Automatic Emotion Recognition: An Investigation of Acoustic and Prosodic Parameters

A thesis submitted for the degree of

Doctor of Philosophy

By

Vidhyasaharan Sethu

Supervisor: **Prof. Eliathamby Ambikairajah**

Co-Supervisor: **Dr. Julien Epps**

School of Electrical Engineering and Telecommunications

The University of New South Wales

November 2009

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……..

Date ……..

COPYRIGHT STATEMENT

Signed …….....

Date …….....

AUTHENTICITY STATEMENT

Signed …….....

Date …….....

Abstract

An essential step to achieving human-machine speech communication with the naturalness of communication between humans is developing a machine that is capable of recognising emotions based on speech. This thesis presents research addressing this problem, by making use of acoustic and prosodic information.

At a feature level, novel group delay and weighted frequency features are proposed. The group delay features are shown to emphasise information pertaining to formant bandwidths and are shown to be indicative of emotions. The weighted frequency feature, based on the recently introduced empirical mode decomposition, is proposed as a compact representation of the spectral energy distribution and is shown to outperform other estimates of energy distribution. Feature level comparisons suggest that detailed spectral measures are very indicative of emotions while exhibiting greater speaker specificity. Moreover, it is shown that all features are characteristic of the speaker and require some of sort of normalisation prior to use in a multi-speaker situation.

A novel technique for normalising speaker-specific variability in features is proposed, which leads to significant improvements in the performances of systems trained and tested on data from different speakers. This technique is also used to investigate the amount of speaker-specific variability in different features. A preliminary study of phonetic variability suggests that phoneme specific traits are not modelled by the emotion models and that speaker variability is a more significant problem in the investigated setup.

Finally, a novel approach to emotion modelling that takes into account temporal variations of speech parameters is analysed. An explicit model of the glottal spectrum is incorporated into the framework of the traditional source-filter model, and the parameters of this combined model are used to characterise speech signals. An automatic emotion recognition system that takes into account the shape of the contours of these parameters as they vary with time is shown to outperform a system that models only the parameter distributions. The novel approach is also empirically shown to be on par with human emotion classification performance.

Keywords: Automatic emotion recognition, group delay features, EMD based weighted frequency, speaker normalisation, contour parameterisation, dynamic modelling.

Acknowledgements

I would like to express my sincere gratitude towards my supervisor Professor Eliathamby Ambikairajah for his unfailing support, guidance, encouragement and advice. This thesis would not have been possible without him. I would also like to thank my co-supervisor Dr. Julien Epps for his numerous suggestions and ideas, and for holding me to a high standard in all my work.

I would also like to thank all the members of the Signal Processing research group at UNSW for their help, Prof. David Taubman, Dr. Deep Sen and Dr. Elias Aboutanios for numerous suggestions and encouragement; Dr. Hadis Nosratighods for discussions on all things signal processing and otherwise; Thiruvaran Tharmarajah, Tet Yap, Karen Kua, Phu Le, Bo Yin, Liang Wang, Ronny Kurniawan, Ning Wang, Reji Mathew, Phyu Khing, Aous Naman, Jonathan Gan, Wenliang Lu, Dan Meng, Yao Wang and Qingqing Meng for their great support and company; and Tom Millet for making sure the lab was much more than just a workplace.

I also wish to thank National ICT Australia for awarding me the NICTA International Postgraduate Award which enabled me to pursue my PhD at UNSW and the School of Electrical Engineering and Telecommunications for supporting me during this period.

Finally, I cannot understate the support I have received from my family and friends during the course of this PhD. In particular, I wish to thank Vikram Rajan and Felicity Allen for their great help on a very short notice in the completion of this thesis; my uncle without whom this PhD would not have been possible; and my sister and my parents for their support, love and encouragement.

To my mother, my sister and to the memory of my father

Acronyms and Abbreviations

AER	**A**utomatic **E**motion **R**ecognition
AI	**A**rtificial **I**ntelligence
AM	**A**mplitude **M**odulation
AR	**A**uto-**R**egressive
BSM	**B**road **S**pectral **M**easures
CDF	**C**umulative **D**ensity **F**unction
CWT	**C**ontinuous **W**avelet **T**ransform
DCT	**D**iscrete **C**osine **T**ransform
DFT	**D**iscrete **F**ourier **T**ransform
DSM	**D**etailed **S**pectral **M**easures
E	**E**nergy (Intensity)
EM	**E**xpectation **M**aximisation
EMD	**E**mpirical **M**ode **D**ecomposition
FD	**F**ractal **D**imension
FF	**F**ormant **F**requencies
FIR	**F**inite **I**mpulse **R**esponse
FM	**F**requency **M**odulation
GCI	**G**lottal **C**losure **I**nstant
GD	**G**roup **D**elay (linear prediction based)
GFCC	**G**ammatone **F**ilter **C**epstral **C**oefficients
GFM	**G**lottal **F**low **M**odel
GMM	**G**aussian **M**ixture **M**odel
HMM	**H**idden **M**arkov **M**odel

HTK	**H**idden markov model **T**ool**K**it
IID	**I**ndependent and **I**dentically **D**istributed
IMF	**I**ntrinsic **M**ode **F**unction
LDC	**L**inguistic **D**ata **C**onsortium
LP	**L**inear **P**rediction
LPRCC	**L**inear **P**rediction **R**esidue **C**epstral **C**oefficients
LTI	**L**inear **T**ime **I**nvariant
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MLE	**M**aximum **L**ikelihood **E**stimation
P	**P**itch
PDF	**P**robability **D**ensity **F**unction
PE	**P**itch & **E**nergy
PhR	**Ph**oneme **R**ate
PNN	**P**robabilistic **N**eural **N**etwork
RAPT	**R**obust **A**lgorithm for **P**itch **T**racking
RC	**R**eflection **C**oefficients
S	energy **S**lope
SC	**S**pectral **C**entroid
SPL	**S**ound **P**ressure **L**evel
SSC	**S**ource **S**pecific **C**ues
SVM	**S**upport **V**ector **M**achines
SZ	energy **S**lope & **Z**ero crossing rate
WF	**W**eighted **F**requency
WS	**W**avelet **S**cale feature
Z	**Z**ero crossing rate (feature)
ZCR	**Z**ero **C**rossing **R**ate

Table of Contents

Chapter 1

Introduction

The development of symbolic language and speech as a means of communications has played a significant role in the evolution of humans. Consequently, speech is probably the most natural and widely used form of interpersonal communication. While in general the primary objective of speech is to convey information encoded as linguistic content, speech is not completely characterised by its linguistic content. Other factors such as the speaker's age, sex, emotional state and cognitive load, collectively referred to as paralinguistic information, influence speech as well. Humans are able to both convey and interpret paralinguistic information in speech with very little effort during the course of any normal conversation.

The development of a machine that is capable of exhibiting the conversational skills of a human being has long been one of the goals of speech processing research. Even without achieving the artificial intelligence (AI) goals of understanding the information conveyed in speech and responding appropriately, the initial ability to recognise linguistic, and to a larger extent paralinguistic, information has not yet been achieved. While a significant volume of research has been carried out over the last six decades in the fields of speech recognition and speaker recognition, research into recognising other aspects of paralinguistic information have only been gaining popularity in recent years[1]. One of these aspects is the emotional state of the speaker and its automatic recognition based on speech is the focus of the work reported in this thesis.

[1] Google Scholar lists 15 papers relevant to emotion recognition published in 1998 compared with more than 100 in 2008.

1.1 Speech Based Emotion Recognition

Human speech is an acoustic waveform, generated by the vocal tract, whose parameters are modulated to convey information. The physical characteristics and the mental state of the speaker determine how these parameters are affected and consequently speech conveys the intended, and on occasion unintended, information. Speech processing research could be described as the effort to determine these parameters, understand how they characterise the information contained in speech, and apply this understanding to practical systems. Even though this knowledge is not explicitly available, the human brain is able to decipher this information from the speech waveform, including the emotional state of the speaker. This ability of a person to recognise the emotional state of the speaker from his or her speech is robust with respect to different speakers, and humans achieve it successfully in many cases, even if it is the first time they have been exposed to that speaker. This suggests the existence of patterns in speech that are characteristic of the emotion being conveyed. These characteristic patterns may exist in many levels, ranging from prosodic and acoustic patterns to patterns in word and language usage, and form the basis for all automatic emotion recognition (AER) systems based on speech.

The importance of such AER systems has increased with the need to improve naturalness and efficiency of speech based human-machine interfaces (Cowie et al. 2001). In general, the aim of an AER system is to extract descriptors that are representative of those patterns in speech that are characteristic of the emotional state of the speaker, while simultaneously unrepresentative of patterns characteristic of all other information. These descriptors (also referred to as features) can then be used to automatically determine the emotional state of the speaker. However, no ideal features are known and the search for the best features (i.e., those that maximise emotion specific information while minimising

dependence on other aspects) is one of the central research themes in the field of speech-based emotion recognition.

Given that ideal features (descriptors) do not exist, pattern classification techniques are used to make a decision about the emotional state based on the chosen features. Herein, based on which aspect of the speech signal they describe, features are broadly categorised into low-level or high-level descriptors. Low-level features describe the acoustic, prosodic or spectral properties of the speech signal, without taking into account the linguistic content explicitly. High-level features on the other hand are based explicitly on the linguistic content without taking into account any variations in the acoustic parameters of the speech signal. Even though evidence suggests that both contain emotion specific information (Lee et al. 2005), in order to limit the complexity of the emotion recognition system, especially given the relative immaturity of the field, a large number of state of the art AER systems do not make use of semantic or linguistic information and rely solely on acoustic, prosodic, and/or spectral features, e.g. (Kwon et al. 2003; Ververidis et al. 2006).

One approach to the search for effective features is to base parameters on some model of speech production and another is to base it on a signal analysis method such as the Fourier transform or the wavelet transform. Given that different features describe different properties of the speech signal, their values vary with any aspect (not only the emotional state) of the speaker that affects these properties. This is obviously undesirable, and ideal features would exhibit minimum variability with respect to other information while retaining emotion specific variability. Hence, both effectiveness of features, and their variability are investigated.

Research in the fields of speech and speaker recognition has provided the speech processing community with established and successful features, and powerful modelling and pattern classification tools. This has allowed for the rapid development of speech

based emotion recognition systems that make use of these tools and features. In a short period of time, even though an exhaustive comparison of all possible combinations is not feasible given the large number of available features and pattern classification techniques, numerous systems with reasonably good performance have been reported. However, this system development approach to research into automatic emotion recognition systems may tend to emphasise rapid performance gains at the expense of an in-depth understanding of why these approaches result in these performances and the relationship of the features used to traditional models of speech.

1.2 Research Issues

An alternative approach is to develop novel features and classification tools specific to the task of emotion recognition. However, one of the major hurdles to this approach is the lack of clear research directions. An attempt to ascertain some of these directions forms a part of this thesis. In particular it looks at the standard source filter model for speech production, interpreting features in terms of the model and studying the variations of these parameters in order to exploit those that are due to and indicative of emotions. It may also be necessary to minimise variations due to other factors.

The principal objective of this thesis is the investigation of emotion-specific patterns in speech parameters, with the aim of using this information in speech based AER systems, but focusing more on developing an understanding of the relationship between these parameters and the emotion being conveyed and less on the actual performance of the AER system. The AER system can be broadly divided in to two stages: (1) A front-end that extracts features from the speech signal, and (2) a back-end that makes a decision based on the features (refer to Figure 1.1).

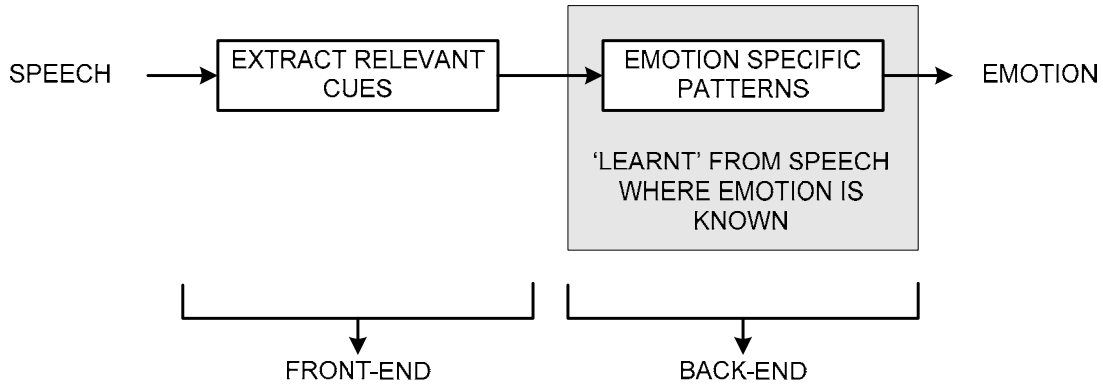

Figure 1.1: Parts of a generic AER (automatic emotion recognition) system

While both stages were implemented in the course of the work described in this thesis, and preliminary comparisons of back-ends are included, a comprehensive evaluation of the best back-end is not part of the scope. The aims of the investigation are:

- To develop and investigate features for automatic emotion recognition to determine which speech parameters (in the framework of a speech production model) are the most representative of emotions.
- To investigate variability in features due to speaker specific information and the normalisation of such speaker specific variability prior to classification.
- To investigate some of the different approaches to emotion classification in order to validate the framework used in the evaluations of features.
- To investigate the use of a three component speech production model, with the intention of employing the model parameters as features. In particular, to determine the importance of taking into account the long-term temporal variations of these parameters.

1.3 Organisation of the Thesis

The remainder of the thesis is organised as follows:

Chapter 2 provides an overview of speech, emotions, speech processing, feature extraction, emotion modelling, and emotion classification. It briefly describes the common approaches to automatic speech based emotion recognition.

Chapter 3 discusses various traditional and novel features as applied to speech based emotion recognition. Different features are based on different aspects of the speech signal. For instance, cepstral coefficients are based on the magnitude spectrum, while pitch is based on the period of the vocal excitation. This chapter reviews the traditional speech production model and attempts to interpret the features in terms of this model.

Chapter 4 investigates the effects of speaker variability on emotion models. In addition to a comparison of the effect of speaker variability in different front-ends, it also proposes a novel speaker normalisation method and it compares speaker variability to phonetic variability.

Chapter 5 describes a few of the modelling approaches adopted for speech based automatic emotion recognition. It includes a preliminary comparison of selected classifiers and an evaluation of a modelling approach that attempts to exploit clustering the feature space that are not related to emotions. It also compares two approaches to modelling the statistics of speech features with different levels of abstraction.

Chapter 6 discusses a three component speech production model as an improvement to the traditional speech production model. The glottal parameters determine the shape of the vocal excitation waveform and influence voice quality while the pitch contour is approximated by linear segments in order to parameterise it. This chapter also discusses subjective and objective evaluations of this model based on listening tests conducted on

re-synthesised speech and automatic classification tests performed using the model parameters as features respectively.

Chapter 7 highlights the achievements of the thesis with respect to the recognition of emotions based on speech. It also discusses opportunities and directions for future work.

1.4 Major Contributions

The research described in this thesis provides original contributions to the automatic recognition of speaker emotional state based on speech. The major contributions can be summarised as follows:

- A novel group delay feature based on an autoregressive (AR) model is proposed for speaker dependent emotion recognition. The group delay highlights formant frequencies and formant bandwidths. The relationship between group delay and formant bandwidth is derived analytically.
- A feature estimated from a definition of instantaneous frequency based on the Hilbert transform and empirical mode decomposition is proposed to characterise spectral energy distribution and its performance in emotion recognition systems is evaluated.
- A comparison of some of the most commonly used prosodic, acoustic and spectral features on the same database with identical back-ends, in speaker dependent and independent scenarios.
- Almost all acoustic and prosodic features are speaker dependent and can result in inefficient estimation of statistics when modelled by classifiers trained on data from multiple speakers. A speaker normalisation technique that is novel in the

AER context is proposed to overcome this problem and its effectiveness is evaluated on 16 distinct acoustic and prosodic features.

- The effect of speaker specific and phoneme specific information on speech based automatic emotion classification is evaluated. Results of this evaluation indicate that speaker variability is more significant than phonetic variability.
- First to use glottal spectral parameters as features in the context of AER, as part of a three component speech production model.
- An AER system that takes into account the shape of the pitch contours is shown to significantly outperform a system that models only the distribution of pitch values.
- The use of glottal and vocal tract parameter contours in addition to pitch contours, is evaluated in the context of emotion recognition. Model parameter approximations are proposed for compact representation as features and a voiced speech synthesis technique based on the model allows for subjective evaluations of the proposed approximations along with objective evaluations based on the use of model parameters as features.
- The performance of an AER system based on the 3 component model parameter contours is shown be comparable to human performance.

1.5 List of Publications

- **Sethu, V.**, Ambikairajah E. and Epps J. (2009) "Pitch contour parameterization based on linear stylization for emotion recognition", in the *Proceedings of INTERSPEECH-09*, pp. 2011-2014.
- **Sethu, V.**, Ambikairajah, E., and Epps J., (2009) “Speaker dependency of spectral features and speech production cues for automatic emotion classification”, in the

Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 4693-4696.

- **Sethu, V.**, Ambikairajah E. and Epps J. (2008) "Phonetic and speaker variations in automatic emotion classification", in the *Proceedings of INTERSPEECH-08,* pp. 617-620.
- **Sethu, V.**, Ambikairajah, E., and Epps J., (2008) “Empirical mode decomposition based weighted frequency feature for speech based emotion classification”, in the *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008,* pp. 5017-5020.
- **Sethu, V.**, Ambikairajah E. and Epps J. (2007) "Group Delay Features for Emotion Detection", in the *Proceedings of INTERSPEECH-07*, pp. 2273-2276.
- **Sethu, V.**, Ambikairajah, E. and Epps, J. (2007)," Speaker Normalisation for Speech-based Emotion detection" in the *Proceedings of 15 International Conference on Digital Signal Processing 2007*, pp. 611-614.
- Le, P. N., Ambikairajah, E., and **Sethu, V.**, (2008) “Speech enhancement based on empirical mode decomposition” in the *Proceedings of 5th IASTED International Conference*, pp. 207-210.
- Wang, Y., An, J., **Sethu, V.**, and Ambikairajah, E. (2007) "Perceptually motivated pre-filter for speech enhancement using Kalman filtering," in the *Proceeding of the 6th IEEE International Conference on Information, Communications and Signal Processing.*

Chapter 2

Speech and Emotions

This chapter discusses the mechanisms involved in human speech production, what emotions are, and how speech could be an indicator of these emotions. It then describes the Emotional Prosody speech corpus, which is used in all the experimental work reported in this thesis. Finally, it provides a brief background to automatic emotion recognition systems. More specifically, it elaborates on what is required for such a system, touching upon different approaches, classifiers and features.

2.1 Human Speech Production

Speech is the vocalised form of human communication. We use it every day almost unconsciously, without devoting much thought to the process. After language processing in the brain which involves conversion of an utterance into phonemes, there are three stages involved: generation of motor commands to the vocal organs; articulation of the vocal organs; and the excitation of the vocal tract by air driven though it by the lungs. Figure 2.1 shows the stages involved in producing human speech (Honda 2003) and Figure 2.2 shows the parts of the human speech production apparatus.

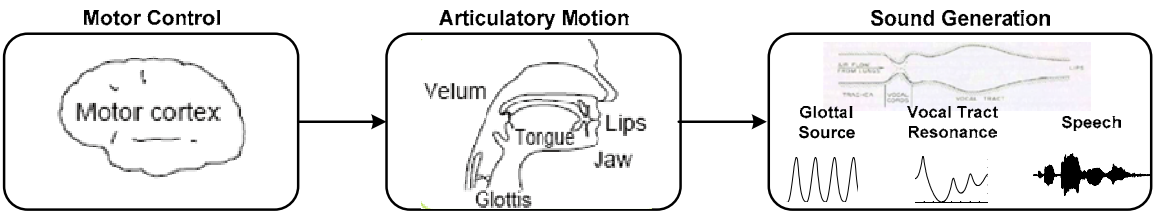

Figure 2.1: Human speech production process, (Honda 2003)

Air flowing through the opening between the vocal folds causes them to draw towards each other until eventually the opening is closed, which causes the air pressure below the folds to increase until they are forced open again. This cyclic opening and closing of the vocal folds modulates the airflow resulting in phonation (production of sound). The sound produced at this stage is characterised only by the fundamental frequency, which is the rate at which the vocal folds vibrate. Alternately, the vocal folds may not vibrate and air flows through a narrow opening, typically created by the position of the vocal folds, tongue, and/or lips, resulting in turbulent airflow and a noise-like sound. Speech characterised by such a sound source is referred to as unvoiced speech (e.g. /s/) as opposed to voiced speech (e.g. /aa/) which requires phonation. The combination of the vocal folds and the space in between the folds is referred to as the glottis.

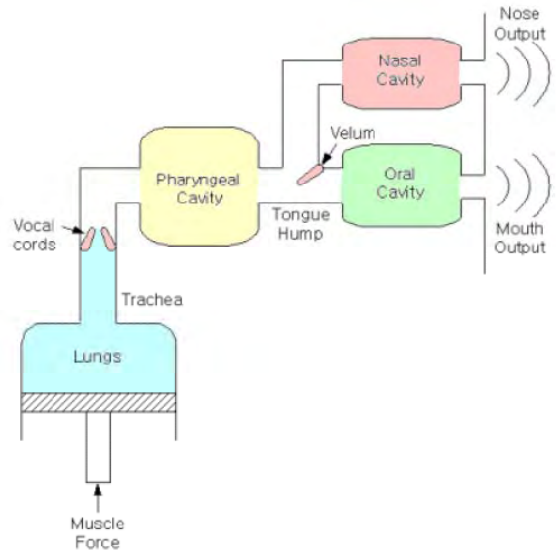

Figure 2.2: Schematic diagram of the human speech production apparatus (Rabiner et al. 1993)

This pulsed or turbulent air stream then excites the vocal tract causing it to resonate at its characteristic frequencies (formants). These characteristic formant frequencies are determined by the shape of the vocal tract which is in turn determined to a certain extent by the position of the jaws, tongue and other parts of the mouth. This allows humans to control the resonance characteristics and consequently the speech sound being produced.

A simplified schematic of the human speech production apparatus is shown in Figure 2.2. Thus speech can be approximated as a signal produced by a sound source, which is then spectrally shaped by the vocal tract. As a result, any physiological changes that affect the organs involved in the process of speech production will have an effect on the speech being produced and the underlying reason for these changes could potentially be determined from the speech.

2.2 Emotion in Speech

In order to build an automatic emotion recognition system, it is essential to have a sense of what an emotion is and how it affects the speech signal. While most people have an informal understanding of emotions, typically built upon the experience of years of interpersonal interactions, a formal framework is a pre-requisite for a thesis on emotion recognition and can be identified from research in the fields of psychology and cognitive sciences. Such a framework must begin with an answer to the question “What are emotions?”. Emotions and feelings, particularly feelings of emotions, are often considered to be the same thing. However, it has been suggested that making a distinction between emotion and the feeling of emotions allows for a testable description (Damasio 2000). One of the most commonly used and accepted frameworks to describe emotions (and one of the few that allows testable hypotheses) is the *component process model* proposed by Scherer (1984).

2.2.1 What are Emotions?

Emotions are specific and consistent collections of physiological responses triggered by internal or external representations of certain objects or situations, such as a change in the person’s body that produces pain, or an external stimulus such as the sight of another

person; or the representation, from memory, of a person, or object, or situation in the thought process. There is some evidence to suggest that the basics of most if not all emotional responses are preset in the genome (Damasio 2000). In a broad sense, emotions are a part of the bio-regulatory mechanism that humans have evolved to maintain life and survive. Emotions form an intermediary layer between stimulus and behavioural reaction, replacing rigid reflex-like response patterns (Scherer 1984) allowing for greater flexibility in behaviour (Tompkins 1962). It has also been postulated that one of the major functions of emotion is the constant evaluation of stimuli in terms of relevance and the preparation of behavioural responses that may be required by these stimuli (Arnold 1960; Scherer 1982). Emotional reactions also serve as a signalling system between organisms and are essential in acquiring new behaviour patterns. It has been pointed out that they are a pre-requisite for learning (Bower 1981; Mowrer 1960).

While the precise composition and dynamics of the responses are specific to an individual (based on environment and individual development), the basic traits are consistent across all humans. In a typical emotion, a part of the brain sends commands to the rest of the body (and other parts of the brain) via chemical molecules in the bloodstream and/or via neuronal pathways, resulting in a global change in the state of the person. Both the body and the brain are profoundly affected by the set of commands, in response to a particular set of sensory patterns (which may have originated internally or externally). This view offered by Damasio (2000) is also consistent with the view inherent in the work of other authors such as Ekman (1992b). However, this view only loosely defines what may be included under the term ‘emotion’ and is not a complete theory of emotions. It should also be noted that while emotions are referred to as ‘states’, they are in fact not static concepts but constantly changing processes.

In terms of the component process model, emotions are treated as psychological constructs consisting of several components each serving distinct functions (Scherer 1984) as listed below. Scherer (1984) also states that there is a fair amount of agreement among researchers that the concept of emotions should indeed encompass all these components.

Component	*Function*
Cognitive stimuli appraisal	Evaluation of an environment
Neurophysiological processes	System regulation
Motivational and behavioural tendencies	Preparation of action
Motor expression	Communication of intention
Subjective feeling	Reflection and monitoring

Based on internal and external stimuli, the state of each of the components is continuously changing (e.g., the sight of a desirable object will change state of the cognitive stimuli appraisal component from ‘seeing an object’ to ‘evaluating it as desirable’; and the state of the motivational component from ‘curious’ to ‘wanting the object’ and so on). An emotion is then conceptualised as a pattern of state changes in these components where each component is influenced by the others (Scherer 1984).

The distinction between emotions and feelings of emotions advocated by Damasio (2000) is inherent in this component process framework as the feeling of an emotion is a state of the subjective feeling component while the emotion itself is a dynamic sequence of states of many components. The distinction between emotions and feelings of emotions is of some importance since emotions include a physiological component while feelings of emotions refer to the private, mental experiences of an emotion and may not be a part of every emotion. It is the physiological and motor expression components, specifically the consistency in the patterns of the changes in their states, which form the basis of

automatic emotion recognition systems. In particular, systematic effects of the emotion specific state sequences of these components on the speech production apparatus forms the basis for speech based AER systems.

2.2.2 Emotions and Speech

Everyday experience tells us that speech is an informative source for the perception of emotions. For instance, talking in a loud voice when feeling very particularly happy, speaking in an uncharacteristically high pitched voice when greeting a desirable person, or the presence of vocal tremor when fearful or sad have all been experienced routinely by a lot of people. This recognition in turn indicates that listeners are able to infer the emotional state of the speaker reasonably accurately – even in the absence of visual cues. Scherer (2003) states that a review of about 30 studies yielded an average recognition rate of about 60%. However, the lack of a common database makes direct comparisons of the recognition rates reported in literature an exercise in futility. Section 2.2.3.3 includes the emotion recognition rates achieved by humans on the data used throughout this thesis.

Based on the definition of emotions as including a physiological component, both voluntary and involuntary effects on the human speech production apparatus can be expected and the characteristics of vocal expression are the net result of these effects. It has been noted that characteristics affecting bodily movement also affect the voice production mechanism and consequently the voice. This is supported by the observation that the vocal expressions of basic emotions is similar in many languages (Fónagy 1981). This work also notes considerable parallels between vocal and physical gestures – for example, an increased tension in the throat causing an increased loudness of speech paralleling an increased tension of the whole body in preparation for an imminent fight. An even more innate ‘frequency code’ with high frequency vocalisation suggesting a submissive attitude and lower frequency vocalisation suggesting greater size and a more

dominant attitude was proposed in (Ohala 1983). Demonstrations suggesting that various aspects of a speaker's physical and emotional state, including age, sex and personality can be identified by voice alone are reviewed in (Kramer 1963). This low-level information is present in even short utterances and could influence the interpretation of the words being uttered, typically identified by "it's not what he said but the way he said it". An analogy from communication interprets the paralinguistic information as an "emotion carrier wave" for the words (Murray et al. 1993). Consequently, emotion can still be recognised even if the linguistic information is not interpreted, this is further supported by the work reported in (Pollack et al. 1960) noting that emotion can be recognised from segments of speech as short as 60ms. Scherer et al. (2001) report an emotion recognition accuracy of 66% on meaningless multilingual sentences by listeners from different cultural backgrounds, and interpret this as evidence for the existence of vocal characteristics specific to emotions.

Various other authors have also hinted at systematic correlations between emotions and acoustic parameters (Darwin 1872; De Gelder 2000; Ekman 1992a; Johnstone et al. 2000). Table 2.1 (reproduced from (Scherer 2003)) and Table 2.2 (reproduced from (Murray et al. 1993)) list some of the relationships between emotions and acoustic parameters as reported in the literature. It should be noted that the relationships that have been reported in literature are not always consistent across all studies and may contradict each other. For instance, Table 2.1 (Scherer 2003) lists heightened intensity for fear while Table 2.2 (Murray et al. 1993) lists normal intensity. However, most relationships are consistent and point towards correlations between emotions and acoustic parameters that can be exploited by an automatic emotion recognition system.

Table 2.1: Synthetic compilation of the review of empirical data on acoustic patterning of basic emotions based on (Johnstone et al. 2000) (reproduced from (Scherer 2003))

	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗(↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	(↗)	
Speech and articulation rate		↗	↗	↘	(↗)	↘

Table 2.2: Summary of human vocal emotion effects (reproduced from (Murray et al. 1993))

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	slightly faster	faster or slower	slightly slower	much faster	very much slower
Pitch average	very much higher	much higher	slightly lower	very much higher	very much lower
Pitch range	much wider	much wider	slightly narrower	much wider	slightly wider
Intensity	higher	higher	lower	normal	lower
Voice quality	breathy, chest tone	breathy, blaring	resonant	irregular voicing	grumbled, chest tone
Pitch changes	abrupt, on stressed syllables	smooth, upward inflections	downward inflections	normal	wide, downward terminal inflections
Articulation	tense	normal	slurring	precise	normal

Reviews of research investigating the effect of emotions on vocal expression can be found in (Cowie et al. 2001; Frick 1985; Murray et al. 1993; Scherer 1986; Scherer 2003).

2.2.3 Emotional Speech Data

A pre-requisite to almost any study on the expression and recognition of emotions via speech is the collection of emotional speech data. However, the lack of common agreement about a theory of emotions complicates this process of data collection. Some

of the broad issues are listed below while a more detailed discussion of emotional speech data-bases can be found in (Douglas-Cowie et al. 2003).

- What are the emotions for which data must be collected ?
- Can data be collected when the emotions are acted out or must emotions be elicited without the speaker being aware of it ?
- Can emotions be considered as discrete labels or are they a continuum ?
- Is it more appropriate to map emotions onto dimensions such as valence, excitation, arousal, etc. and associate dimension values to data rather than emotion names ?

Human languages contain a large number of 'emotion denoting' adjectives. According to (Cowie et al. 2003), the Semantic Atlas of Emotion Concepts (Averill 1975) lists 558 words with 'emotional connotations'. Numbers like these reveal a problem in both collecting data and constructing automatic recognisers that are capable of distinguishing a large number of classes. However, it may be that not all of these terms are equally important and given the specific research aims it could be possible to select a subset of these terms fulfilling certain requirements. A number of such approaches have been proposed including: basic emotions from a Darwinian point of view, which are shaped by evolution to serve functions that benefit survival (Plutchik 1994); emotion categories chosen on the grounds that they are more fundamental than others because they encompass the other emotion categories; and asking people what emotion terms play an important role in everyday life (Cowie et al. 1999).

2.2.3.1 Considering Emotional States

While the aim of the above mentioned approaches is to reduce the number of emotion related terms, it has also been argued that emotions are a continuum and these terms, even

a very large number of them, do not capture every shade of emotion a person can distinguish. The dimensional approach to emotion categorisation is also related to this line of argument in that it describes shades of emotions as points in a continuous two- or three- dimensional space. For instance, in (Cowie et al. 2001), emotional states are described in terms of a two-dimensional circular space, with its axes labelled ‘activation’ (going from passive to active) and ‘evaluation’ or ‘valence’ (going from negative to positive). An important question with the dimensional approach is then if these emotion dimensions capture all relevant properties of the emotion concepts or if they are simplified and reduced descriptions. Opinion is once again divided with Russel et al. (1977) claiming that three dimensions emerging from their factor analysis is “sufficient to define all the various emotional states”, while the opposite view is expressed in (Lazarus 1991). More comprehensive overviews of various descriptive frameworks can be found in (Cowie et al. 2003) and (Schröder 2004).

From an information technology point of view of automatic emotion recognition, a continuum of emotions is an intractable problem at the moment and a finite (and relatively small) number of emotional categories are a necessity. Consequently the two approaches of selecting a set of emotion category labels or using emotion dimensions to describe a finite number of emotional states appear to be equivalent. Given this state of affairs and the lack of agreement on a ‘theory of emotions’, the pragmatic approach of asking people to identify the emotional categories that are most relevant to everyday life is very attractive. Such an approach was adopted in (Cowie et al. 1999) to set up what the authors refer to as a ‘Basic English Emotion Vocabulary’. In a two stage process, they evaluated the probability of various emotion category labels being a part of this basic emotion vocabulary. The 10 most probable emotion labels based on this study are listed

below in decreasing order of their probabilities (i.e., labels that occur higher up on the list have a higher probability of being in the basic emotion vocabulary).

1. Happy
2. Sad
3. Angry
4. Interested
5. Relaxed
6. Worried
7. Pleased
8. Affectionate
9. Bored
10. Confident

In (Cowie et al. 1999), the authors state that many lists based on a priori judgements (such as the other approaches described above) omit terms that appear to be important based on this list, and include others that very few people regarded as useful. This is relevant from the point of view of building an automatic emotion recognition system in that, if it can identify only a finite number of emotions, these should be ones that are most required.

2.2.3.2 Acted vs. Elicited Emotions

On the topic of acted and elicited emotional speech, once again there is no clear consensus. Critics of the acted speech approach question the validity of such data, claiming that such speech may not reflect what people would produce spontaneously. However, this notion is challenged in (Banse et al. 1996) stating that even elicited emotions are ‘acted’, albeit for different reasons. There are numerous advantages in using speech based on acted emotions. Namely, control over the verbal and phonetic content (different emotional states can be produced using the same emotionally neutral utterance); and ease of producing full blown emotions. The high level of control over the linguistic content could also potentially allow direct comparisons of prosodic and voice quality parameters for different emotional states. If acted speech is to be used, it should be noted that actors are able to produce more convincing emotions than non-actors (Schröder 2003).

2.2.3.3 LDC Emotional Speech Corpus

The emotional speech database used in all the experiments reported in this thesis is the LDC Emotional Speech and Transcripts Corpus (Liberman et al. 2002). This database was chosen on the basis of

- Language: It is one of the few English databases available
- Availability: It is publically available (most databases used in emotion recognition studies are not publically available)
- Number of Emotions: It contains a large number of emotional states
- Number of Speakers: It contains speech data from multiple speakers (7 speakers)
- Gender Balance: It contains data from male (3) and female (4) speakers.

It contains audio recordings, recorded at a sampling rate of 22050 Hz, and the corresponding transcripts (word level transcripts that lack time stamps). The recordings were made by professional actors reading a series of semantically neutral utterances consisting of dates and numbers, spanning fourteen distinct emotion categories, selected based on the German study reported in (Banse et al. 1996), and a 'neutral' category that does not involve any emotional state. The categories included are

1. Neutral
2. Hot Anger
3. Cold Anger
4. Panic
5. Anxiety
6. Despair
7. Sadness
8. Elation
9. Happy
10. Interest
11. Boredom
12. Shame
13. Pride
14. Disgust
15. Contempt

Four female and three male actors participated and were provided with descriptions of each emotional context, including situational examples adapted from those used in the German study. Flashcards were used to display series of four syllable dates and numbers

to be uttered in the appropriate emotional category. During the recording, the actors repeated each phrase as many times as necessary until they were satisfied the emotion was expressed and then moved onto the next phrase. Only the last instance of each phrase was included in all the experiments reported herein. This provided about 8 to 12 utterances per speaker for every emotional category. While the phrases recorded for all emotions were not identical, they were very similar to each other and contained numerous words that were common (e.g. 'Two thousand and one' and 'Two thousand and twelve'; or 'December second' and 'December twenty first').

As mentioned previously, not all emotional categories were deemed to be equally relevant in everyday life (Cowie et al. 1999). A comparison of the list of the ten emotional categories deemed most important in everyday life with the emotional categories available in the LDC Emotional speech corpus immediately indicates that happiness, sadness and anger must be included in the experiments. The database contains two versions of anger, hot and cold, and only the more obvious hot anger was used. Neutral was also selected as one of the categories since non-emotional speech is probably more common than emotional speech and since most speech studies (not related to emotions) are based on neutral speech. The other emotional categories that are present on both lists are interest and boredom. Of these, boredom was considered to be important in many of the potential applications of automatic emotion recognition and was chosen as the fifth category but interest was not used in the experiments. The five emotional categories used in the experiments reported in this thesis are listed below. These five emotions are also used in other studies (Huang et al. 2006; Yacoub et al. 2003).

1. Neutral
2. Anger
3. Sadness
4. Happiness
5. Boredom

As part of the work reported in this thesis and in order to act as a reference for all automatic emotion recognition accuracies reported in the rest of the thesis, a listening test was conducted with eleven untrained listeners to determine the accuracy with which humans could classify speech from this database belonging to the above mentioned five emotional categories. The listeners were given an utterance, which they could listen to as many times as necessary, and asked to classify it as one of the five emotions (Neutral, Anger, Sadness, Happiness and Boredom). Each listener classified 15 class balanced utterances (3 from each of the 5 classes) drawn at random from the database. The overall accuracy of all eleven listeners was 63.6 % and the overall confusion matrix is given in Table 2.3. The rows give the actual emotional category and the columns the emotional category into which they were classified (this format is followed in all reported confusion matrices).

Table 2.3: Confusion matrix for five emotions classified by 11 human listeners

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**69.7 %**	3 %	9.1 %	0 %	18.2 %
Anger	3 %	**93.9 %**	0 %	3 %	0 %
Sad	12.1 %	3 %	**57.6 %**	0 %	27.3 %
Happy	39.4 %	3 %	9.1 %	**45.5 %**	3 %
Bored	33.3 %	0 %	15.2 %	0 %	**51.5 %**
Overall Accuracy = 63.6 %					

A few observations can be made from these classification rates. The most obvious one is that anger was identified correctly on almost all occasions and other emotions were mistaken for anger very rarely. Similarly other emotions were mistaken as happiness very rarely even though happiness is often mistaken for no emotion. It is interesting to note that both anger and happiness are emotions associated with a relatively high level of excitation (Osgood et al. 1957) and the other (low-excitation) emotions are rarely mistaken to be one of them. The emotion associated with a low degree of excitation (i.e.,

neutral, sadness and boredom) are however confused with each other on a non-trivial number of occasions.

2.3 Automatic Emotion Recognition

The idea of making a machine that is capable of recognising emotions from speech is not a very new idea (Tolkmitt et al. 1986). However, the popularity of research in this field has grown significantly in recent years, coinciding with the maturing of research in the fields of speech and speaker recognition. A major motivation comes from the desire to develop human-machine interfaces that are more adaptive and responsive to the user's behaviour, thereby making human-machine interactions more natural and closer to human-human interactions (Cowie et al. 2001). For instance, ticket reservation systems that use automatic speech recognition, that are also able to detect annoyance and frustration of the user, could change their response appropriately (Ang et al. 2002). Systems capable of recognising the emotional state of a person based on speech are also useful in call centres (Lee et al. 2005; Petrushin 1999); as diagnostic tools in medicine (France et al. 2000); and as tools to aid in coping with large amounts of speech data in research pertaining to emotions (Mozziconacci et al. 2000).

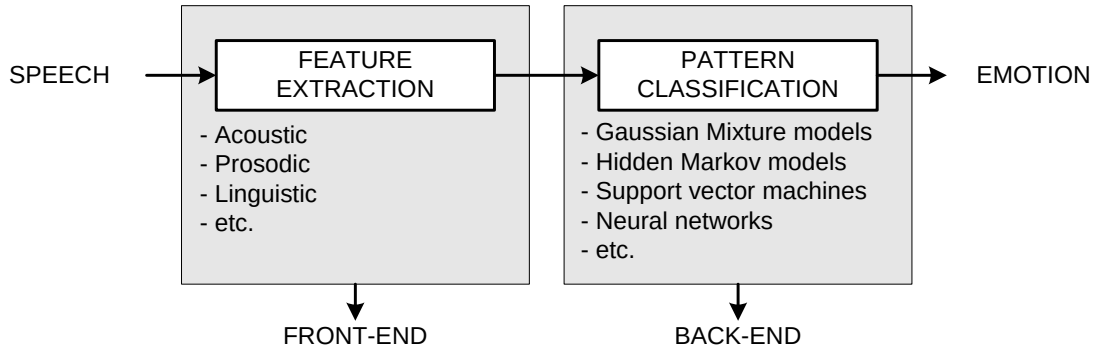

Figure 2.3: Parts of a generic AER (automatic emotion recognition) system

Given that emotions are expressed via speech though numerous cues, ranging from low-level acoustic ones to high-level linguistic content, several approaches to speech based automatic emotion recognition (AER), each taking advantage of a few of these cues, are being explored. However, almost all of them employ a machine learning approach and consist of a front-end which extracts relevant cues (features) from speech and a back-end which models emotion specific patterns of these features (Figure 2.3).

2.3.1 Front-End

As mentioned in section 2.2.2, both voluntary and involuntary effects contribute towards the expression of emotions in speech. The net result of these effects manifests itself as deviations of acoustic, prosodic and linguistic parameters from patterns found in 'neutral speech'. The term 'neutral speech' refers to speech that does not convey any information about the emotional state of the speaker. Acoustic and prosodic parameters are non-verbal parameters such as pitch, loudness, energy spectral distribution, speech rate, etc. which can be extracted without any knowledge of the linguistic content (what is being said) even though they may be related to it. For instance, while pitch is dependent on what is being said as well as other factors (age, sex, emotional state, etc.), it can be extracted without any knowledge of the linguistic content. These features are in most cases short-term features estimated on a frame-by-frame basis. Most of the methods employed in automatic emotion recognition stem from the front-end signal processing developed in the context of speech coding, speech recognition and speaker recognition. Numerous features have been analysed for this task (Barra et al. 2006; Borchert et al. 2005; Lugger et al. 2007; Pantic et al. 2003; Ververidis et al. 2006; Vidrascu et al. 2007; Yacoub et al. 2003) and it would be impossible to list all of them. However the most commonly used acoustic and prosodic features tend to those based on pitch, intensity, cepstral coefficients and speech rate. Although these features are extracted on a frame-by-frame basis, the most

commonly adopted approach in emotion recognition systems is to extract parameters (statistics) from the feature values corresponding to all the frames in an utterance (turn) that is being evaluated. These parameters then form a new feature vector and classification is performed based on this new vector rather than the original features. Commonly extracted statistics include means, standard deviations, quartiles, ranges, extremes, regression coefficients, roll-off points, etc (Schuller et al. 2007; Vlasenko et al. 2007; Yacoub et al. 2003). However, it has been argued that the use of models such as Gaussian mixture models (see section 2.3.2.1) which approximate the probability density functions would intrinsically model these parameters and the frame level features can be used directly (Huang et al. 2006).

Linguistic parameters, unlike the acoustic and prosodic ones, are based on the linguistic content and cannot be extracted without prior speech recognition (automatic or manual). Examples of linguistic parameters include part-of-speech (POS) tags, semantic tags and N-grams, and a number of approaches based on linguistic parameters have been investigated (Ang et al. 2002; Boucouvalas et al. 2002; Cowie et al. 1999; Lee et al. 2002; Litman 2003; Schuller et al. 2009). However, the use of linguistic parameters for automatic emotion recognition is not as widespread as the use of acoustic and prosodic parameters for three predominant reasons, namely: the linguistic approach requires speech recognition which is error-prone and based on a fixed vocabulary; it is language dependent to a much greater extent; and it is significantly more complex than the acoustic and prosodic approach. Measures of the parameters (acoustic, prosodic or linguistic) themselves, or measures of their deviation from the patterns for neutral speech are extracted by an appropriate front-end as features (refer to Chapter 3 for a discussion on features).

2.3.2 Back-End

Given a suitable set of features (cues) representative of the emotional state of the speaker, the role of the back-end is to initially model emotion specific patterns and then perform pattern matching. A number of classification methods have been used for automatic emotion recognition and based on their approach they can be categorised as either generative or discriminative. Generative classifiers try to model the distribution of the training data (features) from each class (emotion) individually (i.e., the models of each class are based only on data from that class and not from any other class). Pattern matching involves estimating some measure of closeness of the unknown data to each of the models and then picking the class whose model is closest to the data. The commonly used generative classifiers are:

- Probabilistic neural networks (PNN) (Specht 1988)
- Gaussian mixture models (GMM) (Reynolds et al. 1995)
- Hidden Markov models (HMM) (Baum et al. 1966; Baum et al. 1970)

Unlike generative classifiers, which attempt to model the entire feature space for each class, discriminative classifiers attempt maximising a discriminative function between the different classes without modelling the distribution of the entire feature space. The main disadvantage of the discriminative classifiers is that their optimal structure has to be selected by trial and error procedures. Some of the commonly used discriminative classification techniques are:

- Linear discriminant analysis (Fisher 1936)
- Polynomial classifier (Specht 1967)
- Recurrent neural networks (Pearlmutter 1995)
- Time-delay neural networks (Lang et al. 1990)

- Multilayer perceptrons (Rosenblatt 1958)
- Support vector machines (SVM) (Vapnik 2000)

Among the classifiers listed above, hidden Markov models, recurrent neural networks and time-delay neural networks are capable of modelling temporal patterns in feature sequences while the other classifiers are insensitive to the temporal order of the features. Both GMMs and HMMs were used in the experiments reported in this thesis and are discussed in this section.

2.3.2.1 Gaussian Mixture Models (GMM)

Appropriate features (cues) are those that enable the separation of classes (emotions) in the feature space (N dimensional vector space determined by the N dimensions of the feature – features are vectors when $N > 1$). Thus each feature vector is a point in this feature space and N dimensional probability density functions (PDFs) describe the distribution of all the feature vectors of each class. An example in a two dimensional feature space with two distinct classes is shown below (Figure 2.4).

Given M classes and their corresponding probability density functions $f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})$, the probability of any feature vector, $\bar{x}$ belonging to a class j is simply $f_j(\bar{x})$. The problem of finding the class, given the feature vector and assuming that all the classes are equiprobable is

$$\arg\max_j \{f_j(\bar{x})\} \tag{2.1}$$

The problems are then estimating the probability density functions themselves, and compactly representing them (the size of the PDF increases exponentially with the dimensionality of the features). Using Gaussian mixture models (GMMs) to model these PDFs provides a solution to both these problem.

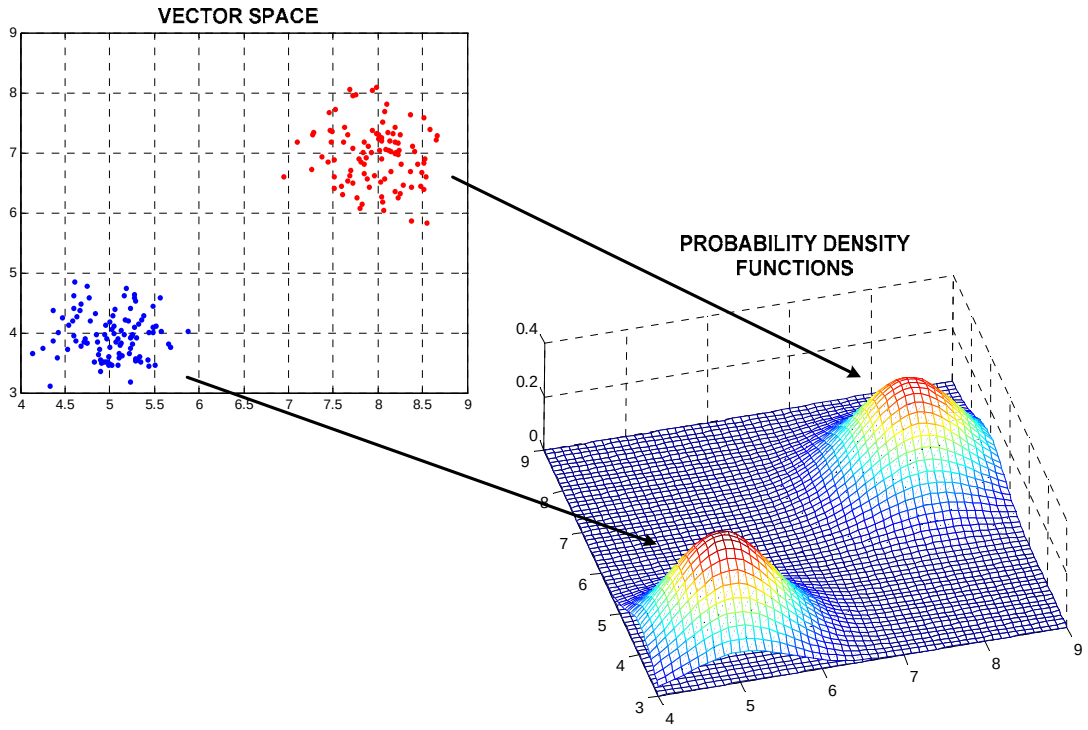

Figure 2.4: Two well separated classes in a 2 dimension feature space and their corresponding probability density functions

A mixture density is a probability density function (PDF) that is expressed as a convex combination (linear combination where all the weights are non-negative and sum to 1, see (2.2) and (2.3)) of other probability density functions. Given a set of probability density functions $p_1(x), \dots, p_n(x)$, referred to as mixture components and corresponding weights $w_1, \dots, w_n$, the weighted sum $q(x)$ is a mixture density

$$q(x) = \sum_{i=1}^{n} w_i p_i(x) \tag{2.2}$$

$$w_i \geq 0 \; and \sum_{i=1}^{n} w_i = 1 \tag{2.3}$$

The mixture components are usually not arbitrary pdfs, but belong to the same parametric family. In the case when they are all normal (Gaussian) distributions, the convex sum is a Gaussian mixture density.

$$q(x) = \sum_{i=1}^{n} w_i \mathcal{N}(x, \mu_i, \sigma_i) \tag{2.4}$$

$$q(x) = \sum_{i=1}^{n} w_i \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \tag{2.5}$$

Any probability density function can be approximated by a Gaussian mixture density given a sufficient number of mixture components n and is referred to as a Gaussian mixture model (GMM). A Gaussian mixture model of n mixtures is parameterised by $3n$ values. Namely, the weights (w_i), means (μ_i) and standard deviations (σ_i) of the n mixture components. An illustration of this for a single dimensional case is shown in Figure 2.5.

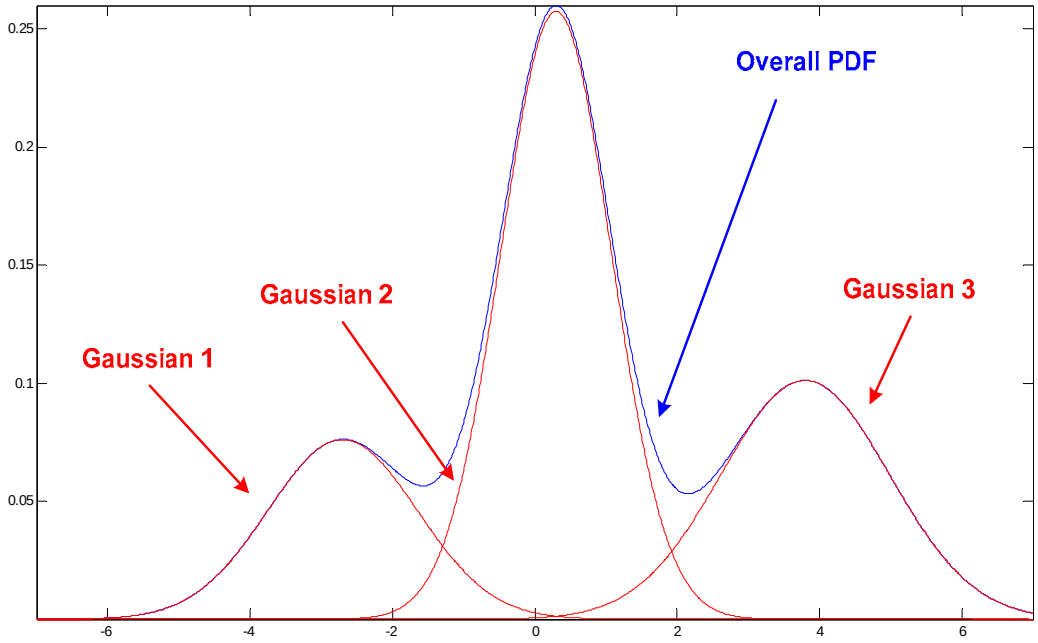

Figure 2.5: A probability density function approximated by a sum of 3 Gaussian mixtures.

In multi-dimensional cases, the means are vectors ($\boldsymbol{\mu_i}$) and covariance matrices ($\boldsymbol{\Sigma_i}$) are used in place of standard deviation.

$$q(\bar{\boldsymbol{x}}) = \sum_{i=1}^{n} w_i \boldsymbol{\mathcal{N}}(\bar{\boldsymbol{x}}, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{2.6}$$

The single dimension model can be considered a special case of the more general multi-dimensional GMM seen in (2.7).

$$q(\bar{\boldsymbol{x}}) = \sum_{i=1}^{n} w_i \frac{1}{(2\pi)^{\frac{N}{2}} |\boldsymbol{\Sigma_i}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\bar{x}-\boldsymbol{\mu_i})^T \boldsymbol{\Sigma_i^{-1}} (\bar{x}-\boldsymbol{\mu_i})} \tag{2.7}$$

Thus, any Gaussian mixture model is then parameterised by the weights, mean vectors and covariance matrices of all its component densities.

$$\boldsymbol{\varpi} = \{w_i, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}, \quad i = 1, \dots, n \tag{2.8}$$

If GMMs are used to model the probability density functions of the feature spaces of each class, the problem of classifying any given feature vector ($\bar{\boldsymbol{x}}$) into one of M classes reduces to

$$\arg\max_j \Pr(\bar{\boldsymbol{x}}|\boldsymbol{\varpi_j}) = \arg\max_j \sum_{i=1}^{n_j} w_{j_i} \frac{1}{(2\pi)^{\frac{N}{2}} \left|\boldsymbol{\Sigma_{j_i}}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\bar{x}-\boldsymbol{\mu_{j_i}}\right)^T \boldsymbol{\Sigma_{j_i}^{-1}} \left(\bar{x}-\boldsymbol{\mu_{j_i}}\right)} \tag{2.9}$$

where, the pdfs of the M classes, $f_1(\boldsymbol{x}), \dots, f_M(\boldsymbol{x})$, are modelled by the Gaussian mixture models, $\boldsymbol{\varpi_1}, \dots, \boldsymbol{\varpi_M}$ and $\boldsymbol{\varpi_j} = \left\{w_{j_i}, \boldsymbol{\mu_{j_i}}, \boldsymbol{\Sigma_{j_i}}\right\}, \quad i = 1, \dots, n_j$.

It should be noted that $\arg\max_j \Pr(\bar{\boldsymbol{x}}|\boldsymbol{\varpi_j})$ and $\arg\max_j \Pr(\boldsymbol{\varpi_j}|\bar{\boldsymbol{x}})$ are not equivalent except when all the classes are equiprobable. In other cases, their relationship is given by the Bayes' theorem as

$$\arg\max_j \Pr(\boldsymbol{\varpi_j}|\bar{\boldsymbol{x}}) = \arg\max_j \{\Pr(\bar{\boldsymbol{x}}|\boldsymbol{\varpi_j}) \Pr(\boldsymbol{\varpi_j})\} \tag{2.10}$$

In order to use Gaussian mixture models for classification the GMMs must first be estimated. The GMM modelling the feature distribution (pdf) of each class is estimated from feature vectors extracted from data that is known to belong to that class (for instance, a Gaussian mixture model for anger is estimated based on features extracted

from speech which known *a priori* to express anger). These data are known as *training data*.

Maximum likelihood estimation (MLE) is used to estimate the model parameters. The likelihood is a function of the model parameters given the observation and is defined as the conditional probability of the observation, given the model. The difference between conditional probability and the likelihood function being that, while the conditional probability has the observation as the independent variable, the likelihood function has the model parameters as the independent variable. Given a set of T independent and identically distributed (IID) feature vectors (observations), $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_T}\}$, and a model $\boldsymbol{\varpi}$, the likelihood of the model is given as

$$\mathcal{L}(\boldsymbol{\varpi}|\boldsymbol{X}) = \Pr(\boldsymbol{X}|\boldsymbol{\varpi}) = \prod_{t=1}^{T} \Pr(\boldsymbol{x_t}|\boldsymbol{\varpi}) \tag{2.11}$$

Maximum likelihood estimation determines the model parameters ($\boldsymbol{\varpi}$) that maximises this likelihood, given the observation (*training data*), $\boldsymbol{X}$. However, this maximisation problem does not have a closed form solution and an iterative procedure called the *Expectation Maximisation (EM) algorithm* (Dempster et al. 1977) is used in most cases.

Often log-likelihoods (log of the likelihood) are used in place of likelihood values to improve numerical precision as the likelihood values tend to be very small (also note that a product of the likelihoods simplifies to a sum of log-likelihoods).

2.3.2.2 Hidden Markov Models (HMM)

A hidden Markov model (HMM) is a doubly stochastic model with an underlying stochastic process that is not directly observable (hidden), but is linked through another set of stochastic processes that produces an observable sequence of symbols. In the context of pattern classification, a sequence of features (observable symbols) is modelled as being generated by a sequence of states (the number of possible states is finite and

unrelated to the number of possible observable symbols) which is not directly observable. At every time instant (corresponding to each of the features in the sequence), the model enters a new state (which may be the same state as the previous one) based on a transition probability distribution which depends on the previous state (Markovian property) and generates the observation (feature) at that instant based on a probability distribution that is associated with that state (regardless of when and how the state is entered). While the number of possible states is always finite, the possible observations (single- or multi-dimensional) may belong to a discrete (and finite) or a continuous set, and thus giving rise to discrete and continuous HMMs respectively.

Any HMM is characterised by the state transition probability distribution, the initial state distribution, and the state observation probability distributions. The state observation pdfs in a continuous HMM are usually modelled by Gaussian mixture models (GMMs) described in section 2.3.2.1. The formal model notations are defined below:

T – length of the observation (feature) sequence

N – number of states

$Q = \{q_1, q_2, \dots q_N\}$ – possible states

x_t – observation at time t

$\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots, \pi_N\}$, $\pi_i = P(q_i|t = 1)$ – initial state distribution

$\boldsymbol{A} = \{a_{ij}\}, a_{ij} = P\left(q_j^t \middle| q_i^{t-1}\right)$ - state transition probability distribution

$\boldsymbol{B} = \{b_j(x)\}, b_j(x) = P\left(x \middle| q_j\right)$- observation probability distribution in state j

The compact notation $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ is used to represent a HMM. Given an observation (feature) sequence $\boldsymbol{X} = \{x_1, x_2, \dots, x_T\}$, the probability of a state sequence $\boldsymbol{I} = \{i_1, i_2, \dots, i_T\}, i_t \in \{1,2, \dots, N\}$ generating it is given by:

$$\Pr(\boldsymbol{X}|\boldsymbol{I},\boldsymbol{\lambda}) = \prod_{t=1}^{T} b_{i_t}(x_t) \tag{2.12}$$

The probability of such a state sequence $\boldsymbol{I}$, on the other hand, is

$$\Pr(\boldsymbol{I}|\boldsymbol{\lambda}) = \pi_{i_1} \prod_{t=2}^{T} a_{i_{t-1}i_t} \tag{2.13}$$

The probability of $\boldsymbol{X}$ and $\boldsymbol{I}$ occurring simultaneously is then given as

$$\Pr(\boldsymbol{X},\boldsymbol{I}|\boldsymbol{\lambda}) = \Pr(\boldsymbol{X}|\boldsymbol{I},\boldsymbol{\lambda}) \cdot \Pr(\boldsymbol{I}|\boldsymbol{\lambda}) \tag{2.14}$$

The probability of the observation sequence $\boldsymbol{X}$**,** given a HMM $\boldsymbol{\lambda}$, is then the sum of the probability of $\boldsymbol{X}$ over all possible state sequences.

$$\Pr(\boldsymbol{X}|\boldsymbol{\lambda}) = \sum_{all\ \boldsymbol{I}} \Pr(\boldsymbol{X},\boldsymbol{I}|\boldsymbol{\lambda}) \tag{2.15}$$

$$\Pr(\boldsymbol{X}|\boldsymbol{\lambda}) = \sum_{all\ \boldsymbol{I}} \Pr(\boldsymbol{X}|\boldsymbol{I},\boldsymbol{\lambda}) \cdot \Pr(\boldsymbol{I}|\boldsymbol{\lambda}) \tag{2.16}$$

Given an observations sequence $\boldsymbol{X}$ and a set of HMMs $\{\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \dots, \boldsymbol{\lambda_M}\}$, the problem of identifying the most probable model can formalised as

$$\arg\max_{\theta}\{\Pr(\boldsymbol{\lambda_\theta}|\boldsymbol{X})\} \tag{2.17}$$

This probability is related to the probability of an observation given a model (eqn. 2.16) by the Bayes' theorem.

$$\Pr(\boldsymbol{\lambda_\theta}|\boldsymbol{X}) = \frac{\Pr(\boldsymbol{X}|\lambda_\theta)\Pr(\lambda_\theta)}{\Pr(\boldsymbol{X})} \tag{2.18}$$

Since the probability of observation $\Pr(\boldsymbol{X})$ is not dependent on the model

$$\arg\max_{\theta}\{\Pr(\boldsymbol{\lambda_\theta}|\boldsymbol{X})\} = \arg\max_{\theta}\{\Pr(\boldsymbol{X}|\boldsymbol{\lambda_\theta})\Pr(\lambda_\theta)\} \tag{2.19}$$

If the model probabilities are not known a priori and the models are assumed to be equally probable, the above relationship further reduces to

$$\arg\max_{\theta}\{\Pr(\boldsymbol{\lambda_\theta}|\boldsymbol{X})\} = \arg\max_{\theta}\{\Pr(\boldsymbol{X}|\boldsymbol{\lambda_\theta})\} \tag{2.20}$$

Hence, in a classification framework if every class is modelled as a hidden Markov model, then a sequence of features (observation) can be classified as belonging to one of the classes using the appropriate relationship from above (eqn. 2.19 or eqn. 2.20). The class models (HMMs) are usually estimated from features extracted from data that is known to belong to the class being modelled (*training data*). The problem of estimating the parameters $\{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$ of a HMM is a difficult one and does not have an analytical solution. Typically iterative procedures, such as the Baum-Welch method, are used. A good overview of hidden Markov models including the Baum-Welch method to estimate the models can be found in (Rabiner et al. 1986).

2.4 Summary

This chapter has provided a brief background to speech, emotions, automatic emotion recognition and data for experiments concerning speech based emotion recognition. It provided an overview of the speech production mechanism in humans before moving on to look at emotions. Section 2.2 set up a working definition of emotions as collections of physiological responses to characteristic internal or external stimuli that are more or less consistent across all humans. It then outlined how emotions thus defined may affect speech which is the basis of any speech based emotion recognition system. Following this, it presented a brief look at some of the issues involved in gathering data for use in the study of emotions and speech based emotion recognition systems. The LDC Emotional Speech and Transcripts corpus which was used in all the experiments reported was then described. This section also presented the results of an experiment performed to evaluate the performance of humans in recognising emotions, in terms of classification accuracy, from the data contained in this corpus.

Finally, this chapter described automatic emotion recognition systems in terms of applications and structure. In particular, it briefly outlined the commonly used classes of features and the ones that will be further explored in this thesis along with the classifiers that will be used.

Chapter 3

Speech Characterisation – Features

This chapter discusses the traditional source-filter model (Fant 1960) of human speech production. While only an approximation, given its relative simplicity the model has been used extensively in almost all aspects of speech processing. The source filter model views speech sounds as being produced by the action of the vocal tract, which is modelled as a filter, on a sound source, either the glottis or some other constriction within the vocal tract (refer to vocal organs depicted in Figure 2.2). An important assumption, fundamental to the model is that the source and the filter are independent. This aids in the analysis of speech sounds, separating the source and vocal tract spectra and allowing for more accurate estimates of speech production parameters. These parameters form the basis for features used in automatic emotion classification systems, and a few of the commonly used ones are discussed in this chapter.

The first section discusses the source-filter model, which is then followed by a description of an automatic emotion recognition system that serves as a common platform to compare the performances of different speech based features. The third section is an overview of some common features and precedes an analysis of novel features proposed for use in emotion recognition systems. The emotion classification accuracies obtained when all the features outlined in sections 3.2 and 3.3 are used individually in the system described in section 3.5 are also included.

3.1 The Source-Filter Model

An overview of the speech production mechanism was given previously in section 2.1. Here the focus is on a commonly used model (Fant 1960) of this speech production mechanism as a linear time invariant (over a short period of time) system excited by an appropriate source.

3.1.1 The Source

Based on the type of excitation, speech sound can be categorised as voiced or unvoiced speech. The source of excitation for voiced speech is the pulsed airflow from the lungs due to periodic vibration of the vocal folds. This is independent of the vocal tract and comprises of a series of glottal pulses. The waveform and spectrum of the glottal source are shown in Figure 3.1 (Harrington et al. 1999).

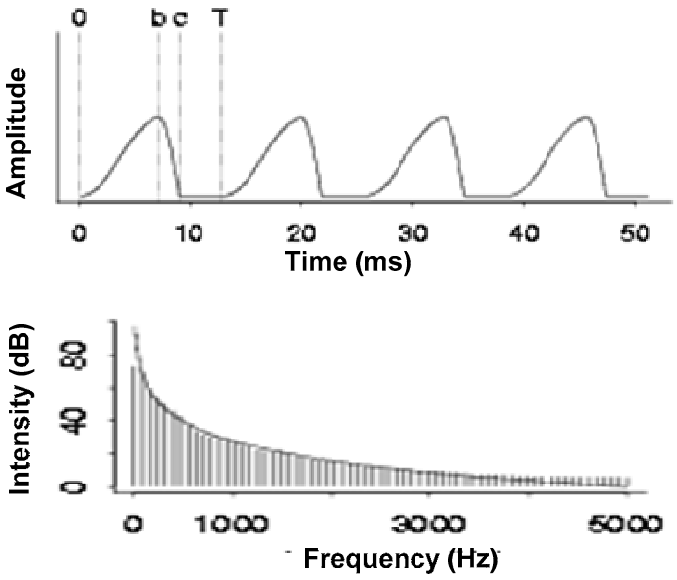

Figure 3.1: A glottal source waveform and the corresponding spectrum (Harrington et al. 1999)

The spectrum of the glottal source is made up of a number of discrete frequency components corresponding to the harmonics of the fundamental frequency of vibration of the vocal folds. The effect of increasing the fundamental frequency on the magnitude

spectrum is to increase the gap between consecutive spectral components, but the overall shape of the spectrum remains unchanged. Pitch is technically the fundamental frequency perceived by a listener, but is often used interchangeably with the fundamental frequency of vibration of the vocal folds (usually denoted as F_0). In this thesis, unless otherwise mentioned, pitch refers to the fundamental frequency of vibration of the vocal folds.

The vocal folds are open in the case of unvoiced speech and do not vibrate. The source of excitation instead is turbulent airflow caused due to a constriction in the vocal tract, which can be at various positions and is caused by the positioning of the tongue, lips, etc. Unlike the periodic excitation of voiced speech, turbulent airflow has no dominant periodic component and has a relatively flat spectrum. It is often described as a noise source, varying randomly.

3.1.2 The Filter

The vocal tract shapes the source waveform to produce the desired speech sound. The shape of the vocal tract is determined by the position of the various articulators such as the position of the tongue, shape of the oral cavity, etc., and this in turn determines how the source waveform is shaped. The vocal tract itself can be considered a continuous tube whose cross sectional area (which is determined by its shape) is a function of position and time. During the course of normal speech, the shape of the vocal tract (cross sectional area) is continuously changing in order to produce the desired sounds. However, when considering very short durations of speech (referred to as frames of speech), typically about 10-20ms, the signal can be considered stationary, i.e., its properties do not change with time. Consequently the shape of the vocal tract can also be considered to be fixed during these intervals. Thus speech can be approximated as a sequence of short segments within which the shape of the vocal tract can be considered to be unchanging, i.e., the cross sectional area can be considered a function of position only and not time, which is

sometimes referred to as the quasi-stationary property of speech. Figure 3.2 taken from (Rabiner et al. 1978) shows a schematic representation of the vocal tract as a tube with a varying cross sectional area (Fant 1960).

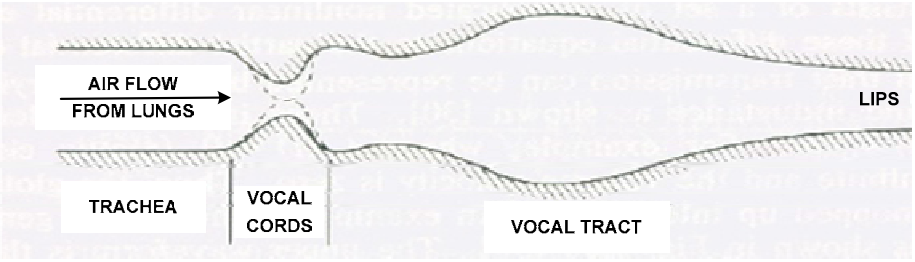

Figure 3.2: Schematic representation of the vocal tract (Rabiner et al. 1978)

The vocal tract can then be further approximated as a concatenation of a number of lossless cylindrical tubes of different cross sectional areas (rather than a tube of continually varying cross sectional area) as depicted in Figure 3.3 (Rabiner et al. 1978). Thus the vocal tract, which is closed at one end by the glottis, can be thought of as having resonant frequencies (which are determined by the lengths and cross section areas of the different sections).

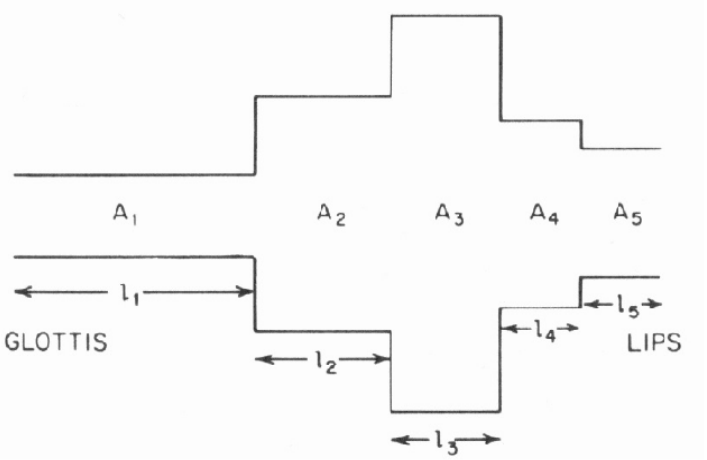

Figure 3.3: Concatenation of 5 lossless acoustic tubes (Rabiner et al. 1978)

The source waveform, either voiced or unvoiced, travelling through this tube is then shaped according to the resonant frequencies. When the shape of the tube is changed the resonant frequencies change as well, resulting in the source waveform being shaped

differently and ultimate producing a different sound. Due to the quasi-stationary property of speech, this approximation of the vocal tract can be considered a linear time invariant system (for the 10-20 ms period) and can be modelled as a fixed filter over this duration. The magnitude response of the filter will reflect the resonant frequencies, which are referred to as formant frequencies, of the lossless tube model (Figure 3.6).

The lips couple the air flow in the vocal tract to the sound pressure wave of the speech waveform and were not considered in the lossless tube model. However, this can be considered to be another filter connected in cascade with the vocal tract filter. The lip-radiation filter has a characteristic spectrum that does not change and can be approximated as a 6dB/octave rise.

3.1.3 Combining Source and Filter

Combining the source and filter models described, the source-filter model of speech production can be schematically represented as shown in Figure 3.4. According to this model speech, $s(n)$ can be viewed as the result of the excitation, $e(n)$ being filtered by the vocal tract filter, $V(z)$ and the lip radiation filter, $R(z)$. The excitation can be either a series of glottal pulses, in the case of voiced speech, or random noise, in the case of unvoiced speech.

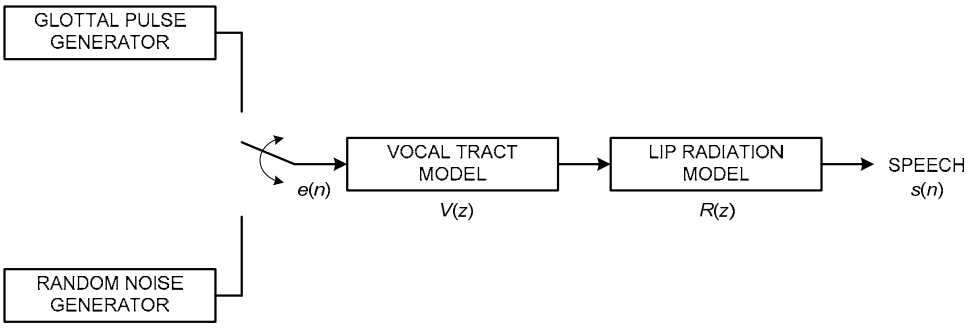

Figure 3.4: The source filter model for speech production.

The model is a linear, time invariant model for the purposes of each 10-20ms frame interval where speech is considered stationary and gives rise to the following relationship.

$$S(z) = E(z)V(z)R(z) \tag{3.1}$$

In the case of voiced speech, the glottal excitation can be further considered the result of the convolution of a train of impulses, separated by the pitch period ($T_p = 1/F_0$, where F_0 is the fundamental frequency), and a single glottal waveform. This is another filtering operation, where the impulse response of the filter is the single glottal waveform.

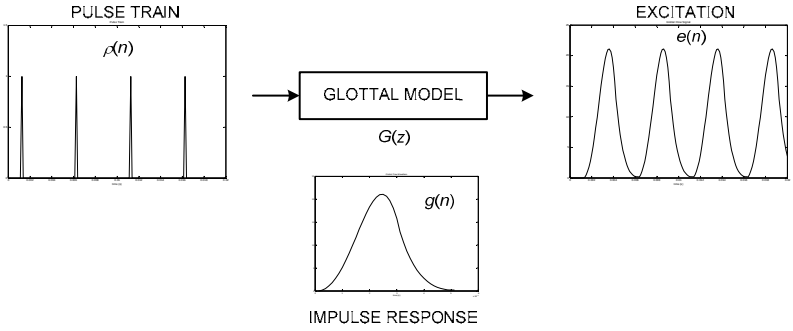

Figure 3.5: Glottal Filter Model

Thus, eqn. 3.1 can be re-written as

$$S(z) = \mathrm{P}(z)G(z)V(z)R(z) \tag{3.2}$$

where, $\mathrm{P}(z)$ is the pulse train and $G(z)$ is the glottal transfer function.

As previously mentioned, the lip radiation filter has a constant magnitude spectrum approximated by a 6dB/octave rise and is usually modelled as a single pole high pass system (eqn. 3.3). The glottal magnitude spectrum is commonly approximated as a 12dB/octave fall and modelled as a two pole low pass system, with both poles at 1 (eqn. 3.4).

$$\begin{aligned} R(z) &= 1 - az^{-1} \\ &\approx 1 - z^{-1}, (\because a \to 1) \end{aligned} \tag{3.3}$$

$$G(z) = \frac{1}{(1 - z^{-1})^2} \tag{3.4}$$

This allows the transfer function voiced speech production system to be approximated in terms of only the vocal tract filter.

$$H(z) = \frac{S(z)}{\mathrm{P}(z)} = \frac{1}{1 - z^{-1}} V(z) \tag{3.5}$$

The vocal tract filter is typically, and quite successfully, modelled as an autoregressive (AR) system. The order of the AR model must be selected appropriately, and a commonly used rule of thumb gives the order as $2 + round(F_s/1000)$, where F_s is the sampling rate. In order to estimate the AR vocal tract model parameters, the speech signal, $s(n)$, is initially high pass filtered with a single pole filter, $H_p(z) = 1 - z^{-1}$, to cancel the effect of the combined glottal and lip radiation models, $G(z)R(z) = 1 / (1 - z^{-1})$. Linear prediction analysis techniques are then used to determine the vocal tract filter model parameters (Makhoul 1975). The high pass filtering of speech prior to analysis in order to study only the vocal tract model is referred to as pre-emphasis. The resonances of the AR model correspond to the resonances of the lossless tube allowing for the formant frequencies (resonant frequencies) to be determined from the speech signal. The magnitude response of a 10th order AR model of 45ms of voiced speech (the phoneme $/\varepsilon/$) is shown in Figure 3.6 and the magnitude spectrum of the pre-emphasised speech signal is given in Figure 3.7.

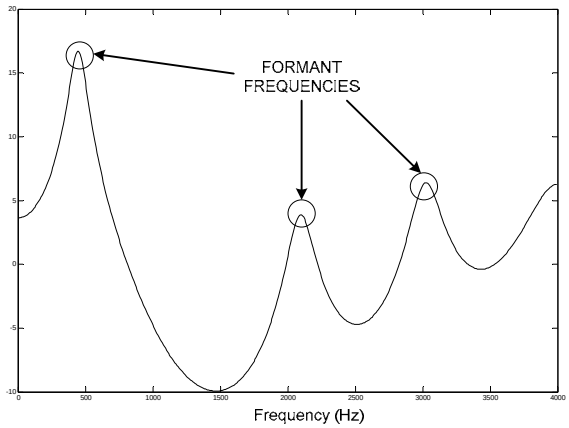

Figure 3.6: Magnitude response of an AR vocal tract model for voiced phoneme $/\varepsilon/$.

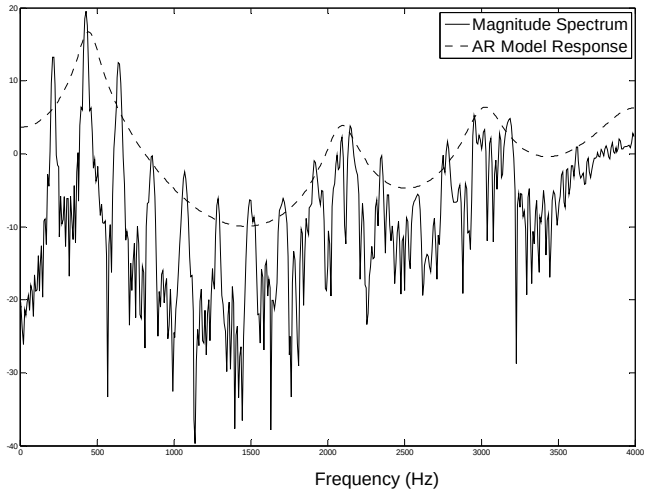

Figure 3.7: Magnitude Spectrum of speech corresponding to phoneme /ε/ and the Magnitude Response of the corresponding AR model (10^{th} order).

It can be seen that the magnitude response of the vocal tract model is the envelope of the magnitude spectrum of the signal. The fundamental frequency (F_0) of the speech signal analysed here was 210 Hz and distinct spectral components can be observed at multiples of 210 Hz in Figure 3.7. From another point of view, the magnitude spectrum of speech signal is the magnitude response of the corresponding vocal tract model sampled at the pitch (F_0) harmonics. This is expected from the speech production model described in this section, since pre-emphasised speech is given by:

$$S_p(z) = (1 - z^{-1})S(z) \tag{3.6}$$

Based on (3.2), (3.3), (3.4) and (3.6)

$$S_p(z) = V(z)P(z) \tag{3.7}$$

Thus, pre-emphasised speech can be considered the response of the vocal tract filter model to an impulse train.

$$s_p(n) = \mathcal{V}\{\rho_N(n)\} \tag{3.8}$$

$$\rho_N(n) = \sum_{k=-\infty}^{\infty} \delta(n - kN) \tag{3.9}$$

where, $\mathcal{V}\{\cdot\}$ is the vocal tract system and $\rho_N(n)$ an impulse train with period N.

The spectrum of pre-emphasised speech is thus given by the product

$$S_p(\omega) = V(\omega)P(\omega) \tag{3.10}$$

where, $V(\omega)$ is the frequency response of the vocal tract model and $P(\omega)$ is the spectrum of the impulse train, $\rho_N(n)$, and is given by

$$P(\omega) = \frac{2\pi}{N}\sum_{k=-\infty}^{\infty}\delta\left(\omega - \frac{2\pi k}{N}\right) \tag{3.11}$$

The period, N, is given by $N = 1/F_0$ and F_0 is fundamental frequency. i.e.,

$$P(\omega) = \omega_0\sum_{k=-\infty}^{\infty}\delta(\omega - k\omega_0) \tag{3.12}$$

where $\omega_0 = 2\pi F_0$.

Hence the spectrum of the impulse train is a sequence of equal magnitude impulses at multiples of the fundamental frequency; and since the spectrum of pre-emphasised speech is the product of the vocal tract response, $V(\omega)$, and $P(\omega)$ as indicated in (3.9), it can be viewed as the vocal tract response sampled at harmonics of the fundamental frequency.

$$S_p(\omega) = V(\omega)\omega_0\sum_{k=-\infty}^{\infty}\delta(\omega - k\omega_0) = \omega_0\sum_{k=-\infty}^{\infty}V(\omega - k\omega_0) \tag{3.12}$$

Most of the commonly used features (but not all) in speech based automatic emotion recognition (AER) systems describe some aspect of speech and consequently some aspect of the source-filter model.

3.2 Typical Features used in AER Systems

This section looks at some of the commonly used features in automatic emotion recognition (AER) system. The features listed in this section have been used in one or more AER systems reported in the literature and shown to be useful for the task of recognising emotions. It also considers them in the context of the traditional source filter

model of speech production whenever possible in order to ascertain the relative significance, if any, of the different aspects/parameters of the speech production system in the context of emotion recognition.

In order to evaluate and compare the effectiveness of different features described in this chapter, it is necessary to quantify their effectiveness. This is done by using them as features in an automatic emotion recognition system and using the classification accuracy of the system as a measure of effectiveness. A 5-class (*neutral*, *anger*, *happiness*, *sadness* and *boredom*) AER system was used for this purpose and is explained in more detail in section 3.5. The system was used in both speaker dependent (training and testing data from the same speaker) and speaker independent (training and testing data from different speakers) scenarios and the overall accuracies in both scenarios are listed along with the feature descriptions in this chapter. The more detailed confusion matrices for all features in both scenarios are presented in Appendix A.

3.2.1 Mel Frequency Cepstral Coefficients (MFCCs)

Mel frequency cepstral coefficients (MFCCs) stem from the more generic filter bank analysis, wherein the signal is filtered by a bank of band pass filters and the energy of the outputs of these filters provide an estimate of the spectrum of the signal. In the case of MFCCs, the filter bank consists of a series of triangular filters, equally spaced in the Mel scale. The Mel frequency scale is given as

$$f_{MEL} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{3.13}$$

Since the triangular filters are arranged linearly in the Mel frequency scale, when looked in the linear scale (Hz), the filters are close to each other and have narrow bandwidths at the low frequency and the spacing and bandwidth increase with frequency. This non-linear analysis of speech is based on the cochlear of the human auditory system.

The energies, or more commonly the log of the energies, of the outputs of the filters provide a low dimensional estimate of the magnitude spectrum and are often used as features in classification systems. In the case of MFCCs, log energies are calculated followed by the discrete cosine transform (DCT) to obtain the cepstrum (a cepstrum can be loosely considered to be the Fourier transform of the log magnitude spectrum). The first few DCT coefficients (around 12) are then used as the features, leading to the name Mel frequency cepstral coefficients. The log energies of adjacent filter bands tend to be correlated and the use of cepstral coefficients reduce this correlation and also allow for further reduction in dimensionality at the cost of finer details of the magnitude spectrum.

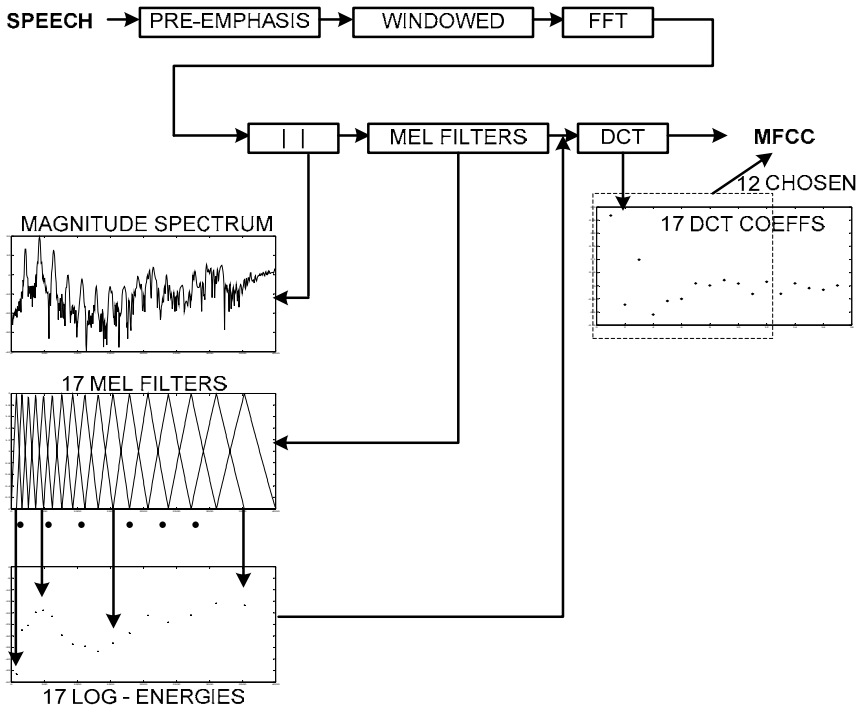

Figure 3.8: Overview of MFCC computation

To compute the MFCCs, the speech signal is initially pre-emphasised to remove the effects of the glottal and lip radiation models and then windowed into short frames (typically 20ms to 30ms). The magnitude spectrum of each frame is then computed and multiplied with the filter response of each of the triangular filters. The averages of each of these 'filtered' spectra are then calculated and their logarithms computed to obtain the

average log-energy within each filter band. A discrete cosine transform (DCT) is then performed and the first N coefficients are selected to obtain the N-dimensional MFCCs. An overview of MFCC computation is shown in Figure 3.8, using 17 triangular Mel scale filters and choosing 12 DCT coefficients to form the final 12-dimensional MFCC vector.

Mel frequency cepstral coefficients contain information about both the source and filter (vocal tract) of the source-filter model. The low frequency filters are closely spaced and have sufficient resolution to capture information about the fundamental frequency (source characteristic), while the entire filter bank spans the magnitude spectrum to obtain an estimate of the spectral envelope (vocal tract characteristics). Also, the first cepstral coefficient is representative of the energy of the signal (another source characteristic).

The overall classification accuracies of a speaker dependent and a speaker independent GMM based automatic emotion recognition (AER) system, described in section 3.5, using MFCCs in the front-end are reported in Table 3.1.

Table 3.1: Summary of Overall Accuracies using MFCCs

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	49.7 %
Speaker **DEPENDENT** AER system	74.8%

3.2.2 Formant Frequencies

The characteristics of the vocal tract determine to a large extent, the sound produced by the speech production apparatus, in turn determining the shape of the magnitude spectral envelope. This shape is typically characterised by a few discrete peaks, which usually occur at the resonant frequencies of the vocal tract. Frequencies at which these peaks occur are referred to as formant frequencies. These are the dominant spectral components in speech and contain a significant amount of information. This makes them very

attractive as features and have been used so in numerous applications. Linear prediction analysis models the vocal tract as an AR (autoregressive) system allows for the estimation of the model parameters. The magnitude response of the AR model (all-pole filter) corresponds to the spectral envelope of the speech signal and allows for the estimation of the formant frequencies. Along with the formant frequencies, the spectral magnitudes of the frequency response of the vocal tract model at these frequencies may also contain information, and are occasionally appended to the formant frequencies to form the feature vector. It should be noted that the formant frequencies are characteristic of the vocal tract model only and do not contain any information about the source (excitation).

The classification accuracies of a speaker dependent and a speaker independent system using a six dimensional feature vector composed of the frequencies and gains of the first three formants as the front-end is reported in Table 3.2.

Table 3.2: Summary of Overall Accuracies using formant information

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	43.7 %
Speaker **DEPENDENT** AER system	58.3%

3.2.3 Reflection Coefficients

Since the vocal tract is modelled as an autoregressive system, the linear prediction (LP) coefficients completely characterise the model. Also, the total number of coefficients is usually small (around 10 per frame for speech sampled at 8 kHz). This makes the use of LP coefficients as features very tempting. However, LP coefficients are very sensitive and can change by large amounts for small changes in the signal and this makes them poor features. An alternative to LP coefficients are reflection coefficients, which are not as sensitive. While LP coefficients are the coefficients of a direct form implementation of

the filter model, the reflection coefficients are obtained from the equivalent lattice form implementation. In the context of the loss tube approximation, where the vocal tract is approximated by a series of concatenated lossless cylinders of different cross section areas, the reflection coefficients characterise the boundaries between adjacent sections. If the cross section area of the i^{th} section is given by A_i, reflection coefficient, Γ_i is given by

$$\Gamma_i = \frac{A_{i-1} - A_i}{A_{i-1} + A_i} \tag{3.14}$$

The Levinson-Durbin algorithm (Rabiner et al. 1978) can be used to obtain both LP and reflection coefficients from the speech signal.

Table 3.3 gives the classification accuracies obtained when using reflection coefficients as features. The AR model order was chosen as 24 (sampling rate of the data was 22.05 kHz) and hence there were 24 reflection coefficients thus giving a 24 dimensional feature vector.

Table 3.3: Summary of Overall Accuracies using reflection coefficients

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	48.9 %
Speaker **DEPENDENT** AER system	71.2%

3.2.4 Pitch

The pitch is a property of the source producing the vocal excitation and is independent of the vocal tract. A single pitch value is determined from every window (frame) of speech. Numerous algorithms have been suggested over the years to estimate F_0 from speech signals. Among these, one of the most popular algorithms is the robust algorithm for pitch tracking (RAPT) proposed by Talkin (Talkin 1995). This algorithm is used to extract pitch for use in all experiments reported in this thesis. Figure 3.9 shows a speech

waveform for the utterance '*thousand*', the corresponding spectrogram and the pitch estimated from the signal using RAPT superimposed on the spectrogram.

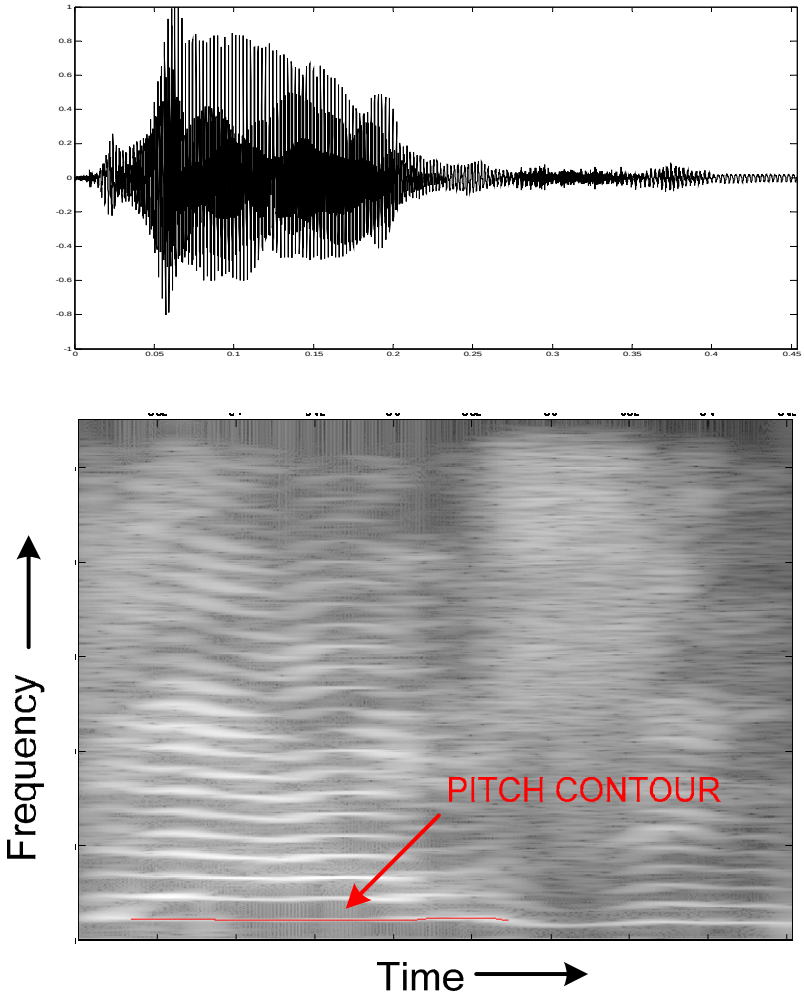

Figure 3.9: The waveform of a speech signal and its spectrogram with the estimated pitch

A single pitch (F_0) value is extracted per frame and the classification accuracies obtained when it is used as a feature in both speaker independent and speaker dependent scenarios are given in Table 3.4.

Table 3.4: Summary of Overall Accuracies using pitch

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	46.6 %
Speaker **DEPENDENT** AER system	51.8%

3.2.5 Intensity (Energy)

The intensity of speech is a measure of the energy contained in speech as it is produced, which in turn is based on the energy of the vocal excitation (since the vocal tract is a passive system). Like pitch, it is a property of the excitation source and the vocal tract filter model is independent of intensity. Loudness of speech as perceived by a listener on the other hand depends on the sound pressure level (SPL) of the sound waves at the eardrum, which is dependent on both the intensity of the speech and the distance of between the speaker and the listener. When analysing recorded speech, loudness depends on the sound pressure level at the microphone which in turn is related to the amplitude of the recorded signal via the microphone transfer function and hence loudness is characterised by the energy of the signal. Consequently, when using loudness as a measure of vocal excitation intensity, an assumption that the speakers are always at the same distance from the microphone is being made. However, since all the data used the experiments reported in this thesis was obtained from the same recording studio, this assumption is reasonable.

Often it is desirable to estimate the change in loudness with time instead of employing a single loudness parameter for an entire utterance. In such cases, rather than energy of the entire speech signal, energy within short frames (windows) may be estimated, and is referred to as short-term energy. Typically 20ms-30ms frames, within which speech can be assumed to be stationary, are used. Given a window, $w[n]$, the short term energy of a speech signal, $s[n]$, within that window is given as

$$E = \sum_{n=-\infty}^{\infty} |s[n]|^2 w[n] \tag{3.15}$$

The short-term energy contour of the speech waveform shown in Figure 3.9 is given below.

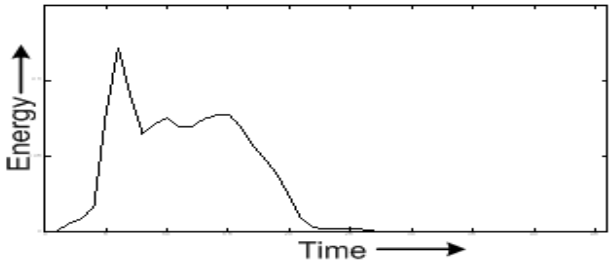

Figure 3.10: Short-Term Energy Contour

Like pitch, intensity is a single dimensional feature with one intensity value per frame. The accuracy of the AER system using only intensity as a feature in speaker independent and speaker dependent tests is given below.

Table 3.5: Summary of Overall Accuracies using energy

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	28.8 %
Speaker **DEPENDENT** AER system	25.2%

3.2.6 Energy Slope (Spectral Slope)

The parameters of the source filter model are representative of all the information contained in speech, both linguistic and paralinguistic. However, in the context of automatic emotion recognition (AER), this is not necessarily advantageous. While emotion specific variability in the parameters form the basis for an AER system, variability due to the linguistic content and other paralinguistic factors (such as age, sex, speaker's identity, etc.) tend to degrade the performance of any AER system that do not make use of linguistic and paralinguistic information as features, albeit to varying degrees based on system configuration. This is particularly true of the parameters of detailed vocal tract model and features characterising the entire spectrum such as MFCCs. Features that characterise some aspect of the spectrum or the vocal tract model, but do not describe it completely may reduce the amount of other variability with respect to emotion

specific variability. Energy slope (also referred to as spectral slope or spectral balance) roughly describes the distribution of energy in the magnitude spectrum, given by the slope of a linear approximation to the magnitude spectrum. Figure 3.11 shows a magnitude spectrum and its linear approximation. The slope of this straight line is the energy slope.

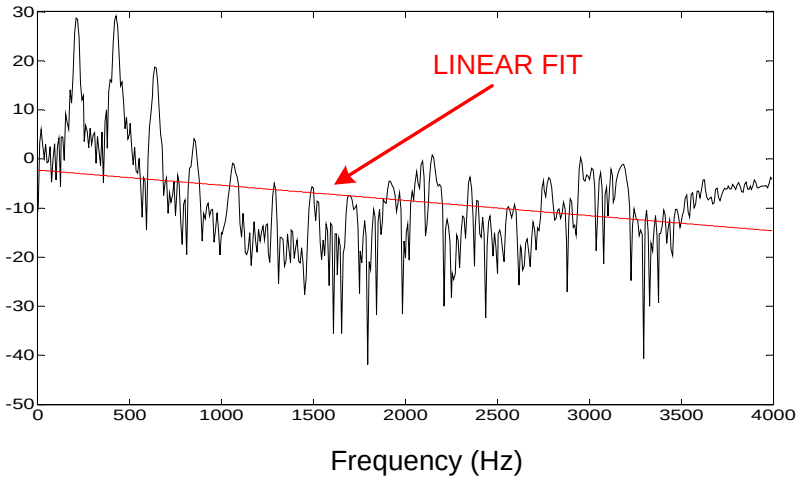

Figure 3.11: Magnitude spectrum of /ε/ and linear fit of the spectrum.

Disregarding the spectral structure even further, an estimate of the slope can be approximated as the ratio of energy in low frequencies to that in high frequencies. In the work reported in this thesis, energy slope was estimated this way and the low and high frequency regions were chosen as 0-1 kHz and 2-11 kHz. This is similar to the energy slope used in (Huang et al. 2006). Table 3.6 reports the accuracies of AER systems using energy slope as their feature.

Table 3.6: Summary of Overall Accuracies using energy slope

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	43.4 %
Speaker **DEPENDENT** AER system	59.0%

3.2.7 Zero Crossing Rate (ZCR)

Zero crossing rate (ZCR) is another parameter that characterises only a part of the spectrum. It serves as a rough estimate of the dominant frequency in the speech signal encapsulated in a single dimensional frame based feature. ZCR has been used as a feature for emotion recognition in (Huang et al. 2006; Lugger et al. 2007). For the work reported in this thesis it was calculated as

$$ZCR = \frac{1}{2}\sum_{n=-\infty}^{\infty} |sign(s[n]w[n]) - sign(s[n-1]w[n-1])| \tag{3.16}$$

where, $s[n]$ is the speech signal and $w[n]$ is the framing window.

The classification accuracies of a speaker independent and a speaker dependent system using only ZCR in their front-ends are given in Table 3.7.

Table 3.7: Summary of Overall Accuracies using ZCR

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	47.1 %
Speaker **DEPENDENT** AER system	46.8%

3.2.8 Spectral Centroid

The spectral centroid is another way to condense the information contained in the speech spectrum. It is the weighted mean frequency, with the spectral magnitudes as weights. The spectral centroid in each frame is single dimensional and was computed as

$$spec\_centroid = \frac{F_s \sum_{i=1}^{N} |X(i)| \cdot i}{N \sum_{i=1}^{N} |X(i)|} \tag{3.17}$$

where, N is the frame size, $X(k)$ is the DFT of the framed speech signal and F_s is the sampling rate.

Table 3.8 reports the classification accuracies obtained when using the spectral centroid, a single dimensional feature, in the front-end of speaker independent and speaker dependent AER systems.

Table 3.8: Summary of Overall Accuracies using spectral centroid

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	40.2 %
Speaker **DEPENDENT** AER system	44.6%

3.2.9 Phoneme Rate

On the face of it, speaking rate appears to be a high-level parameter requiring prior speech recognition. However, it is possible to estimate speech rate based on factors such as the duration of voiced segments, the number of syllables in a period, etc., which do not require speech recognition. Speech rate has been used as a feature in numerous emotion recognition systems (Banse et al. 1996; Dellaert et al. 1996; Lee et al. 2005; Ververidis et al. 2006). Here, speaking rate was estimated from the number of phonemes in each 0.5s window. A phone recogniser developed at the Faculty of Information Technology, Brno University of Technology was used to generate the phonetic labels (Schwarz et al. 2006).

The overall classification accuracy for a speaker independent system and that of a speaker dependent system using phoneme rate are given in Table 3.9.

Table 3.9: Summary of Overall Accuracies using phoneme rate

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	20.6 %
Speaker **DEPENDENT** AER system	23.0%

3.3 Novel Features - AER Performance

This section describes novel features used in the context of automatic emotion recognition. The features described in section 3.2, while used in AER systems, were not originally motivated by such systems. Thus it is possible that they are not the most efficient representations of the information that can be used to discriminate between emotions. The features proposed in this section emphasise information contained in speech differently from the typical features and are investigated to determine if they are better suited for an AER system. While the hypotheses when initially proposing these features were that they would improve the performance of the AER system, failure to do so also provided information about how emotions can be recognised from speech.

3.3.1 Gammatone Filter Cepstral Coefficients (GFCC)

The gammatone filter cepstral coefficients (GFCCs) are identical to the Mel frequency cepstral coefficients except for the analysis filter bank. The GFCCs use gammatone filters instead of the triangular Mel scale filters utilised in MFCCs. The gammatone filters (Aertsen et al. 1980; Flanagan 1960; Katsiamis et al. 2007) are modelled on the cochlear filters unlike the Mel scale filters, whose positions are perceptually motivated but are shaped to minimise computational complexity. The GFCCs, like the MFCCs, will capture source (energy) and filter (spectral envelope) characteristics.

Table 3.10 gives the overall accuracies of a speaker independent and a speaker dependent AER system using gammatone filter cepstral coefficients as features. The feature vector consisted of the first 13 DCT coefficients and 30 gammatone filters were using in the feature extraction process. These parameters are identical to those used in

extraction of MFCCs (section 3.2.1) allowing for a direct comparison of the classification accuracies.

Table 3.10: Summary of Overall Accuracies using GFCCs

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	55.6 %
Speaker **DEPENDENT** AER system	72.6%

3.3.2 Proposed Linear Predictive Model Group Delay

The source-filter model of speech production describes the speech signal as the response of an all-pole filter (AR model) to a suitable glottal excitation. The all-pole filter characterises the vocal tract, and its magnitude response provides information about formant locations, which determine the sound (phoneme) produced, and estimates of formants have been used as features in speech recognition systems. However, when the same phoneme is uttered by the same person in different emotional states, formant positions may not vary much and consequently may not be very useful in distinguishing between the emotional states. Figure 3.12 shows the formant positions obtained from the same phoneme as part of the same word, uttered by the same person in different emotional states. It can be observed that while locations of the first three formants are not very different, the formant bandwidths are very different and produce the difference in the sounds that help distinguish between the two emotions. This change in bandwidth is reflected in the group delays of the corresponding all-pole filters (shown in Figure 3.13). The magnitude of the group delay increases with a reduction in the formant bandwidth and the positions of the group delay's local minima reflect the formant locations.

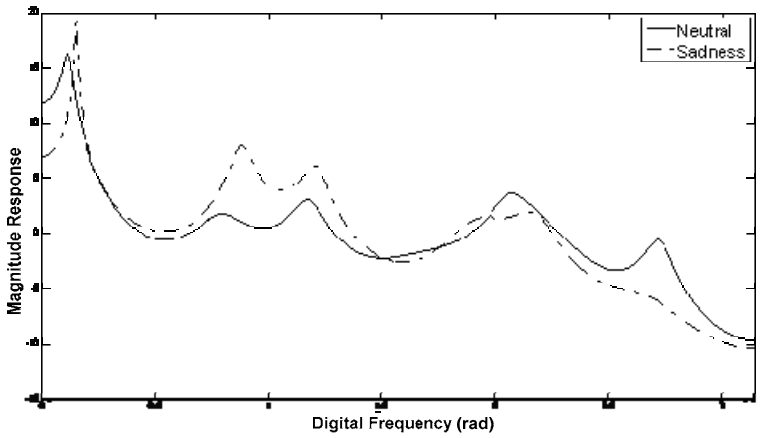

Figure 3.12: Formant locations for /ε/ for two emotions spoken by the same person

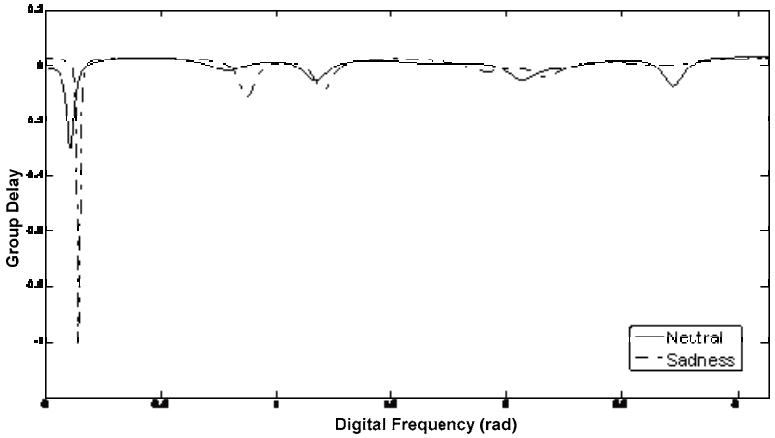

Figure 3.13: Group delay for /ε/ for two emotions spoken by the same person

In order to see the relationship between formant bandwidths and the group delay value at the formant frequency, we examine the transfer function of the all-pole filter that characterises the vocal tract. This can be considered to be a cascade of second order resonators with conjugate poles, with each resonator producing a formant.

$$V(z) = \prod_{i=1}^{M} H_i(z) \tag{3.18}$$

where, $H_i(z)$ is the transfer function of the i^{th} second order resonater.

Setting $z = e^{j\omega}$ gives the frequency response of the all-pole filter.

$$V(\omega) = \prod_{i=1}^{M} H_i(\omega) \tag{3.19}$$

$$|V(\omega)|e^{j\theta(\omega)} = \prod_{i=1}^{M} |H_i(\omega)|\, e^{j\phi_i(\omega)} \tag{3.20}$$

where, $V(\omega)$ is the transfer function of the all-pole vocal tract filter, $H_i(\omega)$ is the transfer function of the i^{th} formant, $\omega \in [-\pi, \pi]$ and M is the total number of formants.

The group delay of the all-pole filter can thus be written as the sum of the group delays of the resonators

$$\theta(\omega) = \sum_{i-1}^{M} \phi_i(\omega) \tag{3.21}$$

Thus, studying the relationship between the group delay value at the resonant frequency and the formant bandwidth for the 2-pole resonator should be adequate. In order to do so, consider the frequency response of a 2-pole resonator.

$$H_i(\omega) = \frac{1}{(1 - re^{j\alpha_i}e^{-j\omega})(1 - re^{-j\alpha_i}e^{-j\omega})} \tag{3.22}$$

where, α_i is the formant (resonant) frequency and $re^{\pm j\alpha}$ are the poles of the system.

The squared magnitude response and consequently the formant bandwidth are computed as

$$|H_i(\omega)|^2 = \frac{1}{[1 + r^2 - 2r\cos(\omega - \alpha_i)][1 + r^2 - 2r\cos(\omega + \alpha_i)]} \tag{3.23}$$

When the poles are near the unit circles, i.e., r is close to but less than 1, the formant bandwidth, $\Delta\omega$, can be approximated as

$$\Delta\omega \approx 2(1 - r) \tag{3.24}$$

Also, from eqn. (3.22), the phase response of the system can be computed as

$$\phi_i(\omega) = -\left[\arctan\left(\frac{r\sin(\omega - \alpha_i)}{1 - \cos(\omega - \alpha_i)}\right) + \arctan\left(\frac{r\sin(\omega + \alpha_i)}{1 - \cos(\omega + \alpha_i)}\right)\right] \tag{3.25}$$

The group delay is obtained by differentiating the phase response with respect to frequency. For the 2-pole resonator, the group delay obtained is as follows

$$\tau_g(\omega) = \left[\frac{r^2 - r\cos(\omega - \alpha_i)}{1 + r^2 - 2r\cos(\omega - \alpha_i)} + \frac{r^2 - r\cos(\omega + \alpha_i)}{1 + r^2 - 2r\cos(\omega + \alpha_i)}\right] \tag{3.26}$$

At the resonant frequency, α_i, the group delay takes the value

$$\tau_g(\omega) = \frac{-r}{1 - r} + \frac{r[1 - \cos(2\alpha_i)]}{1 + r^2 - 2r\cos(2\alpha_i)} \tag{3.27}$$

It can be seen that as the value of *r* approaches 1, the group delay function's value at the formant frequency takes an increasingly negative value since the magnitude of the first term in eqn. (3.27) is always larger than the magnitude of the second term for all $r > 0.2361$.

From equations (3.24) and (3.27) it can be seen that a reduction in the formant bandwidth is reflected by an increasingly larger negative value of the group delay at the formant frequency. Since the overall group delay of the all-pole filter is the sum of the group delays of the resonators, we can expect the group delay to have negative spikes at formant locations, with the magnitudes of these spikes reflecting the formant bandwidths.

In order to estimate the group delay, the all-pole filter parameters are estimated using the LPC algorithm (The model order is set to 24 since the data is sampled at 22.05 kHz). The phase response of this filter is estimated from the first 1024 samples of the impulse response and the group delay is calculated by differentiating this phase response with respect to frequency. Alternatively, equation (3.27) gives the contribution of each complex conjugate pole pair to the overall group delay, which can thus be estimated by adding the contributions of all the poles present in the vocal tract filter.

The group delay computed this way is a vector with a large number of components. In order to represent it compactly, a discrete cosine transform (DCT) is applied to the sequence and the first 10 coefficients ($\beta_0 - \beta_9$) are picked as the elements of the

proposed feature vector. Figure 3.14 shows the steps involved in computing the LPC group delay feature vector.

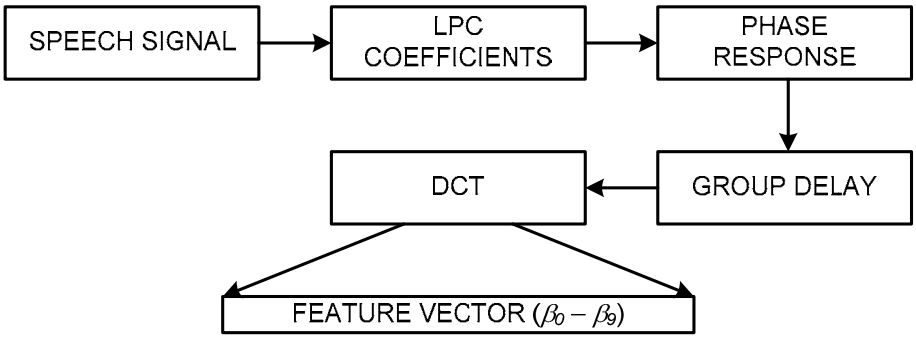

Figure 3.14: LP based Group Delay feature extraction

The ten dimensional group delay feature vector described above was used in speaker independent and speaker dependent AER systems and the resultant classification accuracies are reported in Table 3.11.

Table 3.11: Summary of Overall Accuracies using Group Delay

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	42.9 %
Speaker **DEPENDENT** AER system	69.8%

3.3.3 Frequency Modulation

The cepstral coefficients (both MFCCs and GFCCs) are representative of the magnitude spectrum, in particular the shape of the envelope. However, since information about the average energy in the pass bands of these filters is retained but all information about the shape of the spectrum within these pass bands is discarded, the cepstral coefficients are also dependent on these pass bands and consequently the filter bank. The frequency modulation (FM) feature is an attempt to reduce this dependency by extracting information about the shape of the magnitude spectrum of a signal within the pass bands of the filters in any filter bank. The magnitude spectra of the filtered signals can be

modelled by a 2nd order AR model, whose resonant frequency identifies the dominant peak in the magnitude spectra within the pass bands (Thiruvaran et al. 2008). The FM features are obtained by computing the difference between the 2nd order resonant frequency in each band and the centre frequency of the bands. This taken together with the cepstral coefficients which would contain information about the average energy in each band would roughly characterise the shape of the magnitude spectra in every band. It should be noted that the implicit assumption is that these magnitude spectra contain only one dominant frequency. The 2-pole method for extracting frequency modulation (FM) feature was proposed in the context of speaker recognition in (Thiruvaran et al. 2008).

The classification accuracies obtained when using these FM features with a gammatone filter bank are reported in Table 3.12. However, since the FM features are intended to add detail to cepstral coefficients and minimise the impact of the choice of filter bank, it is more meaningful to consider a combination of cepstral coefficients and FM features extracted using identical filter banks. Such a test was performed and the classification accuracies obtained when using the FM features and GFCCs are in Table 3.13.

Table 3.12: Summary of Overall Accuracies using FM

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	44.4 %
Speaker **DEPENDENT** AER system	64.0%

Table 3.13: Summary of Overall Accuracies using GFCC + FM

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	47.1 %
Speaker **DEPENDENT** AER system	73.4%

3.3.4 EMD based Weighted Frequency (WF)

The recently pioneered empirical mode decomposition (EMD) (Huang et al. 1998) can be used to represent the speech signal as a sum of zero-mean AM-FM components which then allow for the definition of a positive instantaneous frequency for each component based on the Hilbert transform. A major hurdle in the use of the Fourier transform in signal analysis is that the basis functions (sinusoids) are infinitely long and consequently any interpretation of their frequencies as frequencies present in the signal is physically meaningful only when the signal is stationary (within the analysis window). The EMD on the other hand imposes no such restriction. A weighted frequency feature based on the instantaneous frequencies of these components would be low dimensional and contain information about the spectral magnitude distribution.

3.3.4.1 Empirical Mode Decomposition (EMD)

Any real-valued signal can be written as an analytic signal by setting it as the real part of the analytic signal and its Hilbert transform as the imaginary part of the analytic signal

$$z(t) = x(t) + i\mathcal{H}\{x(t)\} \tag{3.28}$$

where, $x(t)$ is the real valued signal and $\mathcal{H}\{\cdot\}$ is the Hilbert transform operator and $z(t)$ is the analytic signal.

From the analytic signal, the instantaneous phase can be obtained and the time derivative of the instantaneous phase is then defined as the instantaneous frequency. The complex analytic function also allows for the definition of instantaneous amplitude.

$$\phi(t) = \arctan\left(\frac{\mathcal{H}\{x(t)\}}{x(t)}\right) \tag{3.29}$$

$$\theta(t) = \frac{d\phi(t)}{dt} \tag{3.30}$$

$$a(t) = \sqrt{x^2(t) + \mathcal{H}\{x(t)\}^2} \tag{3.31}$$

where, $\phi(t)$ is the instantaneous phase, $\theta(t)$ is the instantaneous frequency and $a(t)$ is the instantaneous amplitude.

A problem for most methods of instantaneous frequency estimation occurs when sudden changes in the amplitude or frequency of the signal result in the instantaneous frequency paradoxically taking negative values (Cohen 1995). The empirical mode decomposition (EMD) decomposes any signal as a sum of signals, referred to as intrinsic mode functions (IMF), that have positive instantaneous frequencies as defined in (3.30). A description of the decomposition process and a brief analysis of its application to speech signals is presented in Appendix B.

3.3.4.2 Weighted Frequency Feature

The speech spectrum changes according to the phoneme being uttered, the speaker and the emotional state of the speaker, among other factors. Consequently, the changes in the different IMF instantaneous frequencies due to changes in speech content and the different vocal tract characteristics of different speakers makes using them directly as features for an emotion classifier impossible. However, computing the weighted average of the instantaneous frequencies of the first five modes, with the instantaneous amplitudes acting as weights may give a broad spectral parameter that can be used as a feature. The weighted frequency, $w_f[n]$ is defined as

$$w_f[n] = \frac{\sum_{m=1}^{M} a_m[n]\theta_m[n]}{\sum_{m=1}^{M} a_m[n]} \tag{3.32}$$

where, n is the sample index, $a_m[n]$ and $\theta_m[n]$ are the instantaneous amplitude and frequency of the m^{th} IMF and $M = 5$ (number of modes). The weighted frequency, $w_f[n]$, has as many samples as the frame.

Weighted frequency is computed from the speech signal without any pre-processing and takes into account the spectral shaping imposed by the vocal tract onto the vocal chord excitations. The weighted frequency is indicative of the energy distribution in the speech spectrum (Figure 3.15), taking small values when most of the energy is concentrated in the low frequencies (first formant) and larger values when higher frequencies (higher formants) contain more energy. This information is useful for discriminating between emotions, as shown in Figure 3.15, and similar differences in the weighted frequency due to emotional states were observed in other phonemes as well.

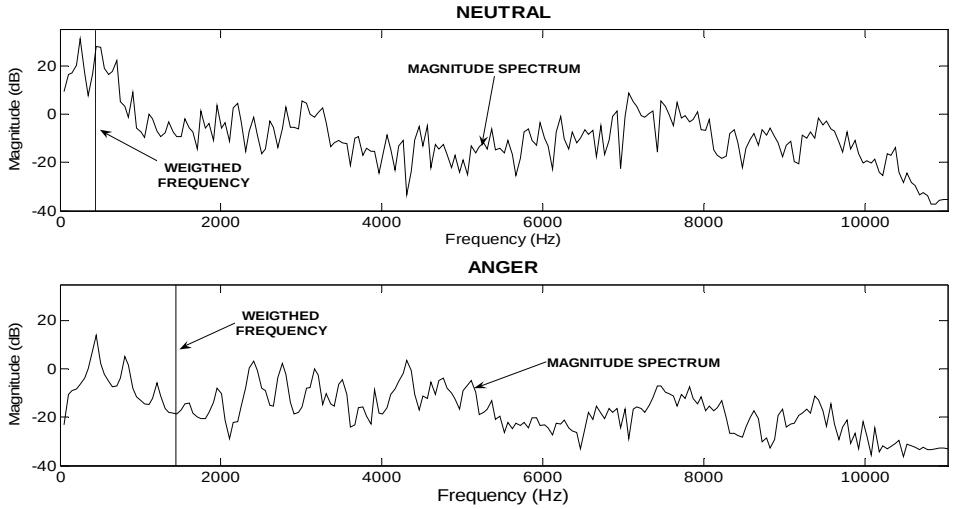

Figure 3.15: Magnitude spectra and average weighted frequency for 20ms frames of speech of phoneme /ε/ for two emotions (a) Neutral; (b) Anger

For use in AER systems, a weighted frequency feature was computed from $w_f[n]$ (not directly feasible as a feature as every sample of the signal will result in an estimate of $w_f[n]$) for 40ms frames using the EMD sifting process and stopping conditions suggested in (Rilling et al. 2003). The discrete cosine transform of this weighted frequency signal in each frame was then obtained and the first three coefficients were selected as a feature vector to represent that frame of data (Figure 3.16).

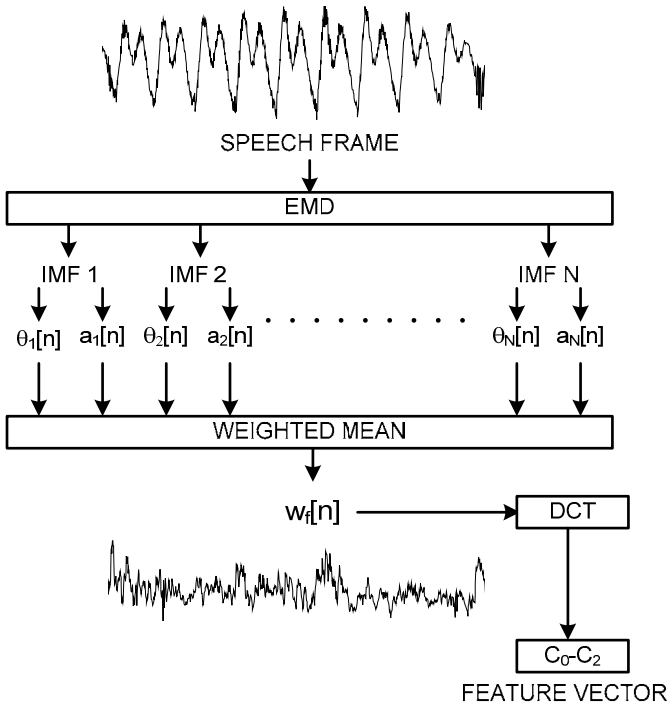

Figure 3.16: Weighted frequency feature extraction.

Table 3.14 gives the classification accuracies of the AER systems using the EMD based weighted frequency as its features.

Table 3.14: Summary of Overall Accuracies using Weighted Frequency

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	47.6 %
Speaker **DEPENDENT** AER system	53.2%

3.3.5 Wavelet Scale based Feature

Instantaneous frequency is not a well defined concept mathematically which makes analysis of the weighted frequency a hard task. A Fourier transform based spectral analysis on the other hand has the advantage of a well defined concept of frequency, but does not have a local time support. Hence, a wavelet transform might be a good compromise, offering both temporal and spectral localisation. The continuous wavelet transform (CWT) of a signal, $f(t)$, is defined as

$$F_{\psi}(a,b) = \int_{-\infty}^{\infty} f(t)\psi\left(\overline{\frac{t-b}{a}}\right)dt \tag{3.33}$$

where, $F_{\psi}(a,b)$ is the wavelet transform, $\psi(t)$ is the mother wavelet and $\bar{z}$ indicates a complex conjugate. a and b are the scale and shift coefficients relating to spectral and temporal localisation.

Similar to the way Fourier coefficients are interpreted as a measure of energy as a function of frequency, the continuous wavelet transform coefficients, $F_{\psi}(a,b)$, can be interpreted as an estimate of energy as a function of scale (frequency) and time. Overviews of wavelet transform and time-scale signal analysis can be found in (Vetterli et al. 1992) and (Allen et al. 2004). Computing the CWT for a finite set of discrete scales allows for the computation of a weighted average frequency (scale), w_{sc}, using the wavelet coefficients as weights in a manner similar to that adopted for weighted frequency in section 3.3.4.2.

$$w_{sc}[n] = \frac{\sum_{\forall \boldsymbol{a}} a \cdot F_{\psi}[a,n]}{\sum_{\forall \boldsymbol{a}} F_{\psi}[a,n]} \tag{3.34}$$

The wavelet scale features used in the experiments reported in this thesis were computed using $[1,2,3 \ldots 64]$ as the scales and the Gaussian wavelet as the mother wavelet. The classification accuracies obtained when using these features in the front-end are reported in Table 3.15.

Table 3.15: Summary of Overall Accuracies using Wavelet Scale feature

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	41.8 %
Speaker **DEPENDENT** AER system	54.7%

3.3.6 LP Residue Cepstral Coefficients (LPRCC)

All features discussed until this point have been discussed in terms of the traditional speech production model discussed in section 3.1. This model however makes certain assumptions about the source characteristics; in particular the assumption about the envelope of the spectrum of the vocal excitation is a gross approximation. A more accurate modelling of the vocal excitation (Doval et al. 2006; Fant et al. 1985) has been shown to be integral for synthesis of natural sounding speech. It is therefore reasonable to assume that features that describe the spectrum of the vocal excitation may be valuable in the context of emotion recognition. Since the vocal tract is modelled as an all-pole filter (AR model), filtering of the speech signal by its inverse filter (all-zero FIR filter) should give an estimate of the vocal excitation (LP residue). The spectrum of this estimate can be represented compactly by the MFCCs of the excitation and is termed the LP residue cepstral coefficients.

A 13 dimensional feature vector comprising of the cepstral coefficients of the LP residue using a Mel filter bank was used in the experiments reported here. The overall classification accuracies of the speaker independent and speaker dependent AER systems using this vector in their front-end are given in Table 3.16.

Table 3.16: Summary of Overall Accuracies using LPRCCs

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	50.0 %
Speaker **DEPENDENT** AER system	68.4%

3.3.7 Fractal Dimension (FD)

The turbulence produced by airflow has previously been discussed as the source of vocal excitation for unvoiced sounds and it is also possible that some amount of turbulence is

present during the production of voiced sounds as well. However, the effect of these turbulences typically manifest as geometrical complexity and fragmentation of the time waveforms of speech; and due to lack of a better approach, are either not modelled or treated as noise. It has been proposed that the theory of fractals can be used to model this complex geometry and in particular use the idea of fractal dimension to quantify the degree of fragmentation (Maragos et al. 1999).

3.3.7.1 Fractal Geometry

The term '*Fractal*' coined by Mandelbrot (Mandelbrot 1983) is based on the concept of *self-similarity*, in which an object appears to be similar to itself when viewed at different scales. A very important parameter for the description of fractals is the fractal dimension, D. Intuitively, it is a quantity that gives an indication of how completely a fractal fills a space.

Suppose, the length of a curve, L, is determined as the number of yardsticks of length ϵ that can fit sequentially along it. The length of a fractal curve would be a function of ϵ and will increase as ϵ decreases, following (approximately) the power

$$L(\epsilon) = k \cdot \epsilon^{1-D}, \quad \text{as } \epsilon \to 0 \tag{3.35}$$

where, D is the fractal dimension of the curve.

3.3.7.2 Minkowski-Bouligand Dimension

This is one of several 'fractal dimensions' that are more or less capable of quantifying the degree of fragmentation of a curve. It is based on Minkowski's idea of finding the length of irregular curves (Maragos et al. 1993):

1. Dilate them with disks of radius ϵ by forming the union of these disks centred on all points of the curve and thus create a "Minkowski cover".
2. Find the area $A(\epsilon)$ of the dilated set at all scales ϵ.

3. Set the length of the curve as $\lim_{\epsilon \to 0} L(\epsilon)$, where $L(\epsilon) = A(\epsilon)/2\epsilon$. If the curve is a fractal, L behaves as in (3.27). Let

$$\lambda(A) \triangleq \sup\left\{p: \lim_{\epsilon \to 0} A(\epsilon)\epsilon^{-p} = 0\right\} \tag{3.36}$$

$$= \lim_{\epsilon \to 0} \frac{\log A(\epsilon)}{\log \epsilon} \tag{3.37}$$

be the infinitesimal order of A.

4. Bouligand defined the dimension, D_M as

$$D_M = 2 - \lambda(A) \tag{3.38}$$

$$= \lim_{\epsilon \to 0} \left(2 - \frac{\log A(\epsilon)}{\log \epsilon}\right) \tag{3.39}$$

The method presented in (Maragos 1991) based on a morphological covering method to estimate D_M was used to estimate the fractal dimension from speech in the experiments reported. The estimation method is discussed in more detail in (Maragos et al. 1999) and (Maragos et al. 1993). Table 3.17 reports the classification accuracies obtained by the AER systems using fractal dimension (FD) as their feature.

Table 3.17: Summary of Overall Accuracies using FD

Classification Test	Accuracy
Speaker **INDEPENDENT** AER system	42.9 %
Speaker **DEPENDENT** AER system	69.8%

3.4 Discussion and Summary

This chapter has discussed the traditional source-filter model of speech production before considering selected features typically used in the front-end of automatic emotion recognition systems. These features are either based on the source-filter model or can be interpreted as being representative of some parameters of the model. Based on this

dependence, the features can be representative of the source (e.g. pitch), filter (e.g. reflection coefficients) or both (e.g.MFCCs).

This chapter also described some novel features for use in AER systems, namely, GFCCs, group delay, FM features, weighted frequency, wavelet scale feature, LPRCCs and fractal dimension. All of these features, apart from fractal dimension, can also be interpreted in the framework of the source-filter model. All of these features were used individually as features for an automatic emotion recognition system (section 3.5) in speaker independent and speaker dependent scenarios to allow for a comparison based on classification accuracies. Table 3.18 lists overall accuracies obtained for all of these features.

Table 3.18: Overall classification accuracies for various features.

Features	Feature Dimension	Speaker Independent	Speaker Dependent
MFCC	13	49.7 %	74.8 %
Formant Frequencies (FF)	6	43.7 %	58.3 %
Reflection Coefficients (RC)	24	48.9 %	71.2 %
Pitch (P)	1	46.6 %	51.8 %
Intensity/Energy (E)	1	28.8 %	25.2 %
Energy Slope (S)	1	43.4 %	59.0 %
Zero Crossing Rate (Z)	1	47.1 %	46.8 %
Spectral Centroid (SC)	1	40.2 %	44.6 %
Phoneme Rate (PhR)	1	20.6 %	23.0 %
GFCC	13	55.6 %	72.6 %
LP based Group Delay (GD)	10	42.9 %	69.8 %
Wavelet Scale Feature (WS)	1	41.8 %	54.7 %
LPRCC	13	50.0 %	68.4 %
Frequency Modulation (FM)	30	44.4 %	64.0 %
FM + GFCC	43	47.1 %	73.4 %
Weighted Frequency (WF)	3	47.6 %	53.2 %
Fractal Dimension (FD)	1	46.3 %	41.0 %

A comparison of the classification accuracies reveals a few interesting trends. The high dimensional features that characterise the spectral shape (and consequently the vocal tract filter model) perform very well in the speaker dependent scenario. These include

- Mel frequency cepstral coefficients (MFCC)
- Formant frequencies (FF)
- Reflection coefficients (RC)
- Gammatone filter cepstral coefficients (GFCC)
- LP based group delay (GD)

Among these MFCCs and GFCCs are very similar to each other and while the use of gammatone filters appears to improve the speaker independent performance, it does not appear to have a similar effect on speaker dependent performance and the difference between the two features do not appear to be very significant. Given that the investigation of the effect of different types of filter banks is a relatively minor detail and beyond the scope of the work presented in this thesis, and the widespread use of MFCCs in speech processing literature, GFCCs are not considered henceforth. The frequency modulation features are not representative of the vocal tract filter on their own, but are intended to improve the spectral resolution of cepstral coefficients. However, comparing the classification accuracies obtained when using GFCCs and a combination of GFCCs and FM, it appears that the added complexity of including the FM features outweighs any advantage.

The low dimensional features that are characteristic of the vocal tract filter, that were considered in this chapter are

- Energy slope (S)
- Zero crossing rate (Z)
- Spectral centroid (SC)

- EMD based weighted frequency (WF)
- Wavelet scale feature (WSC)

Energy slope and ZCR characterise the spectral distribution of energy and the dominant frequency (usually F_1) respectively. However, the other three features are all different types of weighted spectral average and can be expected to characterise similar information. Hence only one of them needs to be considered for any system. Comparing their performances, it can be seen that weighted frequency and the wavelet scale feature perform similarly in the speaker dependent scenario but weighted frequency is better in the speaker independent case. Both outperform spectral centroid. Consequently, only weighted frequency is included in all experiments henceforth.

The source specific features considered are pitch, energy and the LP residue cepstral coefficients. Pitch and energy are complementary and are both used in emotion recognition systems. The LPRCC is an attempt to capture information about the spectral shape of the LP residue, which is an estimate of the vocal excitation. However, explicit glottal – vocal tract separation would be a better approach and is discussed in Chapter 6.

Fractal dimension and phoneme rate cannot be viewed in terms of the source-filter model and consequently can be expected to be complementary to the other features. However, phoneme rate, estimated as described in section 3.2.9, does not appear to contain any emotion specific information. This suggests that estimation method is flawed, or that phoneme rate is not suitable as a feature for a frame based modelling approach. Fractal dimension on the other hand is able to discriminate between the emotions considered here. This is especially interesting since the aspects of speech characterised by the fractal dimension are hypothesised to be caused by non-linear processes that are not modelled by the source-filter model. While non-linear speech production models are not considered in this thesis, these results suggest future directions for research.

3.5 GMM based AER Benchmarking System

The 5-class AER system used to quantify the effectiveness of the features described in this chapter is outlined in this section. Gaussian mixture models were used as the back-end this system. The probability density functions of the features for each emotion were modelled by a GMM (described in section 2.3.2.1), capturing all the statistical information present in them. The LDC corpus (described in section 2.2.3.3) was used in the classification experiments. The data is in the form of short discrete utterances and each utterance was segmented into 20ms frames (unless otherwise mentioned in the description of the features), with 10ms overlap between consecutive frames, prior to feature extraction in these frames. The emotional class models (GMMs) were trained on the feature vectors extracted from all the frames in the training dataset. During testing the likelihood of an utterance belonging to each of the five emotional classes was calculated as the product of the likelihoods (conditional probability of a GMM given a feature vector, refer (2.11)) of the feature vectors (one vector corresponding to each frames) belonging to that model; and the most likely emotion chosen. Since the actual emotional class of each utterance is known, the accuracy of the system can be determined. Figure 3.17 shows an overview of the GMM based AER system.

Speech data from 7 speakers are available in the LDC corpus and both speaker dependent and speaker independent classification tests were performed. In the speaker dependent tests, both training and testing data were from the same speaker. 70% of the utterances from a speaker were used for training the GMMs and the remaining 30% were used for testing. This was repeated 7 times, once for each speaker and the average accuracy of the 7 trials was taken as the overall system accuracy. The speaker independent tests used training and test data from different speakers and were carried out

in a 7-fold cross validation setup. Data from 6 speakers was used for training the back-end and the data from the remaining speaker was used for testing. This was repeated 7 times, using each of the 7 speakers as the test speaker, and the mean accuracy of the 7 trials was taken as the overall system accuracy.

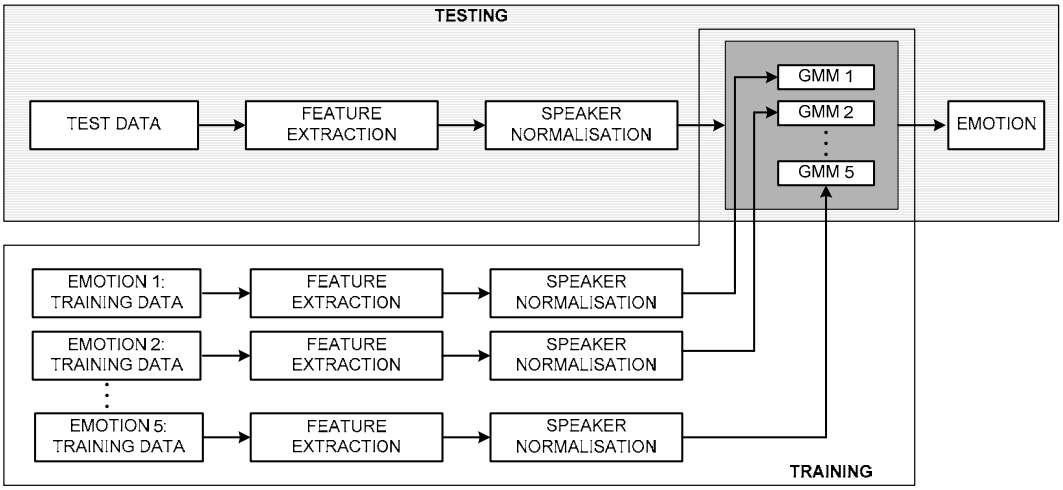

Figure 3.17: Overview of the AER system used in this study

Speaker normalisation was used in the speaker independent tests, to reduce the speaker specific variability in features, and is required when data from multiple speakers are used. Chapter 4 discusses speaker variability and speaker normalisation.

Classification accuracies are reported as overall classification accuracy, i.e. the percentage of test utterances classified correctly, and a table reporting the confusion matrix. The confusion matrix has actual emotions in the rows and target emotions in the columns. Each value is the percentage of utterances belonging to the actual emotion that was classified as the target emotion. i.e., the element v_{ij} of the confusion matrix is the percentage of utterances belonging to emotional class i that were automatically classified as belonging to emotional class j.

It is important to note that pitch (F_0), which is a significant and widely used feature, is defined only for voiced speech and consequently it can be extracted only from frames that contain voiced speech (referred to as *voiced frames*). Thus only voiced frames were used

in training, testing and normalisation in all experiments. This was the case even when pitch was not used as a feature in order to make the systems comparable based on their performance.

Chapter 4

Speaker Variability

The features used in emotion recognition systems are selected to be representative of speech characteristics that vary with the emotional state of the speaker. However, speech also conveys other information including the linguistic content, information about the speaker, etc, and consequently no parameter varies only with the emotional state. The variability in the features used in an AER system that does not contribute towards distinguishing between emotions, usually degrades the performance of the system. The features can exhibit a lot of variability between different speakers in particular. This would not be a problem for a classification system that is trained on data obtained from the target speaker (speaker dependent), but such an expectation is not practical in most cases. This chapter investigates the existence and significance of speaker specific variability. In order to do so, a novel technique for speaker normalisation in AER based on matching the feature distributions for different speakers is proposed. This normalisation technique is then used to compare the speaker variability in different front end configurations. Finally, the chapter also includes a preliminary comparison between the effect of speaker variability and phonetic variability.

4.1 Significance of Speaker Variability

Pitch is a widely used and successful feature in emotion classification problems. Pitch values, however, exhibit a large amount of variation between speakers. Figure 4.1(a-b) shows the probability distributions of the pitch values for two different speakers

expressing no emotion (neutral) and anger, estimated from all utterances from these two speakers present in the LDC corpus. It is clear that while the distributions for neutral and anger are distinct for the speakers, the distributions for speaker 1 are not the same as the distributions for speaker 2, in particular the neutral class in this example. Hence, when the probability distribution of pitch for an emotion is estimated for all speakers, the resultant distribution is multi-modal with a large variance (Figure 4.1c).

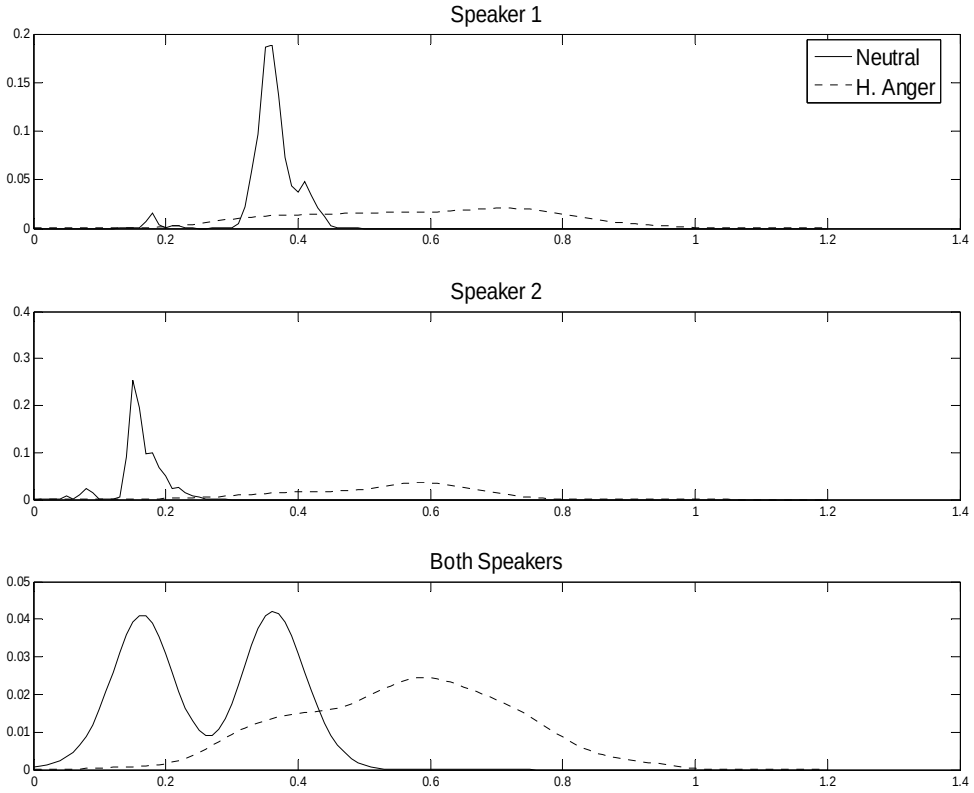

Figure 4.1: Distribution of pitch (a) Speaker 1; (b) Speaker 2; (c) Both speakers together

Although only pitch is illustrated, this is true for all features and is a problem for any speaker independent emotion recognition system. Comparisons of speaker dependent (training and testing data from the same speaker) and speaker independent (training and testing data from different speakers) classification accuracies, for all features, lend support to this observation. Table 4.1 lists the overall classification accuracies obtained by a speaker independent system and a speaker dependent system on the same 5-class classification task. Even though the speaker independent system is trained on a larger

dataset (approx. 6 times larger), the performance of the speaker dependent system is much greater. Given that the tasks and system parameters are identical this difference can be attributed solely to speaker specific variability. A novel speaker normalisation method that makes use of cumulative distribution mapping to match the feature distributions for different speakers is proposed to reduce this variability. This method is also used to investigate the relative speaker variability in different features in section 4.2.

Table 4.1: Comparison of speaker dependent and independent systems (5-class)

Features	Speaker Independent	Speaker Dependent
MFCC	48.7 %	74.8 %
Formant Frequencies (FF)	35.7 %	58.3 %
Reflection Coefficients (RC)	41.8 %	71.2 %
Pitch (P)	36.0 %	51.8 %
Intensity/Energy (E)	22.2 %	25.2 %
Energy Slope (S)	38.9 %	59.0 %
Zero Crossing Rate (Z)	37.6 %	46.8 %
Spectral Centroid (SC)	33.3 %	44.6 %
Phoneme Rate (PhR)	20.4 %	23.0 %
GFCC	46.3 %	72.6 %
LP based Group Delay (GD)	36.0 %	69.8 %
Wavelet Scale Feature (WS)	38.1 %	54.7 %
LPRCC	38.9 %	68.4 %
Frequency Modulation (FM)	40.5 %	64.0 %
Weighted Frequency (WF)	40.7 %	53.2 %
Fractal Dimension (FD)	41.3 %	41.0 %

4.1.1 Cumulative Distribution Mapping

Cumulative distribution mapping is a technique that maps each feature dimension to a predetermined distribution, and was originally suggested as a method to provide robustness against channel mismatch and non-linear noise effects (de la Torre et al. 2002; Pelecanos et al. 2001). Also known as histogram equalisation in image processing literature and feature warping in speech processing literature, it has been used

successfully in speech recognition (de la Torre et al. 2002), speaker verification (Pelecanos et al. 2001) and language identification (Allen et al. 2006). In all three areas it is applied on each utterance (or short segments) based on the assumption the underlying distribution is known (typically Gaussian) and any deviation is due to a distortion that requires normalisation. However, in the proposed method the mapping is estimated from all the data from each speaker and is utilised in a different manner (outlined in section 4.1.2) making no assumptions about the underlying distribution.

Cumulative distribution mapping treats each feature dimension as an independent stream of values, mapping them onto a target distribution (refer to Figure 4.2).

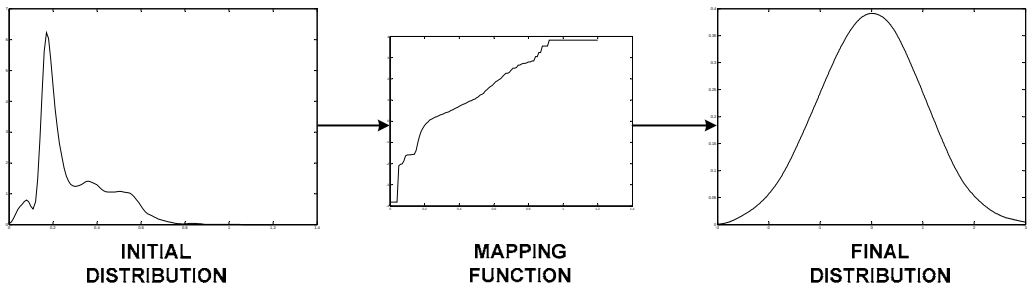

Figure 4.2: Overview of Cumulative Distribution Mapping

Denoting the target distribution as $h(z)$, and the original probability distribution of the feature as $f(y)$, the mapping is defined as

$$\int_{y=-\infty}^{p} f(y)dy = \int_{z=-\infty}^{q} h(z)dz \tag{4.1}$$

where, p is the original feature value and q is the warped feature value.

It is not necessary however to estimate the actual distribution $f(y)$; rather the integrals can be recognised as the cumulative density functions corresponding to the probability distributions.

$$F(p) = \int_{y=-\infty}^{p} f(y)dy \tag{4.2}$$

$$H(q) = \int_{z=-\infty}^{q} h(z)dz \tag{4.3}$$

This reduces (4.1) to

$$F(p) = H(q) \tag{4.4}$$

Since the target distribution, $h(z)$ is known, the corresponding cumulative density function (CDF), $H(x)$, and hence the inverse CDF are also known. Denoting the inverse CDF as $H^{-1}(x)$, the warped feature value is

$$q = H^{-1}\big(F(p)\big) \tag{4.3}$$

Given a large number of feature samples, the value of the CDF corresponding to the original distribution, for any feature value, can be approximated as the ratio of the number of samples lower than that value to the total number of samples. This is accomplished by initially sorting all the feature samples in descending order and indexing them from 1 to N (N is the number of samples). The rank, R, of the feature, p, to be warped is its index after sorting and $N - R$ gives the number of samples lower than it, allowing for the estimation of the cumulative density value.

$$F(p) \approx \frac{N - R}{N} \tag{4.4}$$

This gives the warped value as

$$q = H^{-1}\left(\frac{N - R}{N}\right) \tag{4.5}$$

4.1.2 Proposed Speaker Normalisation

Selecting the target distribution as the standard normal distribution, cumulative distribution mapping is used to map all the features extracted from all the data from each speaker (for all emotions) onto the same region of a new feature space, thereby reducing

any variability introduced by the speakers. In the context of the example pitch distributions shown in Figure 4.1, the overall distribution of the pitch stream (taking into account both emotions) for each speaker is mapped to the standard normal distribution. This preserves the difference between distributions for each emotion for a speaker while normalising the values across speakers.

The distributions estimated from the pitch streams for both speakers after feature warping are shown in Figure 4.3(a-b). It can be seen that the variation between the distributions for both speakers is now much lower. This also results in a reduction in the variance of the overall distribution for each emotion; when estimated from both speakers (refer to Figure 4.3c).

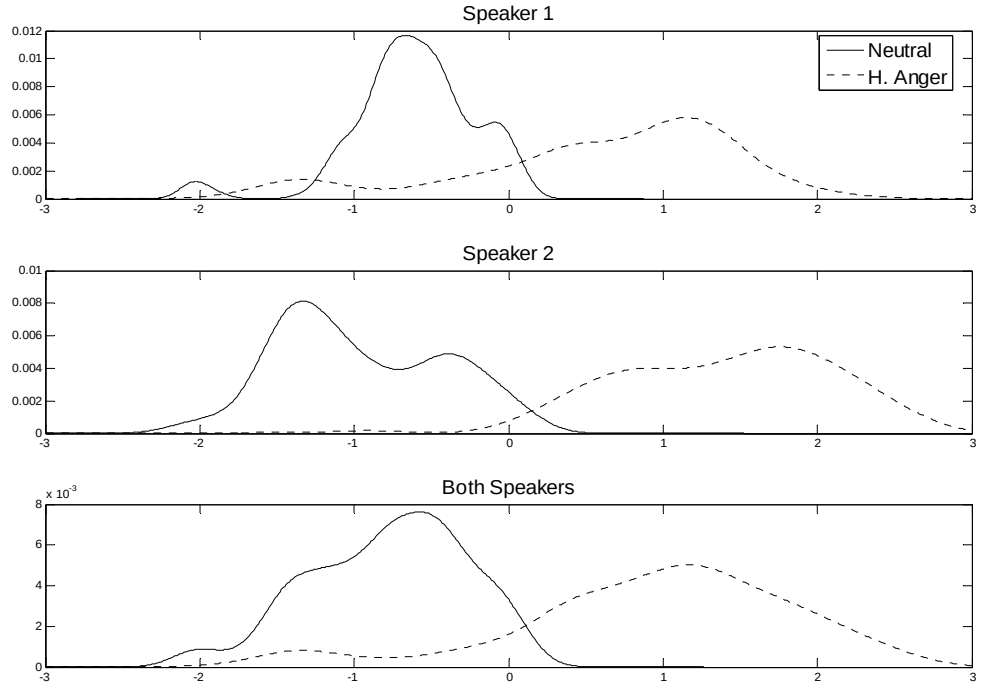

Figure 4.3: Distribution of pitch after normalisation (a) Speaker 1; (b) Speaker 2; (c) Both Speakers

4.1.3 Evaluation

The GMM based automatic emotion recognition (AER) system described in section 3.5 was used to evaluate the performance of this proposed speaker normalisation method. When used in a speaker independent configuration, the training data and the test data are

from different speakers (6 speakers for training and a 7th speaker for testing). Hence, if speaker normalisation is required and if the proposed method is effective, a comparison of the system classification accuracy when using normalisation to the accuracy when normalisation is not used should indicate a significant improvement in system performance. Such a comparison using the GMM based AER system was performed for all the features reported in Chapter 3 and the results are given in Table 4.2.

Table 4.2: Overall classification accuracies for a 5-class GMM based AER system.

Features	Speaker Independent		Speaker Dependent	
	Without Normalisation	With Normalisation	Without Normalisation	With Normalisation
MFCC	48.7 %	49.7 %	74.8 %	73.4 %
Formant Frequencies (FF)	35.7 %	43.7 %	58.3 %	67.6 %
Reflection Coefficients (RC)	41.8 %	48.9 %	71.2 %	74.8 %
Pitch (P)	36.0 %	46.6 %	51.8 %	53.2 %
Intensity/Energy (E)	22.2 %	28.8 %	25.2 %	28.8 %
Energy Slope (S)	38.9 %	43.4 %	59.0 %	57.6 %
Zero Crossing Rate (Z)	37.6 %	47.1 %	46.8 %	49.6 %
Spectral Centroid (SC)	33.3 %	40.2 %	44.6 %	47.5 %
Phoneme Rate (PhR)	20.4 %	20.6 %	23.0 %	23.0 %
GFCC	46.3 %	55.6 %	72.6 %	74.8 %
LP based Group Delay (GD)	36.0 %	42.9 %	69.8 %	71.9 %
Wavelet Scale Feature (WS)	38.1 %	41.8 %	54.7 %	54.7 %
LPRCC	38.9 %	50.0 %	68.4 %	68.4 %
Frequency Modulation (FM)	40.5 %	44.4 %	64.0 %	62.6 %
Weighted Frequency (WF)	40.7 %	47.6 %	53.2 %	57.6 %
Fractal Dimension (FD)	41.3 %	46.3 %	41.0 %	41.0 %

The classification accuracies in Table 4.2 strongly indicate that speaker variability is a problem in emotion recognition and needs to be addressed. They also indicate that the proposed speaker normalisation method is able to reduce this variability and improve the performance of AER systems. The proposed normalisation method improves the classification accuracy of the system for all features, except phoneme rate. However, given the extremely poor phoneme rate accuracy (20% is random separation for a 5 class

problem), this feature can be safely ignored. These results indicate that provided a feature is capable of discriminating between emotional classes, the proposed technique can improve performance by normalising speaker variability.

4.2 Speaker Dependency of Features

Different features may have different levels of speaker dependent and emotion dependent characteristics. This would produce differing performances in speaker dependent (trained on data from target speaker) and speaker independent (training and testing data come from different speakers) systems. Also, in some cases the information contained in a particular feature set could be complementary to the information in another set. This section attempts to compare the speaker variability of such features and determine if some or any of them are complementary.

Based on whether the features are representative of parameters of the speech production model (Section 3.1) or the speech spectrum, they can be categorised as speech production cues or spectral features. Spectral features can be further classified into broad and detailed spectral measures based on the level of spectral detail contained in them. While it has been shown that speaker variability in features significantly lowers the performance of a speaker independent system (section 4.1.2), different features capture different amounts of the speaker's characteristics and consequently not all of them are affected to the same degree. It should be noted that the distinction between spectral features and speech production cues is only a loose way of grouping the features and is not set in stone, particularly when considering that parameters of the speech production model ultimately affect the speech spectrum.

4.2.1 Source Specific Cues (SSC)

Features based on the parameters of the vocal excitation in the context of the source-filter model constitute source specific cues. The source parameters are represented by pitch and energy, each of which is a single dimensional feature. In this section, they are taken together to form a 2 dimensional feature vector that parameterises the vocal excitation. This 2 dimensional feature does not contain any information about the vocal tract.

4.2.2 Detailed Spectral Measures (DSM)

The vocal tract is modelled by an all-pole filter and characterised by its resonant frequencies (formant frequencies). Features that contain detailed information about the spectral characteristics of the vocal tract are referred to as detailed spectral measures (DSM). They are typically high dimensional when compared to broad spectral measures. Mel frequency cepstral coefficients (MFCCs), LP-based group delay (GD), formant frequencies (FF) and reflection coefficients (RC) belong to this category. It should be noted that while the group delay, formant frequencies and reflections coefficients are related to the spectral envelope of the signal and characterise only the vocal tract, MFCCs contain information about both the vocal excitation and the vocal tract.

4.2.3 Broad Spectral Measures (BSM)

These are features derived from the spectrum, but exclude a lot of detail in an attempt to reduce variability that could degrade the performance of the AER system. The broad spectral measures tend to be low dimensional and describe only a part of the spectrum. Energy slope and zero crossing rate are taken together to form a rough estimate of the spectral distribution of energy in the signal. In (Huang et al. 2006), they were proposed as additions to pitch and energy in a speaker independent system. Taken together, they form a 2 dimensional vector. Two other estimates of the spectral energy distribution that could

be considered broad spectral measures are the EMD based weighted frequency (refer to section 3.3.4) and spectral centroid (refer to section 3.2.8). However, given that they are both very similar to each other (both are some form of weighted spectral averages) and that weighted frequency outperforms spectral centroid, only weighted frequency is considered.

4.2.4 Performance Comparison

The GMM based automatic emotion recognition (AER) system described in section 3.5 was used in both speaker dependent and speaker independent contexts to compare the speaker variability inherent in the features described in above. The accuracies obtained are reported in Table 4.3.

Table 4.3: Comparison of emotion classification accuracies for individual features

	Features	Speaker Dependent	Speaker Independent	
			Without Normalisation	With Normalisation
SSC	Pitch + Energy (PE)	51.1 %	37.6 %	46.6 %
DSM	MFCC	74.8 %	48.7 %	49.7 %
	Formant Feature (FF)	58.3 %	35.7 %	43.7 %
	Reflection Coefficients (RC)	71.2 %	41.8 %	48.9 %
	LP based Group Delay (GD)	69.8 %	36.0 %	42.9 %
BSM	Energy Slope + ZCR (SZ)	57.6 %	40.2 %	51.6 %
	Weighted Frequency (WF)	53.2 %	40.7 %	47.6 %

As can be seen from these accuracies, the best performing features for the speaker dependent and independent systems are different. More interestingly, the different groups of features are affected similarly by speaker variability. Pitch and Energy (PE), the only features that are based completely on source characteristics, perform moderately well in both speaker dependent and speaker independent (with normalisation) scenarios and have the smallest difference in accuracies in both scenarios. On the other hand, MFCCs, reflection coefficients and group delay which all describe the vocal tract characteristics in

detail exhibit a large difference between speaker dependent and speaker independent scenarios. The Formant feature, which is also characteristic of the vocal tract, exhibits a similarly large difference between speaker dependent and independent performance. The broad spectral measures, namely the weighted frequency and the energy slope - zero crossing rate features, are also affected in terms of classification accuracy due to speaker variability. However, the difference in the case of the broad spectral measures is much lower that those of the detailed spectral measures (MFCC, group delay and reflection coefficients) or formant information. Moreover, it can be seen that speaker normalisation does not improve the performance of the features that describe the vocal tract characteristics in detail as significantly as it does for the other features.

Since the different groups of features are characteristic of different aspects of the speech production model, certain combinations of features (concatenation of the individual feature vectors to make a larger feature vector in the front end) would be complementary and should lead to improved system performance. Such a comparison of feature combinations is reported in Table 4.4. An exhaustive comparison of all possible combinations is neither feasible nor necessary. Features that describe similar information such as MFCCs, group delay and reflection coefficients (detailed vocal tract characteristics) will not benefit from being combined with each other. However, combining the source specific features (pitch and energy) with vocal tract specific features (e.g. reflection coefficients) can be expected to result in improved performance. Thus combinations of features from the three different categories (source specific cues, detailed spectral measures and broad spectral measures) were used as front-ends and the overall classification accuracies of the systems are reported.

Combinations of MFCCs with other features give the best speaker dependent classification accuracies. However, when compared to the MFCC alone speaker

dependent system (Table 4.3), the improvements are small or non-existent. This is most likely due to the fact that MFCCs model both filter and source characteristics and combining them with other features add little extra information. The other detailed spectral measures (group delay, formant information and reflection coefficients) show some improvement (very small in the case of reflection coefficients) when combined with pitch and energy or with the broad spectral measures (BSM).

Table 4.4: Comparison of emotion classification accuracies for feature combinations

	Features	Speaker Dependent	Speaker Independent	
			Without Normalisation	With Normalisation
SSC + BSM	PE + SZ	61.2 %	34.7 %	52.9 %
	PE + WF	59.7 %	36.2 %	56.4 %
SSC + DSM	PE + MFCC	71.2 %	45.0 %	58.2 %
	PE + FF	57.5 %	36.2 %	57.7 %
	PE + GD	69.1 %	43.1 %	54.2 %
	PE + RC	71.9 %	47.9 %	58.2 %
BSM + DSM	SZ + MFCC	74.1 %	45.0 %	50.8 %
	SZ + FF	60.4 %	38.6 %	48.9 %
	SZ + GD	72.7 %	39.7 %	48.4 %
	SZ + RC	74.8 %	48.2 %	52.7 %
	WF + MFCC	77.0 %	43.7 %	48.9 %
	WF + FF	62.6 %	41.3 %	49.5 %
	WF + GD	71.2 %	42.9 %	47.1 %
	WF + RC	72.7 %	50.0 %	56.1 %

In the case of the speaker independent system, feature combinations of the source specific cues (pitch and energy) with the other features give the best results. Among these, combining detailed spectral measures (DSMs) with pitch and energy appears to be more effective than combining broad spectral measures (BSMs) with pitch and energy. It is also interesting to note that reflection coefficients are the best performing DSMs in a speaker independent scenario, outperforming the others when combined with the same feature (i.e., RC + X outperforms or matches MFCC + X, GD + X and FF + X).

The accuracies of the different features in a five-class emotion classification reported in this section suggests that MFCCs are very discriminative but are also very characteristic of the speaker, and that they do not lend themselves well to speaker normalisation. Since most practical emotion classification systems would need to be speaker independent, MFCCs may not be the front-end of choice, unlike in speech recognition and speaker recognition systems.

4.3 Phonetic and Speaker Variations

This section reports an experiment performed to determine if some phonemes are more conducive to emotion classification than others. In order to achieve this, an emotion classifier was setup and the independent classification accuracies for different phonemes were determined. It was expected that if certain phonemes expressed the emotion being conveyed better than others, the classification accuracies for those phonemes would be correspondingly higher than those of other phonemes. The effect of speaker variability in this context was also investigated.

4.3.1 Phoneme Recognition

In order to examine the effect of phonetic content on classifier performance it was essential to determine the phoneme associated with every frame of data. The phoneme recogniser developed at the Faculty of Information Technology, Brno University of Technology (Schwarz et al. 2006) was applied to generate phonetic labels from the data. The dominant phoneme in each frame (the phoneme with the longest duration in the frame when more than one was present) according to the labels was then associated with the frame, as seen in Figure 4.4.

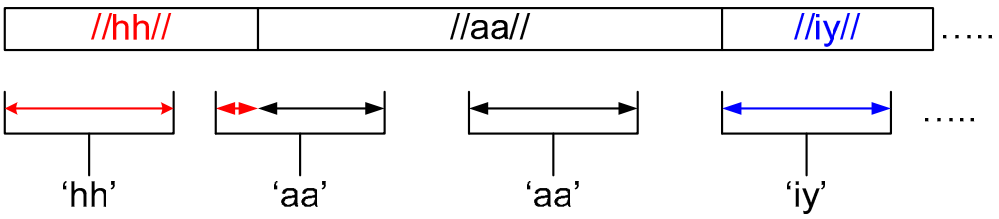

Figure 4.4: Frame level phonetic labelling

The phoneme set consists of 39 phonemes, as described in (Schwarz et al. 2006). However, since reliable pitch estimation is rarely possible from stops, affricates and fricatives, they were all combined as a single phoneme group (the phonemes *b, d, g, p, t, k, dx, jh, ch, s, sh, z, f, th, v,* and *dh* were grouped together and labelled as *fr*). Also, frames labelled as silences or pauses were not included in the experiments. This gave a total of 23 classes. Informal tests on the TIMIT database indicate the phone recogniser had an accuracy of about 74% for these 23 classes.

4.3.2 Classification System

Since the aim of the experiment was to study the difference between different phonemes with respect to automatic emotion recognition, a decision about the emotion could not be made for every phrase, as in all the previously reported results, since each phrase would contain many phonemes. However, since the AER systems (both speaker dependent and independent systems) described in Section 3.5 compute likelihood scores for feature vector from every frame prior to making a decision about a phrase, these scores can be used to classify each frame (instead of each phrase) as belonging to one of the five emotional classes. Since each frame is also associated with a phoneme (Section 4.3.1), this allowed for the study of phoneme specific system performance and such a system was used in this experiment and the classification accuracies for all the phoneme classes listed in Table 4.5 were determined for speaker dependent and speaker independent cases.

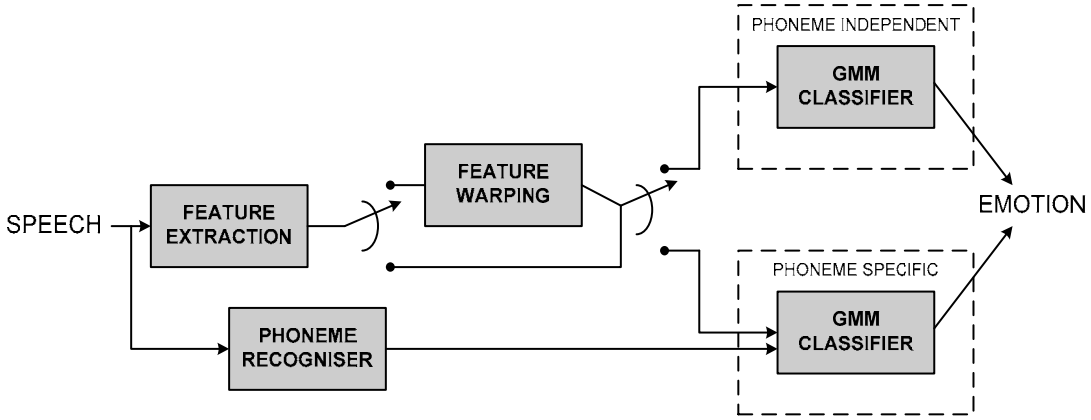

Figure 4.5: Phonetic and speaker variation test system overview

Table 4.5: Number of Test Frames in each phonetic class

Phonemes	Number of Frames		Phonemes	Number of Frames	
	Speaker Dependent	Speaker Independent		Speaker Dependent	Speaker Independent
fr	556	1973	***ey***	190	671
m	226	897	***ae***	383	983
n	678	2558	***aa***	106	296
ng	5	27	***aw***	258	610
l	623	2178	***ay***	596	2413
r	56	184	***ah***	372	1420
w	4	46	***oy***	0	0
y	3	3	***ow***	18	119
hh	344	1063	***uh***	5	13
iy	899	2772	***uw***	263	881
ih	1131	3791	***er***	140	888
eh	256	888			

Since the speech spectrum determines the sound (phoneme) being uttered and is in turn determined by the shape of the vocal tract, features that are characteristic of the vocal tract can be expected to have phoneme specific information. Using such features in this experiment could bias the results in favour of phonemes which were more closely clustered in that feature space. Hence a feature vector comprising of pitch, energy and weighted frequency, the best performing feature combination in a speaker independent scenario (refer to Table 4.4), that did not contain any detailed vocal tract information was chosen as the front-end for this experiment.

Table 4.6: Phonetic accuracies for speaker dependent (SD) and speaker indepdent (SI) systems

	Accuracy					Accuracy			
	With Normalisation		Without Normalisation			With Normalisation		Without Normalisation	
	SD	SI	SD	SI		SD	SI	SD	SI
fr	51.3 %	50.1 %	48.6 %	39.2 %	*ey*	46.3 %	42.0 %	39.0 %	26.5 %
m	52.2 %	34.5 %	54.4 %	30.4 %	*ae*	46.2 %	47.4 %	46.2 %	20.8 %
n	39.8 %	38.9 %	38.8 %	27.0 %	*aa*	67.9 %	60.8 %	61.3 %	51.4 %
ng	0 %	29.6 %	0 %	33.3 %	*aw*	42.6 %	56.1 %	44.2 %	32.1 %
l	45.6 %	39.5 %	44.8 %	26.3 %	*ay*	56.5 %	53.5 %	50.7 %	38.3 %
r	69.6 %	46.7 %	33.9 %	34.2 %	*ah*	47.8 %	42.5 %	48.1 %	30.6 %
w	100 %	63.0 %	100 %	32.6 %	*oy*	-	-	-	-
y	100 %	100 %	100 %	100 %	*ow*	22.2 %	53.8 %	0 %	24.4 %
hh	36.6 %	31.7 %	33.4 %	27.2 %	*uh*	100 %	100 %	20.0 %	38.5 %
iy	50.1 %	41.5 %	46.4 %	32.0 %	*uw*	33.1 %	40.3 %	35.7 %	30.0 %
ih	49.6 %	41.6 %	48.5 %	28.8 %	*er*	32.9 %	34.9 %	37.9 %	24.9 %
eh	41.8 %	45.6 %	42.6 %	36.0 %	***Overall***	**47.5 %**	**43.8 %**	**45.8 %**	**31.6 %**

From Table 4.5, it can also be seen that the rate of occurrence of some phonemes is higher than that of others, particularly semi-vowels and vowels. This is because better pitch estimates can be obtained from these phonemes than the others and only frames with pitch estimates were used in the experiments. Also, the accuracies for phonetic classes with very few test frames convey little or no useful information since they are easily affected by a few frames being misclassified (phonetic classes *//ng//*, *//w//*, *//y//*, *//oy//*, *//ow//*, *//uh//* can be safely ignored). Their low rates of occurrence also mean their contribution to the overall accuracy is negligible.

From Table 4.6 it can be seen that feature warping has very little effect on a speaker-dependent system, as expected. However in the speaker-independent case, feature warping plays a very significant role. This suggests that variations in the features between different speakers are quite large and much better modelling can be achieved when some sort of normalisation is used to reduce this variability.

The Gaussian mixture models used for each emotional class in all of the abovementioned experiments were trained on data from all phonetic classes. It might be

argued that better modelling may be achieved if a separate GMM was trained for every phonetic class for every emotion. During testing, since every test frame is associated with a particular phonetic class, likelihood estimation and consequently classification is performed only over the five GMMs associated with the five emotions for that phonetic class. Such an experiment was performed for the speaker-independent case (there was insufficient training data to do this in a speaker-dependent manner) and the results are given in Table 4.7.

Table 4.7: Phonetic accuracies for a speaker independent system using phoneme specific GMMs

Phonemes	**Accuracy**		**Phonemes**	**Accuracy**	
	With Normalisation	**Without Normalisation**		**With Normalisation**	**Without Normalisation**
fr	50.8 %	45.8 %	*ey*	46.4 %	21.8 %
m	30.2 %	38.0 %	*ae*	45.0 %	37.1 %
n	39.2 %	30.3 %	*aa*	78.0 %	59.1 %
ng	29.6 %	33.3 %	*aw*	46.4 %	32.1 %
l	43.6 %	29.9 %	*ay*	46.3 %	33.2 %
r	35.9 %	43.5 %	*ah*	40.5 %	37.1 %
w	50.0 %	39.1 %	*oy*	-	-
y	100 %	100 %	*ow*	63.9 %	33.6 %
hh	28.6 %	25.4 %	*uh*	100 %	38.5 %
iy	44.2 %	37.5 %	*uw*	26.7 %	23.6 %
ih	39.4 %	30.2 %	*er*	45.8 %	39.5 %
eh	33.2 %	31.1 %	***Overall***	**42.2 %**	**33.6 %**

Comparing the accuracies of the systems using phoneme-specific GMMs with those that use phoneme independent emotion models, the difference appears to be very small. This tends to suggest that the phoneme-specific models are very similar to the phoneme-independent models, indicating that for these features phonetic variability is very small and certainly much less significant than speaker variability. It is important to note that the features used in this experiment were chosen on the basis that they did not characterise the detailed spectral shape and consequently the phonetic content of speech. Hence this

made it difficult to determine whether the similarity of the phoneme-specific GMMs to the phoneme-independent GMMs is because of the lack of phoneme-specific information in the features or because the information being modelled by the emotion models is different from that modelled by phoneme recognisers. In essence, are these experiments biased towards this similarity, or is the similarity an inherent property of the emotion models? To clarify this, speaker independent emotion classification was performed with phoneme-independent and phoneme-specific GMMs using a MFCC based front end.

Table 4.8: Phonetic accuracies for a MFCC based Speaker-Independent system (with normalisation)

Phonemes	Accuracy		Phonemes	Accuracy	
	Phoneme Independent GMMs	Phoneme Specific GMMs		Phoneme Independent GMMs	Phoneme Specific GMMs
fr	40.6 %	44.1 %	*ey*	35.9 %	38.3 %
m	32.9 %	20.0 %	*ae*	46.0 %	37.7 %
n	28.1 %	35.2 %	*aa*	68.6 %	77.0 %
ng	25.9 %	25.9 %	*aw*	51.0 %	43.0 %
l	34.2 %	36.0 %	*ay*	43.5 %	48.0 %
r	29.4 %	32.6 %	*ah*	38.6 %	40.1 %
w	37.0 %	37.0 %	*oy*	-	-
y	100 %	100 %	*ow*	47.1 %	61.3 %
hh	34.0 %	40.5 %	*uh*	92.3 %	92.3 %
iy	35.6 %	38.1 %	*uw*	43.0 %	35.6 %
ih	35.2 %	36.9 %	*er*	30.4 %	28.5 %
eh	34.0 %	40.5 %	***Overall***	**37.4 %**	**38.8 %**

The classification accuracies of the MFCC based systems are reported in Table 4.8 and from these accuracies it can be observed that once again there is very little difference between phoneme-specific and phoneme-independent emotion models. This lends support to the argument that even when phoneme-specific information is present in the features, they are not modelled by the emotion models; in turn supporting the observation that speaker variability is a more significant problem in emotion modelling than phonetic

variation. A summary of the overall accuracies of the different systems based on the pitch, energy and weighted frequency front-end is shown in Table 4.9.

Table 4.9: Summary of overall accuracies

	Speaker Dependent		**Speaker Independent**	
	Without Norm.	*With Norm.*	*Without Norm.*	*With Norm.*
Phoneme Independent GMMs	45.8 %	47.5 %	31.6 %	43.8 %
Phoneme Dependent GMMs	-	-	33.6 %	42.2 %

4.4 Summary

This chapter initially presented a technique that is novel in the context of AER to reduce the variance in data that arises due to differences in speaker characteristics, in order to improve the performance of a speaker independent emotion classification system. The proposed method involved the use of cumulative distribution mapping to transform the data from each speaker such that they are all mapped to the same distribution. This resulted in the data retaining their separation in the feature space due to emotional classes but not due to speaker specificity. Experimental results (Table 4.2) indicated that the proposed method improves the performance of the speaker independent system for all features.

In the next section, the proposed normalisation technique was used to study how different features are affected by speaker variability. The performance of features and feature combinations from three broad categories of feature types were compared (refer Table 4.3 and Table 4.4). These accuracies revealed that MFCCs are very discriminative but are also very characteristic of the speaker. Moreover, the results indicated that source specific features such as pitch and energy lend themselves more to normalisation than

detailed spectral features such as MFCCs and group delay. The results also tend to suggest that MFCCs may not be the front-end of choice in a speaker independent system.

This chapter then examined whether features extracted from speech corresponding to certain phonemes are more discriminative of emotions than features extracted from other phonemes in both speaker-dependent and speaker-independent systems. The classification accuracies (refer Table 4.6) indicated that this is the case, and differences between emotions are better conveyed by some phonemes than others. However, the accuracies of the emotion models were affected to a larger extent by differences between speakers than they were by difference between phonemes. For example, the high classification accuracies for frames associated with the phoneme //aa//, indicates that the classifier is able to make better decisions from features corresponding to //aa//; but classifiers trained on speech frames corresponding to the phoneme //aa// perform similarly to classifiers trained on all frames corresponding to all phonemes.

Chapter 5

Static Classification Approaches

The automatic emotion recognition system described in section 3.5 is based on a Gaussian mixture model (GMM) back-end. This approach is useful for comparing different front-ends and for studying the effect of variability unrelated to emotion. However, the fact that GMMs are only one of many possible back-ends prompts the question, how do they compare to other classification approaches? Tools such as *N*-grams, though not explicit classifiers, are used to model temporal patterns while some classifiers such as hidden Markov models are used to model both statistical and temporal patterns. Classifiers that explicitly model temporal patterns are usually termed dynamic classifiers (e.g. HMMs, refer to section 6.2.4 for a description of a system that uses HMMs to model pitch contours), while those that model only statistical patterns are termed static classifiers. Section 5.1 compares a few commonly used static classifiers in the context of emotion recognition to determine if any of them are significantly better than the others. Other questions about the back-end that arise are:

- Is a single stage classifier sufficient, or would pre-classification improve recognition rates?
- Given that features are extracted for short frames, should the back-end model these features directly or should utterance (turn) level statistics be computed from the features and these statistics modelled?

The experiments performed to try and ascertain the answers to these questions are reported in this chapter. While the limited size of the database precludes high degrees of

certainty, the results obtained can provide some indications as to which approaches may offer more promise compared to the others.

5.1 Comparison of Static Back-Ends

A number of different classifiers have been used in various speech processing applications. Some of them such as Gaussian mixture models and support vector machines are used to statistically model the distribution of the features. It should be noted that while static classifiers model only the statistical patterns in the feature space, the features themselves could be chosen to be representative of temporal patterns in speech and thus enable the static classifiers to model temporal patterns indirectly. The static classifiers compared in this section are

- Gaussian mixture models (GMM) - (Reynolds et al. 1995)
- Probabilistic neural networks (PNN) - (Specht 1988)
- Support vector machines (SVM) - (Vapnik 2000)

In the experiments reported in this thesis where Gaussian mixture models were employed, they were implemented using HTK (Young et al. 1995) and unless mentioned otherwise, all GMMs utilised 4 mixtures. Support vector machines were implemented using the SVM^{light} toolkit (Joachims 2003) and utilised a radial basis kernel (with parameter, $\gamma = 0.015$). The implementation from the neural network toolbox of MATLAB was used for probabilistic neural networks with the spread of the radial basis functions set at 0.15.

Since these classifiers require different techniques for combining frame level decisions/scores to obtain an utterance (turn) level decision, and since the aim is to compare the capabilities of these classifiers to model the feature space, only frame level

classification accuracies, as in section 4.3, are reported. Also as in section 4.3, pitch, energy and weighted frequency are used in the front-end. The classification accuracies obtained when using these three back-ends in a speaker independent scenario are reported in the following tables.

Table 5.1: Confusion matrix for **GMM** based AER system using P + E + WF

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**47.2 %**	0.6 %	23.3 %	5.6 %	23.3%
Anger	1.7 %	**70.5 %**	5.5 %	20.3 %	2.0 %
Sad	27.0 %	2.1 %	**35.1 %**	18.0 %	17.9 %
Happy	6.8 %	25.4 %	21.9 %	**39.3 %**	6.6 %
Bored	36.4 %	1.8 %	22.1 %	10.9 %	**28.8 %**
Overall Accuracy = 43.1 %					

Table 5.2: Confusion matrix for **PNN** based AER system using P + E + WF

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**14.0 %**	0.7 %	7.8 %	7.1 %	70.4%
Anger	6.2 %	**72.5 %**	1.9 %	21.3 %	4.0 %
Sad	6.0 %	3.4 %	**12.3 %**	21.7 %	56.6 %
Happy	1.4 %	27.4 %	8.8 %	**42.4 %**	20.1 %
Bored	7.3 %	1.9 %	14.1 %	14.5 %	**62.3 %**
Overall Accuracy = 44.1 %					

Table 5.3: Confusion matrix for **SVM** based AER system using P + E + WF

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**13.3 %**	1.9 %	11.3 %	4.5 %	69.0 %
Anger	1.2 %	**87.9 %**	4.2 %	3.4 %	3.3 %
Sad	7.2 %	8.7 %	**29.0 %**	11.3 %	52.9 %
Happy	4.9 %	51.1 %	15.6 %	**12.9 %**	15.6 %
Bored	9.6 %	7.5 %	19.2 %	6.7 %	**56.9 %**
Overall Accuracy = 40.5 %					

These overall accuracies of all three systems are very close to each other and do not shed much light on the differences. The confusion matrices on the other hand reveal some

differences between the performances of these classifiers, while some trends are common to all three cases. For instance, anger and happiness are confused with each other far more often than with any other emotions in all three cases. Also, neutral speech is not recognised very well by the PNN or the SVM. It should be noted that the classifiers can distinguish between classes only if such a distinction is possible in the feature space and trends that are common to all three cases could be a result of the features chosen for the front end. Trends that are consequences of the choice of front-end should be ignored in a comparison of back-ends.

In order to help distinguish back-end specific trends from feature related trends, the classification experiments were repeated with a Mel frequency cepstral coefficients (MFCC) based front-end. The accuracies obtained when using MFCCs (which characterise the speech spectrum in detail) instead of pitch, energy and weighted frequency (which lack detailed spectral characterisation) are given in Table 5.4, Table 5.5 and Table 5.6. Given that the features are representative of different aspects of speech, the front-end specific trends can reasonably be expected to be different in both cases.

Since two front-ends modelling different aspects of speech are used, trends common to both can reasonably be attributed to the back-end. This enables a comparison between the three back-ends and the following observations can be made:

- The overall accuracies for all three back-ends are close to each other
- The individual emotion specific classification accuracies vary and GMMs have the best all round performance with no significantly low accuracy for any of the classes
- Happiness is often confused with anger by all three system, and for both front-ends. Given that humans very rarely exhibited this confusion, the confusion in the

automatic classification systems indicates the distinction between the two emotions are not sufficiently well modelled at either the feature level or the classifier level (or both).

- Both SVM and PNN appear to be biased towards boredom, giving a high accuracy for that emotion, at the cost of misclassifying neutral and sadness as boredom in many instances. The reason for this bias is not clear, but it could be a result of insufficient training data.

Table 5.4: Confusion matrix for **GMM** based AER system using MFCC

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**43.2 %**	0.4 %	22.7 %	7.7 %	16.0%
Anger	3.7 %	**60.7 %**	6.8 %	25.3 %	3.6 %
Sad	26.8 %	4.2 %	**27.5 %**	13.7 %	27.8 %
Happy	14.7 %	23.5 %	17.5 %	**29.6 %**	14.7 %
Bored	30.9 %	2.3 %	25.4 %	12.6 %	**28.8 %**
Overall Accuracy = 37.1 %					

Table 5.5: Confusion matrix for **PNN** based AER system using MFCC

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**30.7 %**	1.1 %	18.4 %	12.5 %	37.3%
Anger	2.8 %	**49.1 %**	7.9 %	32.2 %	8.1 %
Sad	13.7 %	4.8 %	**24.3 %**	19.7 %	37.5 %
Happy	9.2 %	21.1 %	16.1 %	**34.5 %**	19.2 %
Bored	18.7 %	3.2 %	25.8 %	15.8 %	**36.6 %**
Overall Accuracy = 35.6 %					

Table 5.6: Confusion matrix for **SVM** based AER system using MFCC

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**29.4 %**	1.8 %	9.3 %	10.0 %	49.6%
Anger	1.1 %	**75.0 %**	2.7 %	16.3 %	5.0 %
Sad	9.6 %	8.6 %	**13.0 %**	16.4 %	52.5 %
Happy	7.0 %	37.6 %	6.7 %	**26.2 %**	22.5 %
Bored	17.1 %	4.3 %	16.1 %	13.1 %	**49.4 %**
Overall Accuracy = 39.9 %					

These trends suggest that a back-end based on Gaussian mixture models is in fact a good choice. The lack of any obvious bias in GMMs towards any of the classes simplifies the interpretation of any classification results, since it eliminates the necessity to compensate for any back-end specific trends.

5.2 Pre-Classification

The training of static classifiers involves separation of the feature space into distinct regions and associating them to the different classes. This is done based on the distribution of the training data in the feature space. Classification then reduces to a problem of associating a feature vector to one of the regions. Ideally, every region in the feature space would correspond to a particular class. The speaker normalisation technique proposed in section 4.1.2 is an attempt to reduce the number and size of clusters caused by speaker specific variability while retaining emotion specific clusters. Another way to address this problem is to perform some sort of pre-classification to identify and separate large clusters not related to emotions and then perform emotion recognition within these clusters.

The classification system used in the experiment reported in section 4.3, using phoneme specific GMMs, in fact performs this kind of pre-classification. It initially separates phonetic clusters and then performs emotion classification within these clusters. However, the phonetic clusters imposed on the feature space are based on knowledge of the linguistic structure of speech and not estimated from the distribution of the features. This may be the reason the classification accuracy of the system did not improve.

An unsupervised clustering algorithm can instead identify data clusters in the feature space without the imposition of any prior assumptions. Once identified, all features can be pre-classified into one of these clusters. Cluster specific emotion modelling and

classification can then be performed. Self-organising maps (SOM), also referred to as Kohonen maps, can be used to 'learn' the structure of the distribution of data in the feature space and hence identify clusters (Kohonen 1982; Kohonen 1997; Kohonen 1998).

Such an experiment was performed, and cluster specific and cluster independent GMM based emotion models were trained in a manner similar to the experiment reported in section 4.3. A comparison of cluster specific and cluster independent classification accuracies however revealed that cluster specific modelling resulted in no significant performance gains in either the AER system that used pitch, energy and weighted frequency (P + E + WF) or the AER system that used MFCCs as its features. The details of this experiment along with classification accuracies are reported in Appendix C.

5.3 Frame based vs. Turn based Static Modelling

The AER system described in section 3.5 uses frame based features and a Gaussian mixture model (GMM) based back-end to model the distributions of these features for the different emotional classes. An alternative approach, that is very prevalent, is to estimate statistics from the set of features extracted from all the frames in each phrase (turn) and to model these statistics instead of the feature values (refer to Figure 5.1). These two approaches are termed '*frame based*' and '*turn based*' respectively (Vlasenko et al. 2007). The turn based static modelling approach is popular since the high level of abstraction of the features results in information reduction, which avoids phonetic over-modelling that may occur in dynamic modelling (Schuller et al. 2003; Vlasenko et al. 2007). The frame based approach used in the system described in section 3.5 also ignores temporal patterns while modelling the probability distribution of the features (using GMMs) which should implicitly capture the statistics that are computed in the turn based

approach (Huang et al. 2006). The statistics utilised in the experiments reported in this section are:

- Mean
- Maximum
- Minimum
- Standard Deviation
- Median
- 1st Quartile
- 3rd Quartile
- Skewness
- Kurtosis

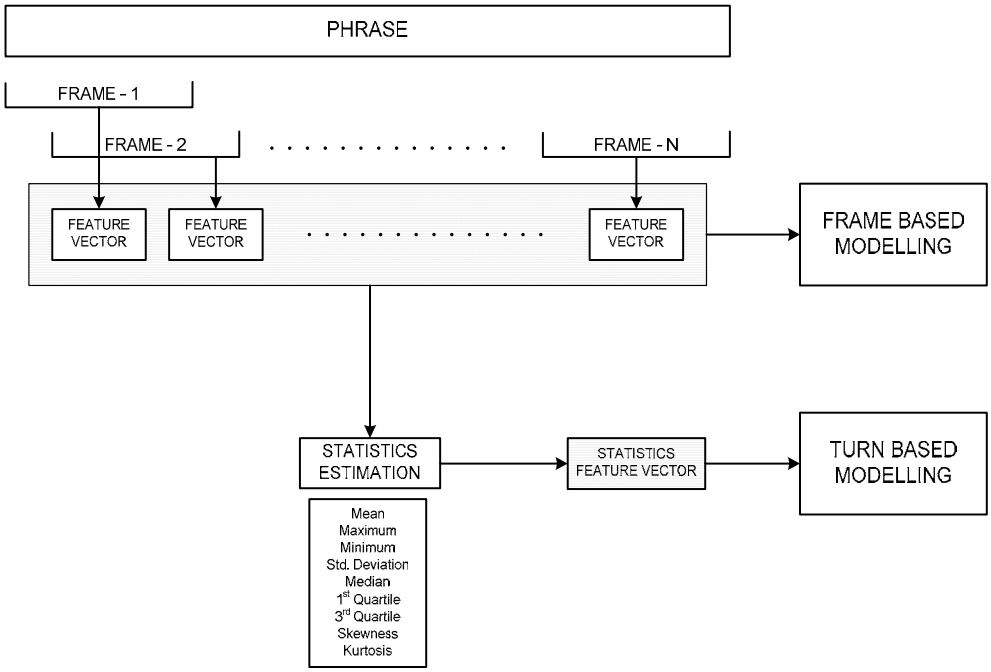

Figure 5.1: Frame and Turn based modelling approaches

It should be noted that both approaches are types of *static* modelling, where only statistics of the features are modelled, as opposed to *dynamic* modelling, where temporal patterns are modelled as well. In order to compare the two approaches, turn based systems with two front-ends - one with pitch, energy and weighted frequency and the other with MFCCs - were set up where:

- Phrases were divided into frames and a feature vector extracted from each frame

- Statistics of features from all frames in each phrase were computed to form a new vector, thus giving one statistics vector per phrase.
- This system was used in a speaker independent scenario (the size of the database was too small to train a speaker dependent setup) and the training and test datasets were setup as described in section 3.5.
- The speaker normalisation technique proposed in section 4.1.2 was applied prior to estimation of the statistics.

The performances of these systems were compared to those of two frame based systems (as described in section 3.5) using the same front-ends. Table 5.7 and Table 5.8 report the accuracies of the pitch, energy and weighted frequency based systems.

Table 5.7: Confusion matrix for **TURN** based AER system using P + E + WF

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**43.4 %**	0 %	15.1 %	3.8 %	37.7%
Anger	0 %	**73.0 %**	1.4 %	21.6 %	2.7 %
Sad	8.1 %	0 %	**31.1 %**	21.6 %	39.2 %
Happy	2.4 %	17.7 %	8.2 %	**63.5 %**	8.2 %
Bored	18.5 %	2.2 %	28.3 %	8.7 %	**42.4 %**
Overall Accuracy = 51.3%					

Table 5.8: Confusion matrix for **FRAME** based AER system using P + E + WF

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**66.0 %**	0 %	17.0 %	0 %	17.0%
Anger	0 %	**79.7 %**	1.4 %	18.9 %	0 %
Sad	23.0 %	0 %	**48.7 %**	13.5 %	15.9 %
Happy	1.2 %	16.5 %	15.3 %	**65.9 %**	1.2 %
Bored	37.0 %	0 %	27.2 %	6.5 %	**29.4 %**
Overall Accuracy = 56.4%					

The accuracies obtained using the MFCC based systems are reported in Table 5.9 and Table 5.10.

Table 5.9: Confusion matrix for **TURN** based AER system using MFCC

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**24.5 %**	0 %	18.9 %	7.6 %	49.1%
Anger	0 %	**79.7 %**	1.4 %	18.9 %	0 %
Sad	2.7 %	0 %	**51.4 %**	17.6 %	28.4 %
Happy	0 %	18.8 %	11.8 %	**57.7 %**	11.8 %
Bored	6.5 %	0 %	22.8 %	21.7 %	**48.9 %**
Overall Accuracy = 54.0%					

Table 5.10: Confusion matrix for **FRAME** based AER system using MFCC

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**45.3 %**	0 %	26.4 %	0 %	28.3%
Anger	0 %	**73.1 %**	1.4 %	25.7 %	0 %
Sad	16.2 %	0 %	**35.1 %**	10.8 %	37.8 %
Happy	3.5 %	17.7 %	11.8 %	**54.1 %**	12.9 %
Bored	18.5 %	0 %	29.4 %	14.1 %	**38.0 %**
Overall Accuracy = 48.9%					

The higher level of abstraction of the features in the turn based approach can be expected to ignore phonetic variability to a larger extent at the cost of higher information loss. Consequently the turn based approach can be expected to perform well when the front-end captures a significant amount of phonetic detail. Of the two front-ends used in the experiment reported here, the MFCC based front-end captures vocal tract specific information and will be more dependent on the phonetic structure of speech and the turn based approach should be more suited to it. From the reported accuracies it can be seen that the turn based approach is indeed better than the frame based on for MFCC, while the frame based approach is superior when using pitch, energy and weighted frequency which are representative of the phonetic content to a much smaller extent.

Given that the turn based approach is better when the features are phoneme dependent, the extensive use of MFCC as features explains the prevalence of the turn based approach in emotion recognition literature. However, as indicated by the

classification accuracies in Table 5.7 and Table 5.8, the turn based approach may not necessarily be the best one. It should be noted that the number of mixtures in the Gaussian mixture models were held constant to facilitate direct comparison of the features and increasing the number of mixtures does result in a small improvement in the performance of the frame based system that makes use of MFCCs. However this improvement is small and the accuracy of the best performing system was less than that of the frame based system using pitch, energy and weighted frequency.

5.4 Summary

This chapter sought to address three issues, namely, the choice of classifier, the use of a pre-classifier and the issue of turn vs. frame based classification. In the first section, Gaussian mixture models, probabilistic neural networks and support vector machines, three commonly used static classifiers, were compared in terms of classification accuracy. The overall classification accuracies of AER systems based on all three were also similar. However, class (emotion) specific recognition rates indicated a bias towards one class (boredom) in the PNN and SVM based system, suggesting that GMM based back-end was the best option of the three.

The next section investigated the potential for multi-stage classification. Specifically, it investigated the use of a pre-classifier to compensate for non-emotion specific clustering of data in the feature space. Unlike the phoneme specific classification system (section 4.3.2) and the speaker dependent AER system, both of which can be considered pre-classification, this section investigated pre-classification based on self-learning without any assumptions of a prior framework (e.g. phonetic structure or speaker specific characterisation of speech). A comparison of a system that made use of pre-classification to one that did not in terms of emotion recognition rates indicated that when low

dimensional features were used, there was no advantage in performing pre-classification. When a high dimensional feature (MFCC) was used, pre-classification conferred a small advantage. However this advantage was too small to draw any conclusions one way or the other.

The final question addressed in this chapter was that of comparing a turn (utterance) based approach to a frame based approach. The turn based approach involved estimation of statistical parameters pertaining to the distribution of features extracted from all the frames in a turn (utterance) and then using the back-end to model a feature space comprising of these statistics. The frame based approach on the other hand involved using the back-end to model the feature space comprising of the features extracted from the frames directly and allowing the back-end to model the statistics implicitly. A comparison of recognition rates obtained using both approaches once again indicated that the dimensionality of the features played a role, with the higher level of abstraction of the turn based approach suiting high dimensional features.

Chapter 6

Speech Parameterisation for Emotion Recognition

The source-filter model described in Section 3.1 is widely used in speech processing literature. It models the vocal apparatus as a linear system excited by a series of glottal pulses, often approximated as the response of a second order low pass system to an impulse train, and has been particularly successful in describing voiced speech. The automatic emotion recognition (AER) systems described in this thesis all make use of only voiced speech and the source-filter model plays an integral role in understanding them. All the features explored in this thesis, with the exception of fractal dimension, are interpreted in the light of the source-filter model (see Chapter 3). Briefly, the characteristics of the vocal apparatus described by this model that are captured by the commonly used features for AER systems are

- Fundamental frequency – described by the period of the impulse train
- Vocal Intensity – described by the amplitude of the impulses
- Shape of the vocal tract – Spectral properties/coefficients of the vocal tract filter model

Two shortcomings common to the features that characterise the various aspects of this model are that (i) they do not contain information pertaining to the shape of the glottal pulses (since the glottal model is always a two pole low pass system with both poles at

unity); and (ii) they do not characterise the temporal variations of the model parameters. This chapter explores:

- The use of the spectral representation of glottal source models, thereby increasing the flexibility of the source-filter model and allowing for explicit seperation of the effects of the glottal source and vocal tract on speech giving rise to a 3-part source-filter model. Also, in contrast to the most other investigations of an explicit glottal model, spectral parameters are used instead of temporal parameters.
- The use of temporal patterns of the model parameters, which adds a temporal dimension to the model. These temporal patterns are longer term patterns than the ones typically characterised by delta parameters (which are commonly utilised).
- The use of discrete contours of model parameters as features in an AER system. Thus allowing for an investigation of the significance of these temporal patterns in the context of AER. A comparison of the performance of an AER system using model parameter contours to that of one modelling the probability distributions of the parameter values would give an indication of this significance.

6.1 The Glottal Source – Spectral Extension

As previously mentioned in section 3.1, voiced speech is modelled as the response of a LTI system modelling the vocal tract to a periodic excitation after it passes through a lip radiation model. The glottal flow is further modelled as the response of a filter to a periodic train of impulses (with the impulses separated by the fundamental period); where the impulse response of the filter is the shape of a single glottal flow pulse. While it is acknowledged that the magnitude response of the glottal filter is low-pass, approximating it by $G(z) = 1/(1 - z^{-1})^2$, as it is commonly done, is equivalent to the assumption that

the shape of the glottal pulse is always the same. As a result, the effects of changes to this shape are not taken into account.

This shape has been associated with certain characteristics of speech that are subsumed under the cover term '*voice quality*' (Childers et al. 1991). While there is no generally accepted definition of voice quality, the term has been used to refer to the auditory impressions of the listener that are not accounted for by measurable parameters. For instance, voice types such as hoarse, harsh and breathy are considered voice qualities. Thus, the incorporation of a more detailed model of the glottal source is required for improved speech analysis. The importance of glottal flow modelling in parametric synthesis of natural sounding speech has been well established (Cabral et al. 2007; Childers et al. 1991). The modification of glottal voice quality factors has been shown to be important for the synthesis of emotional (expressive) speech in (d'Alessandro et al. 2003). While not commonly used, glottal parameters have also been used in an emotion classification framework (Rui et al. 2009) and were shown to be useful in distinguishing between emotional category pairs that had statistically similar pitch values.

Often instead of the glottal flow, its derivative is modelled. The reason for this is twofold:

- The lip radiation model can be approximated by a differentiator as in the traditional source-filter model and using the glottal-flow derivative automatically takes lip radiation into account.
- Some aspects of the shape of the glottal-flow derivative can be visible from the acoustic speech waveform (e.g. peak of the glottal flow derivative may be visible in the speech waveform).

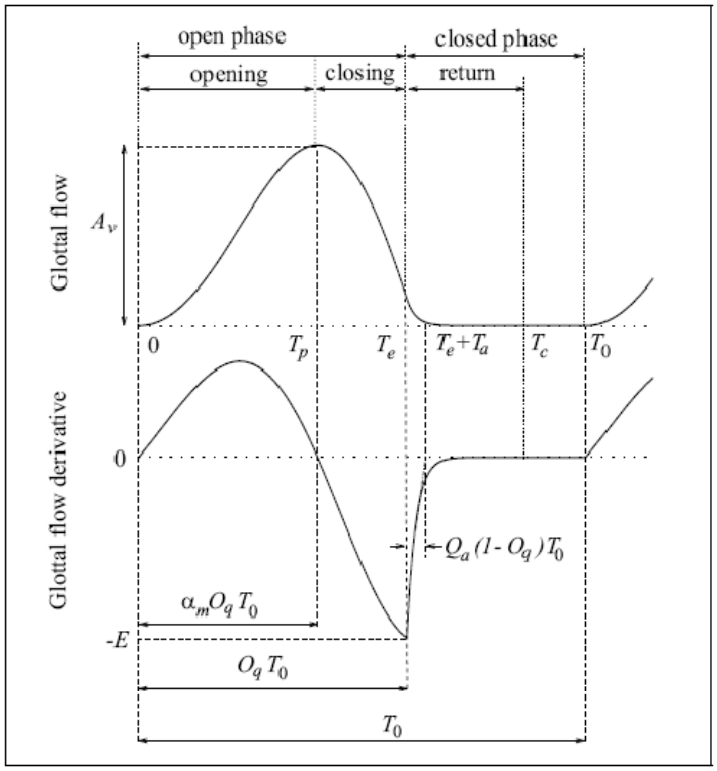

Figure 6.1: Parameters of a typical glottal flow signal and its derivative, (Doval et al. 2006).

The time-domain models of glottal flow (or its derivative) are described in terms of phases of the glottal flow signal. Figure 6.1, taken from (Doval et al. 2006), shows a typical glottal flow signal and its derivative during one fundamental period, T_0. It illustrates the following phases of the signal:

- **Open phase** when the vocal folds are open and glottal flow is present. This phase is further divided into two phases,
  - the **opening phase,** where the glottal flow increases from its baseline to its maximum value as the vocal folds open; and
  - the **closing phase,** where the glottal flow decreases from its maximum value as the vocal folds close until the glottal closure instant (GCI), which is the point at which the glottal flow derivate attains is negative minimum.
- **the closed phase,** when the vocal folds are closed and the glottal flow is at the baseline DC value. It should be noted that when there is a smooth closure, the

glottal flow derivative is continuous at the GCI and changes smoothly from its negative minimum to its baseline value, resulting in a **return phase**.

The significant time domain glottal flow model (GFM) parameters, as seen from Figure 6.1, are thus

- T_0 – The fundamental period
- A_v – Amplitude of voicing, the maximum amplitude of glottal flow
- T_p – The time at which A_v is reached
- E – Maximum excitation, the minimum negative value of the glottal flow derivative
- T_e – Glottal Closure Instant (GCI), the time at which the flow derivative reaches E
- T_c – Closure instant, time at which the glottal flow reaches its baseline DC value

Pioneering work on voice source signal modelling was done by Fant (Fant et al. 1985), Rosenberg (Rosenberg 1971) and others. Following this signal analysis approach, most of the glottal-flow models proposed have been time domain models, which describe the shape of the glottal excitation signal in the time domain. The most commonly used time-domain glottal models are

- KLGLOTT88 model (Klatt et al. 1990)
- Rosenberg C model (Rosenberg 1971)
- R++ model (Veldhuis 1998)
- LF model (Fant et al. 1985)

Time domain models have certain advantages since the model parameters can be directly related to temporal phases of the glottal flow signal. Also, time domain models lend themselves well to the study of glottal activity using time domain analyses such as

electroglottography, high-speed cinematography and electromyography. However, in some applications, a frequency-domain approach can be desirable. For instance, the main spectral parameters for synthesising voices with different voice qualities were found to be: 1. Spectral tilt; 2. Amplitude of the first few harmonics; 3. Increase of the first formant bandwidth; 4. Noise in the voice source; (Hanson 1995; Klatt et al. 1990).

6.1.1 The Glottal Spectrum

Based on the source-filter model, the speech spectrum can be viewed as the product of the vocal tract frequency response, the lip radiation frequency response, the glottal flow spectrum and a spectrum of an impulse train, with the impulses separated by the fundamental period. The spectrum of an impulse train is another series of impulses (Dirac comb) in the frequency domain separated by the fundamental frequency. This Dirac comb spectrum gives rise to the harmonic structure of speech while the envelope of the speech spectrum is determined by the magnitude responses of the vocal tract, lip radiation and glottal flow models (see Figure 6.2).

That is:

$$S(\omega) = R(\omega) \cdot V(\omega) \cdot G(\omega) \cdot P(\omega) \tag{6.1}$$

where, $S(\omega)$ is the speech spectrum, $R(\omega)$ is the lip radiation response, $G(\omega)$ is the glottal flow response, $V(\omega)$ is the vocal tract response and $P(\omega)$ is the spectrum of the impulse train (Dirac comb).

As previously mentioned the lip radiation is often modelled as a differentiator and combined with the glottal flow. Together they are considered to be the glottal flow derivative. The four commonly used time-domain glottal flow models also model the glottal flow derivative. In (Doval et al. 2006), all four are described within a common framework and their spectra analysed; suggesting that the magnitude spectrum of the

glottal flow derivative can be stylised by three asymptotic lines with +6dB/oct, -6dB/oct and -12dB/oct slopes as shown in Figure 6.3. Such a stylised representation allows for a very compact characterisation of the glottal flow derivative magnitude spectrum since the stylised spectrum can be uniquely identified based on three values, F_g, F_c and A_g. The compact representation also lends itself to use as a feature in a classification system.

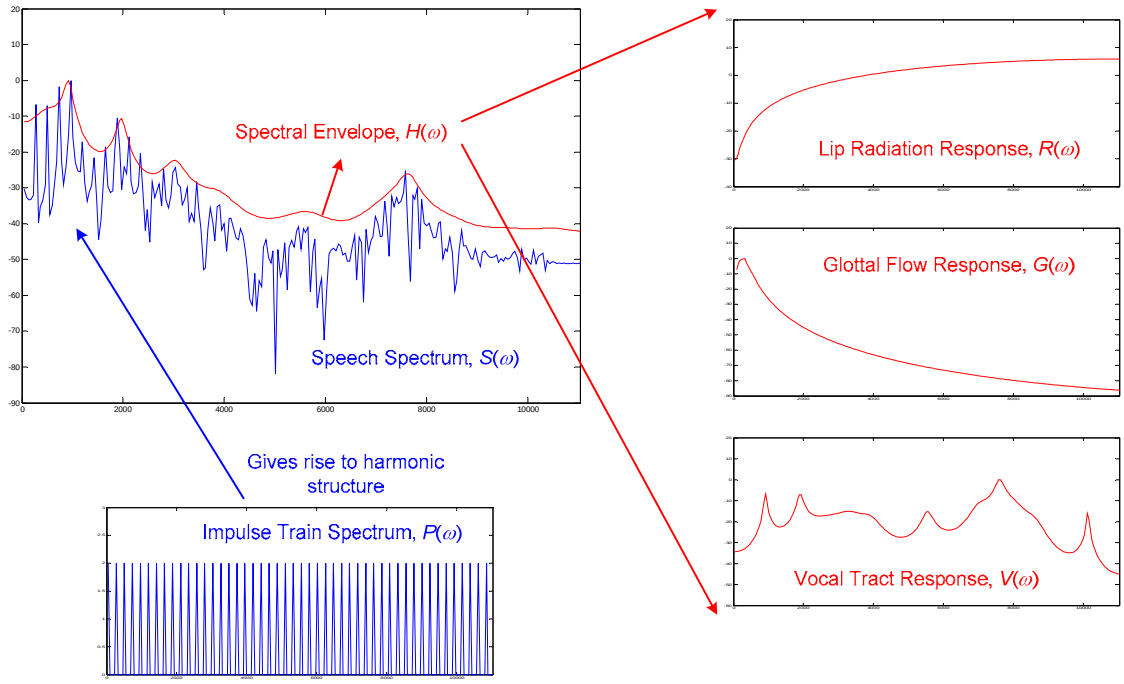

Figure 6.2: Spectral Structure of Speech

6.1.1.1 Estimation of Glottal Spectral Parameters

While it is obvious that the use of a glottal flow (or glottal flow derivative) model results in a more accurate modelling of the speech production apparatus when compared to the two-pole (with both poles at unity) approximation outlined in section 3.1.3, the estimation of the glottal flow signal from the speech signal is not a well defined problem and lacks an analytical solution. However, numerous techniques have been proposed over the years which are based on the properties of the glottal flow signal (Alku 1991; Cabral et al. 2008; Frohlich et al. 2001; Hui-Ling et al. 1999; Riegelsberger et al. 1993; Vincent et al. 2005). The iterative adaptive inverse filtering (IAIF) method (Alku 1991) was used to

estimate the glottal flow derivative in all the work reported in this thesis. The IAIF method can be used pitch synchronously (with variable window lengths based on pitch) or asynchronously (with fixed window lengths). For pitch synchronous IAIF, the DYPSA algorithm (Naylor et al. 2007) was used to detect glottal closure instants (GCI) and hence identify window boundaries. Given an estimate of the glottal flow derivative, the best stylised fit to its magnitude spectral envelope (Figure 6.3), in terms of minimum mean squared error (MMSE), was estimated via a brute force search of the 3-dimensional parameter space. An overview of the glottal spectral parameter estimation method is given in Figure 6.4.

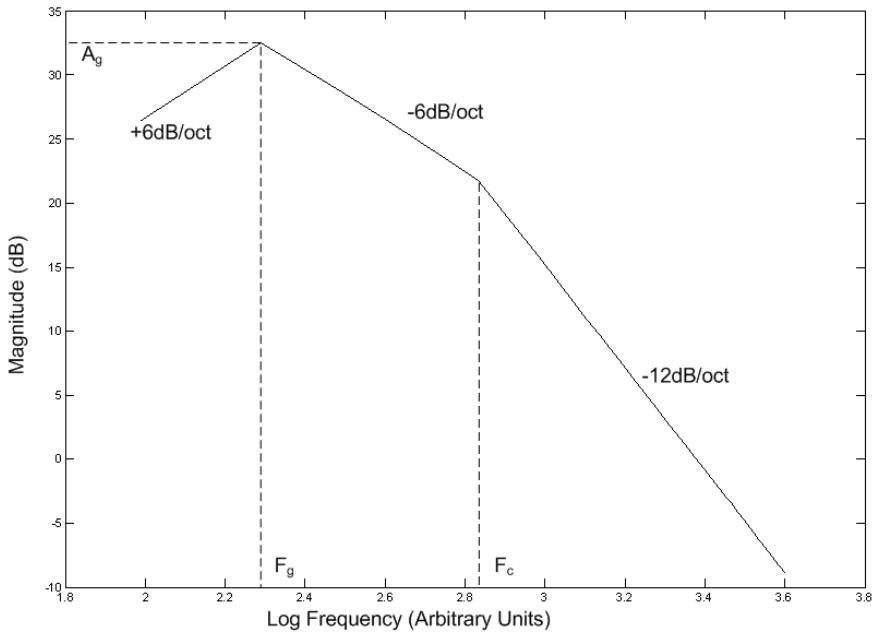

Figure 6.3: Stylised glottal flow derivative magnitude spectrum

Preliminary visual comparisons of the spectral fit obtained and the glottal flow spectrum across consecutive frames showed glottal flow spectra with different errors resulting in similar spectral fits (particularly the location of the glottal formant). This indicates the spectrum fitting process is robust (to a certain extent) to errors in the glottal flow derivative estimation process that result from incomplete removal of the formant structure, particularly in terms of identifying the glottal formant. However, the estimation

of the corner frequency, F_c, is affected to a much larger degree by the errors and the estimated values were not very reliable. Therefore, F_c was ignored as a glottal parameter and only the glottal formant frequency and magnitude, F_g and A_g, were used in the work reported in this thesis.

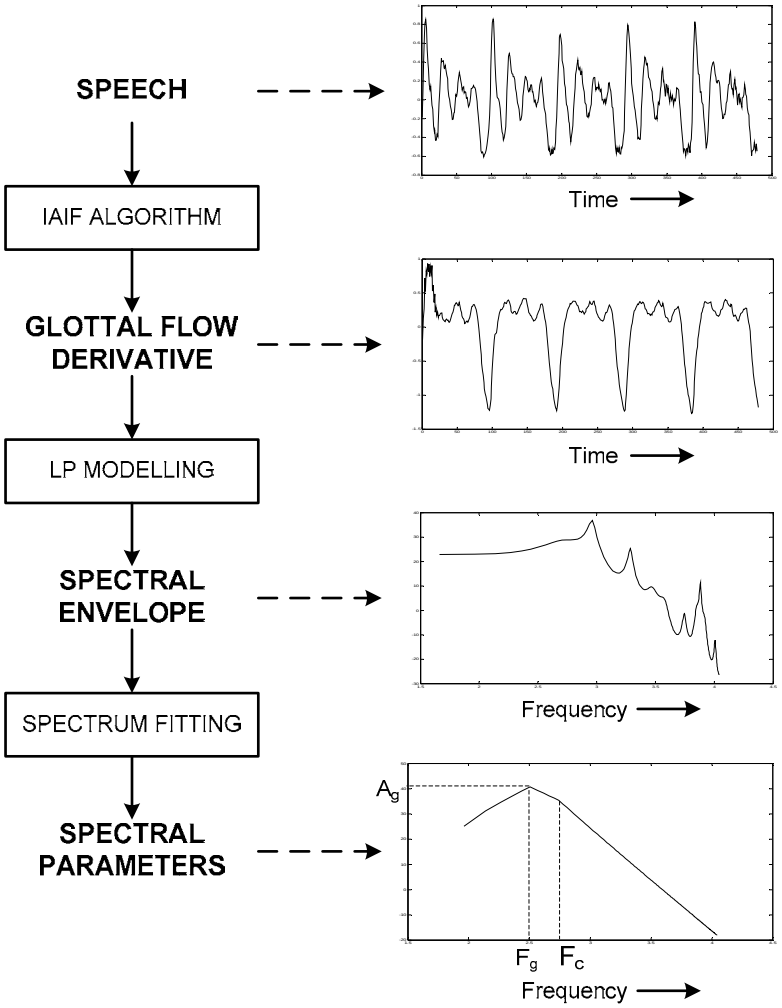

Figure 6.4: Overview of glottal spectral parameter estimation procedure

6.1.1.2 Glottal Parameters as static features

The AER system outlined in section 3.5, used to evaluate the features outlined in Chapter 3, can be used here to evaluate the utility of the glottal parameters. A 2-dimensional feature vector comprised of the two glottal formant parameters, $[F_g, A_g]$, was used as the front-end and a GMM based classifier was used to make turn level decision in both

speaker dependent and independent scenarios. The classification accuracies obtained in the two scenarios are reported in Table 6.1 and Table 6.2. These accuracies are substantially above random chance (20%) and indicate that the shape of glottal flow derivative contributes towards the expression of emotions. This is also in accordance with the findings in (Rui et al. 2009) that glottal features, albeit different ones from those proposed here, exhibit statistically significant difference between emotions.

Table 6.1: Confusion matrix for Speaker **INDEPENDENT** AER system using Glottal Parameters

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**58.5 %**	0 %	24.5 %	1.9 %	15.1%
Anger	0 %	**81.1 %**	0 %	17.6 %	1.4 %
Sad	47.3 %	1.4 %	**20.3 %**	9.5 %	21.6 %
Happy	1.2 %	21.2 %	5.9 %	**56.5 %**	15.3 %
Bored	48.9 %	0 %	13.0 %	14.1 %	**23.9 %**
Overall Accuracy = 46.6 %					

Table 6.2: Confusion matrix for Speaker **DEPENDENT** AER system using Glottal Parameters

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**81.3 %**	0 %	0 %	0 %	18.8%
Anger	3.6 %	**67.9 %**	0 %	25.0 %	3.6 %
Sad	21.4 %	3.6 %	**60.7 %**	3.6 %	10.7 %
Happy	6.3 %	9.4 %	12.5 %	**65.6 %**	6.3 %
Bored	17.1 %	0 %	37.1 %	11.4 %	**34.3 %**
Overall Accuracy = 59.0 %					

6.2 Temporal Patterns of Pitch

A second aspect of the source-filter model that is not captured very well by the features outlined in Chapter 3 is information about the variation of the model parameters with time within an utterance, or within segments of speech. The estimation of deltas and shifted deltas allow for a limited description of the temporal evolution of a parameter, these are

still short term descriptions, involving only adjacent frames (or a few frames in the case of shifted deltas). Parameter contours are representative of the variations over an entire utterance and are characterised by a much longer duration. The most common and probably best studied of these is the pitch (F_0) contour, which is essentially pitch as a function of time, and is the focus of this section. While most studies agree on the importance of global prosodic parameters such as F_0 level, F_0 range, loudness and rate of speech, F_0 contours are taken into account less frequently in the context of emotion recognition even though they have been shown to play an important role in the expression and perception of emotion (Burkhardt et al. 2000; Mozziconacci et al. 1999).

The significance of the temporal patterns can be gauged from a comparison of an AER system that makes use of these patterns with one that does not. In the section, the pitch contour is parameterised and the estimated parameters are employed as features. This system is then be compared with that described in Section 3.5 using pitch values as features, and the results are reported in Section 6.2.4.3.

6.2.1 Contour Parameterisation

Linear stylisation of F0 contours (Mertens et al. 1995; Ravuri et al. 2008; Wang et al. 2005) is commonly carried out to make them simpler to analyse, but has the additional advantage of making their representation more compact than that of the original contour. Approximating the pitch contour in each voiced segment by a straight line enables the representation of that contour using three parameters. This is different from typical F0 contour stylisations (Mertens et al. 1995; Wang et al. 2005) since each voiced segment is approximated by a single linear segment rather than a piecewise linear approximation in order to utilise a single vector to characterise the contour in each voiced segment (refer to section 6.2.4.1). This is the simplest form of contour parameterisation and if an AER

system that makes use of such a representation outperforms a system that models the distribution of pitch values without taking into account any temporal information, it is reasonable to suppose that the shapes of the pitch contours contain emotion specific information.

The RAPT algorithm for pitch estimation (Talkin 1995) was used to estimate pitch contours from speech. A separate voicing activity detector (VAD) was used to identify voiced segments prior to linear curve fitting of the pitch contours in these segments as in (Ravuri et al. 2008). The linear approximation in each segment is represented by the slope of the line (*s*), the initial offset (*b*) and the length of the segment (*x*) as shown in Figure 6.5. Thus the pitch contour of any speech sample can be represented by *3N* parameters, where *N* is the number of voiced segments in the utterance.

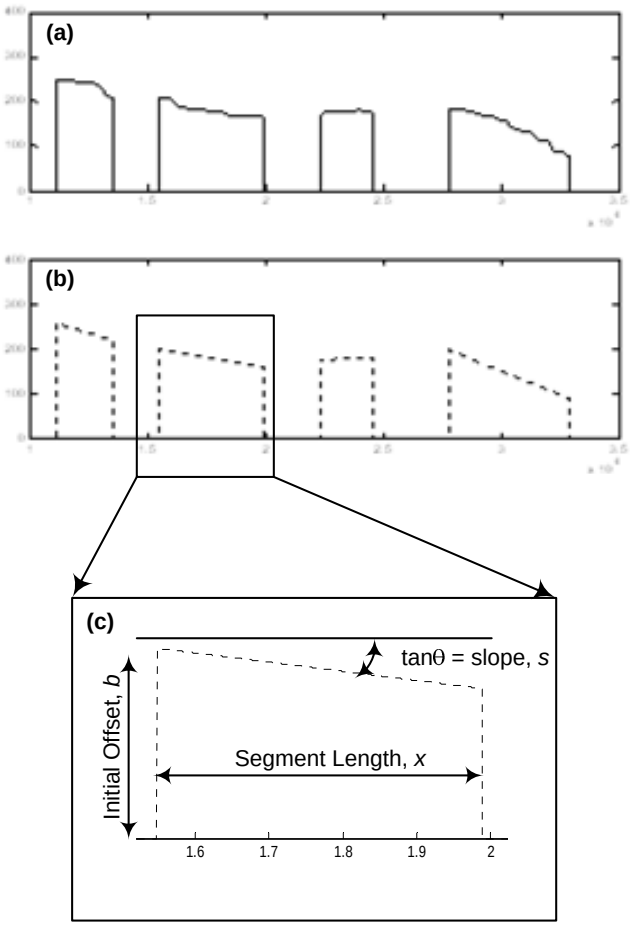

Figure 6.5: (a) Estimated F0 contour (b) Linear approximation of F0 contour (c) Linear model parameters - b, s and x

A subjective comparison of:

- speech synthesised using the approximate F0 contour
- speech synthesised using the actual estimated F0 contour, and
- the actual speech sample

was performed to determine if the linear approximations of F0 segments retain information representative of emotions from the actual speech signal, and hence to determine their utility as AER features.

6.2.2 A Voiced-Speech Synthesis Technique

In the work reported in this section, the purpose of speech synthesis is to enable subjective comparisons of speech samples that use the estimated pitch contour with speech samples whose pitch contours have been replaced by linear approximations. Given the sole focus on pitch, a synthesis method based on a non-stationary AM-FM type representation of speech was used. This method is very close to the sinusoidal representation (McAulay et al. 1986).

$$s(t) = \sum_{k=1}^{N} A(kf(t), t) \sin\left(\int_{0}^{t} kf(\tau) d\tau \right) \tag{6.2}$$

where $f(t)$ is the F0 contour, N is the number of harmonics and $A(f, t)$ is an estimate of the spectral magnitude as a function frequency and time.

The representation of speech as a sum of harmonic sinusoids as given in (6.2) is directly dependent on the pitch contour, and allows for synthesis with both the estimated contour and its linear approximation. The spectrogram of the speech signal was used to determine the amplitude of the sinusoids for all the synthesis reported in this work.

However, other estimates such as the LPC spectrum or more complex forms reported in (Kawahara et al. 1999), for example, may also be used.

6.2.3 Subjective Evaluation

Speech data from the LDC Emotion Prosody speech corpus (Liberman et al. 2002) was used in the subjective evaluations. Two listening tests were performed to determine whether the linear approximations to pitch contour segments retained sufficient information about the emotion being expressed by speech. Both tests were taken by the same eleven untrained listeners. Synthetic speech used in both listening tests was produced from spectrograms estimated from speech samples taken from the LDC database, and either the actual pitch contours estimated from these samples or linear approximations of the estimated contours.

6.2.3.1 Accuracy of linear approximation

The first test compared speech re-synthesised using the linear approximations with speech re-synthesised using pitch contours estimated from the original samples. The eleven untrained listeners were given two utterances, which they could listen to as many times as they needed to, and asked to give a non-fractional score between 1 and 5 depending on how close the two utterances were to each other. The scores were described as follows.

- 5 – Utterances are indistinguishable
- 4 – Utterances sound very similar
- 3 – Utterances sound moderately similar
- 2 – Utterances have very little similarity
- 1 – Utterances are completely dissimilar

The listeners were also asked to consider only how close the two utterances were to each other and to not take into account any other factors such as intelligibility, quality, clarity of emotional expression, etc. Each listener rated 30 comparisons, of which 15 were control, where both utterances were identical (both were speech re-synthesised using the estimated pitch contour). For the other 15 comparisons one utterance was speech re-synthesised using linear approximations to pitch contours and the other utterance was speech re-synthesised using estimated pitch contours. Re-synthesised speech using estimated pitch contours was used instead of actual speech so as to negate the effect of some quality loss due to the re-synthesis method, which is independent of approximations to the pitch contour. The utterances for the 15 control and 15 comparisons were chosen to produce 3 samples of the five emotions in each set but were otherwise selected randomly from the database. The scores given by each listener were normalised using the mean control score of that listener as given below.

$$\widehat{S_i} = \frac{S_i \times 5}{C_i} \tag{6.3}$$

where $\widehat{S_i}$ is the adjusted score for the i^{th} listener, S_i is the actual score and C_i is the mean control score.

Table 6.3: Subjective comparison scores (Range 1-5, with 5 indicating two versions were indistinguishable)

Emotion	Mean Score
Neutral	4.67
Anger	4.32
Sadness	4.84
Happiness	4.30
Boredom	4.87
Overall	**4.60**

The mean comparison scores for each of the five emotions and the overall mean comparison score are listed in Table 6.3. The high scores across all five emotions

indicates that the use of linear approximations is more or less indistinguishable from the use of the estimated pitch contours.

6.2.3.2 Emotion classification – Human

In the second test, listeners were given a sample of speech, which they could listen to as many times as necessary, and asked to classify it as one of the five emotions (Neutral, Anger, Sadness, Happiness and Boredom). Each listener classified 45 utterances, comprising three versions each of three samples drawn from each of the five emotions. The first version was the actual speech sample from the database, the second version was speech re-synthesised using the estimated pitch contour and the third version was speech re-synthesised using linear approximations to the pitch contours. The 45 utterances were presented in random order to the listeners. The confusion matrices for the three versions are given in Table 6.4, Table 6.5 and Table 6.6.

Table 6.4: Confusion matrix for original speech

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**69.7 %**	3 %	9.1 %	0 %	18.2 %
Anger	3 %	**93.9 %**	0 %	3 %	0 %
Sad	12.1 %	3 %	**57.6 %**	0 %	27.3 %
Happy	39.4 %	3 %	9.1 %	**45.5 %**	3 %
Bored	33.3 %	0 %	15.2 %	0 %	**51.5 %**
Overall Accuracy = 63.6 %					

Table 6.5: Confusion matrix for re-synthesised speech using actual estimated pitch contour

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**78.8 %**	3%	9.1 %	0 %	9.1 %
Anger	3 %	**78.8 %**	0 %	18.2 %	0 %
Sad	18.2 %	6.1 %	**51.5 %**	0 %	24.2 %
Happy	39.4 %	0 %	12.1 %	**39.4 %**	9.1 %
Bored	18.2 %	0 %	24.2 %	6.1 %	**51.5 %**
Overall Accuracy = 60.0 %					

Table 6.6: Confusion matrix for re-synthesised speech using linear approximations of pitch contours

	Neutral	Anger	Sad	Happy	Bored
Neutral	**60.6 %**	3 %	18.2 %	3 %	15.2 %
Anger	9.1 %	**72.7 %**	0 %	18.2 %	0 %
Sad	24.2 %	6.1 %	**39.4 %**	0 %	30.3 %
Happy	39.4 %	9.1 %	12.1 %	**30.3 %**	9.1 %
Bored	21.2 %	0 %	15.2 %	0 %	**63.6 %**
Overall Accuracy = 53.3 %					

From these confusion matrices, it can be seen that the class confusion patterns across the five emotions are more or less consistent for all three versions. However, anger is not identified as well in both re-synthesised versions as it is in the actual speech sample, even though it is still the most accurately recognised emotion in all three cases. The most likely reason for this drop in accuracy is that voice quality factors are not preserved very well by the re-synthesis method adopted in this investigation. There is a drop in accuracy for sadness as well, but it is not as significant as the drop for anger. Happiness is not very well recognised even in the first case, making it hard to infer anything from the results. The recognition rates for boredom and neutral are more or less consistent for all three cases. The recognition and confusion rates in the second and third cases are similar, indicating that the linear approximations to the pitch contours are able to capture a significant amount of the information that the pitch contours contain about the emotion being expressed.

To summarise, (i) the loss of voice quality as a result of the synthesis method led to a drop in recognition rates; and (ii) similar recognition rates in the second and third case indicate that the linear approximations are able to preserve emotion-specific information in pitch contours to a large extent.

6.2.4 Automatic Classification System

6.2.4.1 Front-End

An automatic emotion classification system based on the linear approximations to the pitch contour was constructed to help determine whether they were able to capture emotion-specific information. As shown in Figure 6.6, linear approximations to segments of the pitch contour of each utterance were determined. Each linear segment was represented by a three-dimensional vector comprising the slope of the linear fit (*s*), the initial offset (*b*) and the length of the segment (*x*) (refer Figure 6.5). Thus the entire utterance was represented by a sequence of *N* 3-dimensional vectors, and served as the front-end for the classification system.

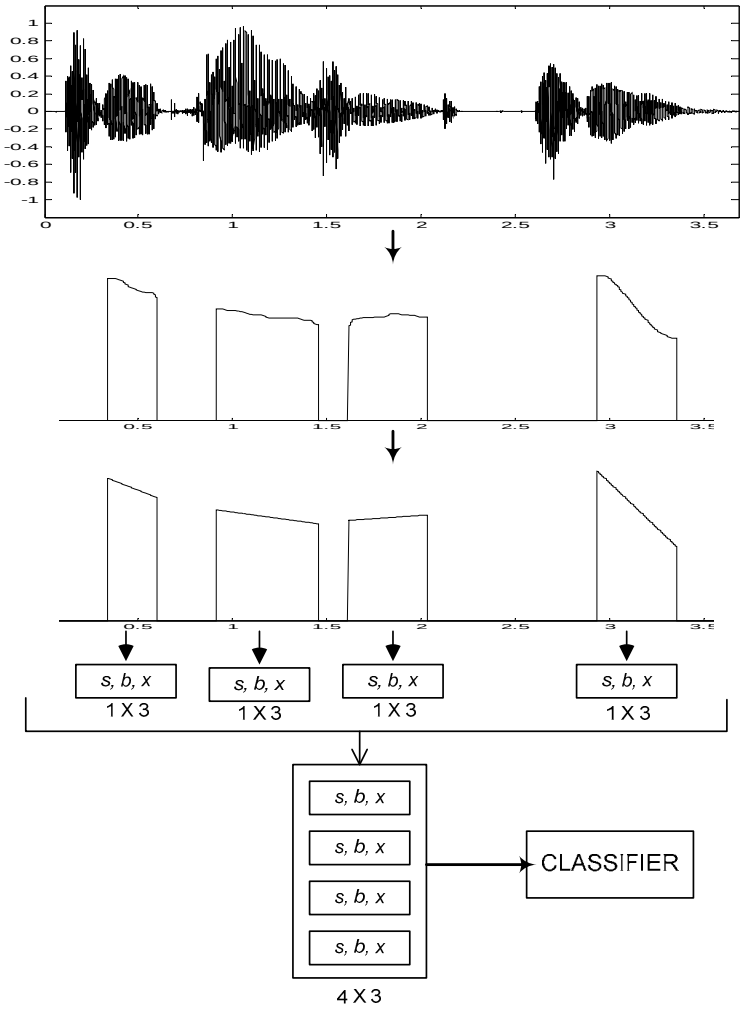

Figure 6.6: An overview of the system (s – slope, b – initial offset, x – segment length)

6.2.4.2 Back-End

Since the front-end produces a sequence of vectors for every utterance, the system requires a back-end that can model such sequences, and therefore a hidden Markov model (HMM) based back-end was chosen. An overview of hidden Markov models was provided in section 2.3.2.2. The number of states in the HMMs is determined by how much of the temporal variation in the contours between segments must be modelled and by how many segments were present in the utterances. A 3-state HMM has sufficiently many states to model the variations in the initial, central and terminal sections of the contour for the utterances employed herein without over-fitting and losing the ability to generalise. Preliminary experiments supported this choice. Each state was represented by a 4-mixture Gaussian mixture model. The speaker normalisation method outlined in section 4.1.2 was applied to the features prior to modelling and classification.

6.2.4.3 Classification Accuracy

The automatic classification system was implemented only in a speaker-independent configuration. Training and test datasets were chosen in the same manner as in the speaker-independent configuration outlined in section 3.5. i.e., all experiments were repeated 7 times in a 'leave-one-out' manner, using data from each of the 7 speakers as the test set in turn, and the data from the other 6 speakers as the training set. The accuracies reported are the means of the seven trials.

Table 6.7: Confusion matrix for the HMM based automatic emotion classification system

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**59.6 %**	0 %	19.2 %	6.4 %	14.9 %
Anger	0 %	**78.9 %**	2.8 %	16.9 %	1.4 %
Sad	13.1 %	1.6 %	**49.2 %**	14.8 %	21.3 %
Happy	1.4 %	32.9 %	16.4 %	**43.8 %**	5.5 %
Bored	6.5 %	0 %	28.6 %	10.4 %	**54.6 %**
Overall Accuracy = 57.1 %					

Comparing the accuracies reported in Table 6.4 with those reported in Table 6.7, it can be seen that the recognition rates for anger, sadness and boredom are similar, while the automatic classifier is not as good as humans in recognising neutral speech but is better than humans at recognising happiness. When comparing confusion rates, it can be seen that automatic classification and human classification are very different from each other, suggesting that the information contained in the pitch contours are used in different ways. While this is interesting, it suggests that direct comparisons of the two sets of accuracies must be done with a lot of care.

As previously mentioned, the value of modelling the pitch contour rather than just the statistical distribution of pitch values (without taking into consideration any temporal dependence) can be estimated by comparing the performance of a GMM based classification system that uses pitch values as its features, with the performance of the system described above. In fact, the performance of such a GMM based system is reported in section 3.2.4 and repeated here (Table 6.8) for convenience.

Table 6.8: Confusion matrix for the GMM based automatic emotion classification system (No temporal pattern)

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**47.2 %**	0 %	24.5 %	3.8 %	24.5%
Anger	0 %	**75.7 %**	1.4 %	23.0 %	0 %
Sad	31.1 %	1.4 %	**50.0 %**	12.2 %	5.4 %
Happy	0 %	29.4 %	18.8 %	**47.1 %**	4.7 %
Bored	47.8 %	2.2 %	21.7 %	8.7 %	**19.6 %**
Overall Accuracy = 46.6 %					

In the GMM-based system, the mixture models capture all the statistical information present in these values but not the temporal information that may be contained in the shape of the pitch contours. On the other hand, the HMM-based system that models the feature sequences based on linear approximations to the pitch contour captures this

temporal information along with the pitch values. Thus, the contribution of temporal information towards recognising emotions can be determined by comparing the recognition rates of these two systems.

Comparison of Table 6.7 with Table 6.8 shows that the performance of the HMM-based system was much better than that of the GMM-based system: in terms of overall accuracy, 56.4% as opposed to 46.6 %. This lends support to the claims in (Burkhardt et al. 2000) and (Mozziconacci et al. 1999) that temporal information contained in the shape of the pitch contour is useful in conveying and perceiving emotions and that this information is preserved by the linear approximations to the pitch contours to a large extent. A comparison with a GMM-based system that uses delta-pitch along with pitch as its features (overall accuracy of 53%) also indicated that the addition of temporal information increases the recognition rate. The overall classification accuracies for both human and automatic classification reported in this section are listed in Table 6.9.

Table 6.9: Summary of overall accuracies

Classification Test	Accuracy
Human – Actual Speech	63.6 %
Human – Re-synthesised with estimated F0	60.0 %
Human – Re-synthesised with approximate F0	53.3 %
Automatic – using temporal information	57.1 %
Automatic – without any temporal information	46.6 %

6.3 Temporal Parameter Contours

While pitch contours have been analysed in a large number of contexts, other parameters of the speech production model (source-filter model) also vary with time and contours of these parameters can also be considered, i.e., we can generalise the temporal modelling approach of section 6.2 to any frame-based parameter. For instance, another commonly

considered set of contours are the formant contours. Given that the source-filter model describes speech as the response of an all pole filter modelling the vocal tract to the glottal flow excitation, all aspects of the model can be compactly represented by a small number of model specific parameters. In the case of the vocal tract, the formant frequencies which are the dominant frequencies of the system can be used to parameterise it. Also, given that the vocal excitation is a series of glottal pulses, the shape of the glottal flow derivative (considering the glottal flow and lip radiation together) can be parameterised by F_g and A_g as outlined in section 6.1.1, while the period of each pulse is parameterised by pitch. Thus voiced speech can be described by a small set of parameters (Figure 6.7), each of which evolves with time.

Also, if these model parameter contours in each voiced segment of speech can be further parameterised by linear approximations, as previously outlined for pitch contours (section 6.2.1), an utterance can be compactly represented by a sequence of vectors, $\mathbb{V}$:

$$\mathbb{V} = \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \vdots \\ \boldsymbol{v_N} \end{bmatrix} \tag{6.4}$$

where, N is the number of voiced segments in the utterance and $\boldsymbol{v_i}$ is a vector corresponding to the i^{th} voiced segment.

$$\boldsymbol{v_i} = \begin{bmatrix} s_{j_i} & b_{j_i} & s_{j_i} & b_{j_i} & \dots & x_i \end{bmatrix} \tag{6.5}$$

where, x_i is the length of the i^{th} voiced segment; s_{j_i} and b_{j_i} are the linear coefficients (slope and bias) that describe the contour of parameter j in the i^{th} voiced segment; and j is one or more of the 3-part source filter model parameters (Figure 6.7), i.e.

$$j \in \{F_1, F_2, F_3, A_1, A_2, A_3, F_g, A_g, pitch\} \tag{6.6}$$

Note that the lengths of the contours are identical for all model parameters within each voiced segment. Hence only one element, x_i, in each vector is required to represent

contour length. This representation also lends itself for use as a front-end to a classification system.

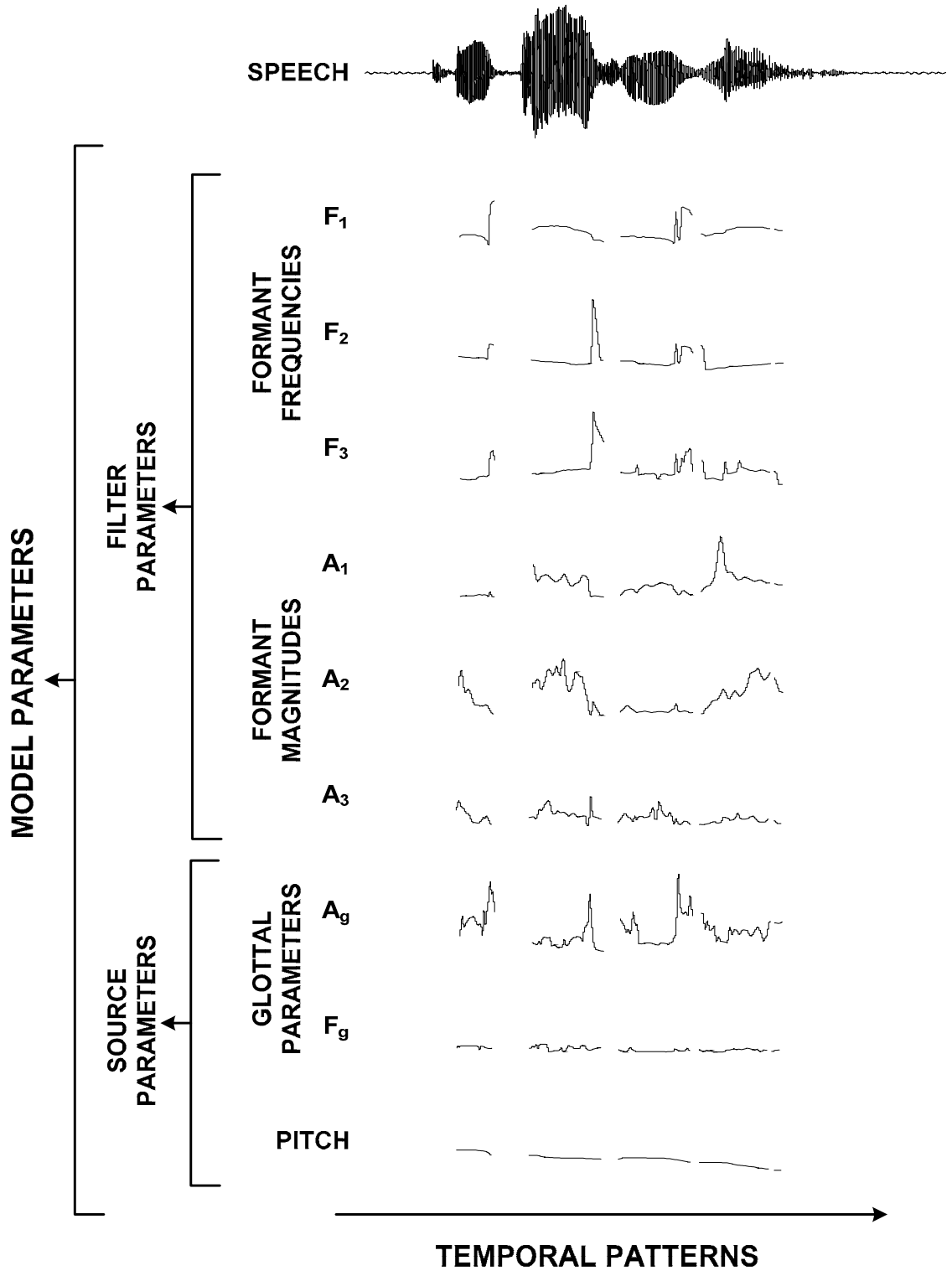

Figure 6.7: Parameter Contours – elements of the 3-part source-filter model based speech characterisation

6.3.1 Evaluating the use of contours as features

A comparison of the performances of AER systems that make use of parameter contours to those that model the parameter distributions without taking into account temporal patterns can be expected to provide some insight into the significance of the temporal patterns, both in a broad sense and on a feature-by-feature basis. Section 6.2.4 described such a comparison for pitch contours. Similar comparisons were performed here for glottal and filter (vocal tract) parameters and the results are reported in Table 6.10, Table 6.11, Table 6.12 and Table 6.13. F_g and A_g, describing the glottal formant, were used as glottal parameters and the frequencies and magnitudes of the first three formants, $[F_1, F_2, F_3, A_1, A_2, A_3]$ were used as vocal tract parameters. As in the previous section, a 3-state HMM based system (with a 4 mixture GMM in each state) using parameterised contours as features that made use of the temporal patterns was compared to a GMM based system (section 3.5) that did not make use of temporal information.

Table 6.10: Confusion matrix for HMM based AER system using contours of glottal parameters

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**53.2 %**	0 %	14.9 %	6.4 %	25.5 %
Anger	1.4 %	**80.3 %**	0 %	18.3 %	0 %
Sad	16.4 %	3.3 %	**31.2 %**	6.6 %	42.6 %
Happy	9.6 %	19.2 %	6.9 %	**50.7 %**	13.7 %
Bored	13.0 %	1.3 %	31.2 %	9.1 %	**45.5 %**
Overall Accuracy = 52.6 %					

Table 6.11: Confusion matrix for GMM based AER system using glottal parameters (No temporal pattern)

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**58.5 %**	0 %	24.5 %	1.9 %	15.1%
Anger	0 %	**81.1 %**	0 %	17.6 %	1.4 %
Sad	47.3 %	1.4 %	**20.3 %**	9.5 %	21.6 %
Happy	1.2 %	21.2 %	5.9 %	**56.5 %**	15.3 %
Bored	48.9 %	0 %	13.0 %	14.1 %	**23.9 %**
Overall Accuracy = 46.6 %					

Table 6.12: Confusion matrix for HMM based AER system using contours of formant parameters

	Neutral	Anger	Sad	Happy	Bored
Neutral	**55.3 %**	2.1 %	14.9 %	17.0 %	10.6 %
Anger	4.2 %	**46.5 %**	0 %	26.8 %	9.9 %
Sad	6.6 %	8.2 %	**32.8 %**	24.6 %	27.9 %
Happy	1.4 %	21.9 %	15.1 %	**37.0 %**	24.7 %
Bored	2.6 %	9.1 %	20.8 %	27.3 %	**40.3 %**
Overall Accuracy = 41.6 %					

Table 6.13: Confusion matrix for GMM based AER system using formant parameters (No temporal pattern)

	Neutral	Anger	Sad	Happy	Bored
Neutral	**49.1 %**	0 %	9.4 %	15.1 %	26.4%
Anger	2.7 %	**73.0 %**	2.7 %	20.3 %	1.4 %
Sad	16.2 %	6.8 %	**35.1 %**	14.9 %	27.0 %
Happy	12.9 %	15.3 %	15.3 %	**44.7 %**	11.8 %
Bored	15.2 %	9.8 %	25.0 %	8.7 %	**41.3 %**
Overall Accuracy = 48.2 %					

A summary of the overall classification accuracies of the GMM and HMM based systems using pitch, glottal and vocal tract parameters as features is given in Table 6.14.

Table 6.14: Summary of overall accuracies

Parameters	**Overall Accuracy**	
	Modelling Contours	**Modelling Statistics**
Pitch	**57.1 %**	46.6 %
Glottal	**52.6 %**	46.6 %
Vocal Tract	41.6 %	**48.2 %**

A comparison of the performances of the three classes of parameters indicates that parameter contours related to the vocal source (pitch and glottal parameter contours), are more suitable as features for an AER system than formant contours. Two possible reasons for this are

- The formant contours contain a significant amount of phonetic and speaker specific information which degrade emotion classification performance. The comparison of speaker dependent and speaker independent performance of formant features (Table 4.1) supports this conjecture.
- The pitch and glottal contours do not vary significantly within each voiced segment and can be approximated well by straight lines, but formant contours exhibit a larger degree of variation. This is suggested by informal observations of the contours and their corresponding linear approximations.

6.3.2 Alternative Contour Description

It was observed that in some cases the linear curve fitting was inaccurate due to either

- outliers in the parameter values estimated caused by errors in the estimation stage, or
- the contour shape being more complex and a straight line not being an adequate approximation.

In the first case, both the bias and the slope values can be very different from ideal values, even if the linear fit is the best possible one in a least squares sense. One way of overcoming this problem is to use the value of the midpoint of the linear fit to describe it rather than the slope and bias (Figure 6.8). Since outliers typically affect only one small section of the contour, the sensitivity of the midpoint will be lower than that of the bias and/or slope (Figure 6.9). The cost of using the midpoint rather than the slope and bias is a reduction in the amount of information available to the back end.

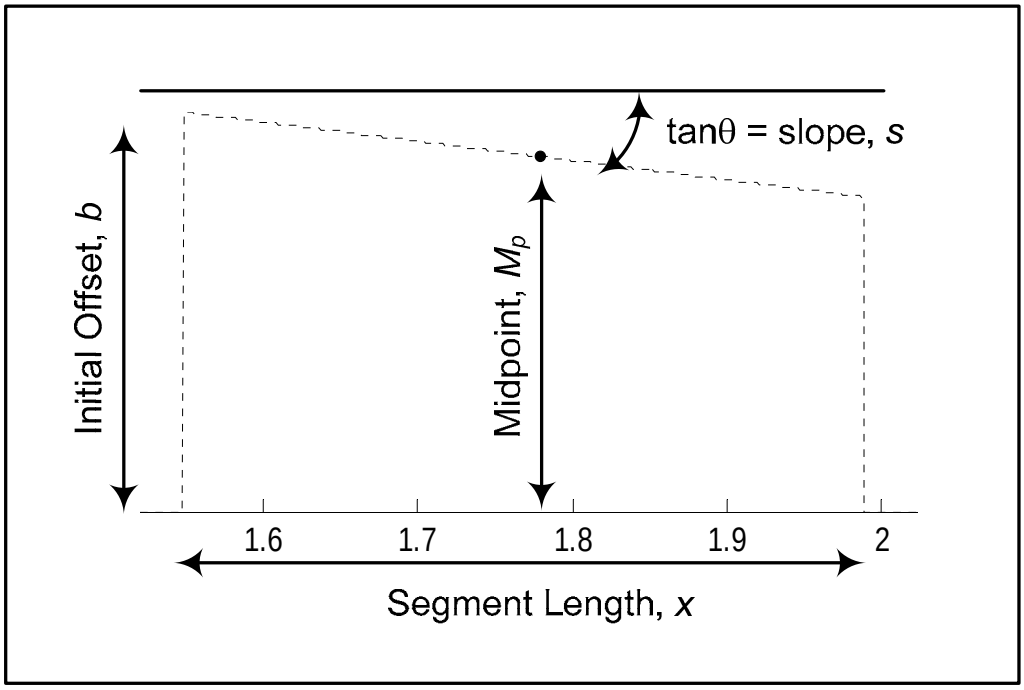

Figure 6.8: Midpoint, M_p may be used in place of Offset, b, and slope, s.

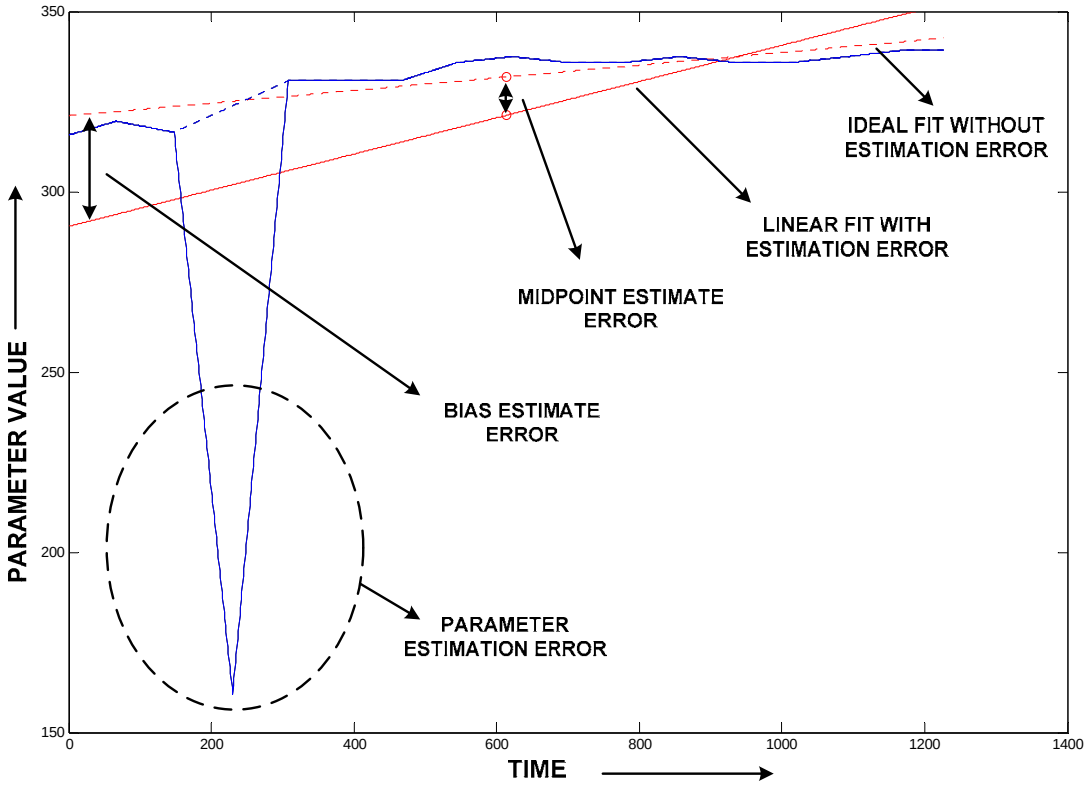

Figure 6.9: Midpoint error vs. Bias error in case of outlier.

In order to determine if any of the model parameter contours would benefit from the midpoint description rather than the slope and bias description, AER systems that used

the midpoints as features were implemented and their performances compared to those obtained when the slope and bias description was used (see section 6.3.1). The classification accuracies obtained are reported in Table 6.15, Table 6.16 and Table 6.17. A summary of the overall accuracies, comparing the midpoint description to the slope-bias description is given in Table 6.18.

Table 6.15: Confusion matrix for the AER system using pitch contours (MIDPOINT)

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**63.8 %**	0 %	14.9 %	4.3 %	17.0 %
Anger	0 %	**77.5 %**	2.8 %	18.3 %	1.4 %
Sad	13.1 %	0 %	**39.3 %**	26.2 %	21.3 %
Happy	4.1 %	27.4 %	16.4 %	**49.3 %**	2.7 %
Bored	10.4 %	1.3 %	28.6 %	7.8 %	**52.0 %**
Overall Accuracy = 56.2 %					

Table 6.16: Confusion matrix for the AER system using glottal parameters (MIDPOINT)

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**72.3 %**	0 %	14.9 %	0 %	12.8 %
Anger	0 %	**83.1 %**	0 %	15.5 %	1.4 %
Sad	6.6 %	1.6 %	**36.1 %**	9.8 %	45.9 %
Happy	1.4 %	23.3 %	2.7 %	**54.8 %**	17.8 %
Bored	16.9 %	0 %	33.8 %	15.6 %	**33.8 %**
Overall Accuracy = 55.0 %					

Table 6.17: Confusion matrix for the AER system using vocal tract parameters (MIDPOINT)

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**42.6 %**	4.3 %	25.5 %	14.9 %	12.8 %
Anger	0 %	**66.2 %**	8.5 %	16.9 %	8.5 %
Sad	6.6 %	8.2 %	**37.7 %**	11.5 %	36.1 %
Happy	8.2 %	19.2 %	19.2 %	**37.0 %**	16.4 %
Bored	3.9 %	7.8 %	29.9 %	18.2 %	**40.3 %**
Overall Accuracy = 45.0 %					

Table 6.18: Summary of overall accuracies

Parameters	Overall Accuracy	
	Slope-Bias Description (from Table 6.14)	**Midpoint Description**
Pitch	**57.1 %**	56.2 %
Glottal	52.6 %	**55.0 %**
Vocal Tract	41.6 %	**45.0 %**

Comparing the accuracies of the systems using the midpoint descriptions with that of the systems using the slope-bias description, it can be seen that the midpoint description improves the systems based on glottal parameter and vocal tract parameter contours. However, the slope-bias approach appears to be better suited for modelling pitch contours.

6.3.3 Combining Contours in an AER System

The results from the sections 6.2.4.3, 6.3.1 and 6.3.2 indicate that the contour modelling approach is better than the static distribution modelling approach for pitch and glottal formant parameters. Also, the pitch, glottal and vocal tract parameters describe independent (based on the source-filter model) and distinct components in the speech production mechanism and are hence independent of each other. Consequently, their contours can be expected to be complementary (unless the back end models identical information from the different parameter contours in each voiced segment). This suggests that a system, such as the one described in section 6.2.4 to model pitch contours (and used to model the other parameter contours in section 6.3.1), can be used to model all the contours by simply concatenating the contour descriptors (slope-bias or midpoint) to form the feature vector. It should be noted that while a static distribution modelling approach was better that the contour modelling approach for the vocal tract parameters, they would still contribute towards a combined system if they are complementary to the other

parameters. Such a system was constructed and its performance evaluated in a speaker independent manner (using leave one speaker out cross validation) and the classification accuracies obtained are reported in Table 6.19. This system uses the slope-bias description for pitch contours and the midpoint description for glottal and vocal parameter contours. The choice of midpoint vs. slope-bias description was based on the results reported in Table 6.18 and a summary of the overall accuracies of all contour modelling (HMM based) systems is given in Table 6.20.

Table 6.19: Confusion matrix for the HMM based AER system using all model parameters

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**55.3 %**	0 %	23.4 %	4.3 %	17.0 %
Anger	0 %	**81.7 %**	0 %	16.9 %	1.4 %
Sad	3.3 %	0 %	**57.4 %**	14.8 %	24.6 %
Happy	0 %	23.3 %	5.5 %	**60.3 %**	11.0 %
Bored	3.9 %	0 %	32.5 %	7.8 %	**55.8 %**
Overall Accuracy = 62.6 %					

Table 6.20: Summary of overall accuracies

Classification Test	Accuracy
Human – Actual Speech	63.6 %
Automatic – using pitch contours (slope-bias)	57.1 %
Automatic – using glottal parameter contours (midpoint)	55.0 %
Automatic – using vocal tract parameter contours (midpoint)	45.0 %
Automatic – using all model parameters	62.6 %

The classification accuracies obtained by the system making use of all the model parameter contours is higher than accuracies obtained by any of the individual systems, as expected. It should also be noted that the emotion specific classification accuracies (diagonal elements of the confusion matrix) are all higher than 55% indicating that the system does not suffer from any inherent bias against one or more of the emotions.

Moreover the performance of the combined system compares very well against human classification performance.

The best performing AER that did not make use of any temporal information (static modelling) had an overall classification accuracy of 59.0 %. This system modelled the distributions of all three component (pitch, glottal and vocal tract) parameters. The best performing AER system (also static modelling and did not make use of any temporal information) that made use of features discussed in Chapter 3 used pitch, energy and MFCCs (PE + MFCC) and had an overall classification accuracy of 58.9 % when the number of Gaussian mixtures in the GMM based back-end was optimised empirically.

6.4 Summary

This chapter has taken a second look at the traditional source-filter model widely used in speech processing tasks and in particular the assumption about the vocal excitation that is inherent in common feature extraction procedures. Namely, it has examined the assumption that the glottal spectrum can be modelled as the response of a system with a fixed transfer function even though literature indicates that the shape of the vocal excitation and consequently the shape of the glottal spectrum varies and determines voice quality. The first section of this chapter described the glottal flow spectrum and proposed the use of glottal formant frequency and magnitude as a 2-dimensional feature in emotion recognition systems.

While the source-filter model parameters can be viewed as characterising the speech spectrum at any instant in time, the longer-term (more than the few frames involved in delta features) temporal evolution of these parameters are typically not taken into account in automatic emotion recognition literature. The modelling of the temporal patterns of pitch contours in particular is the focus of the next section. Here the use of a simple

technique based on linear approximations of pitch contours was proposed, and sequences of these linear coefficients were employed to model the pitch contour for classification purposes. Listening tests validated the use of these linear approximations. A comparison of the classification accuracy of a system that models pitch contours to that of another system that models only the statistical distribution established that the shape of the contours does contain emotion specific information.

Finally, the idea of modelling contours was extended from pitch to the other model parameters in the third section of this chapter. Since these parameters tend to vary significantly more than pitch and/or contain more errors in their estimation, a less descriptive but more robust description of the linear approximations was introduced. Comparisons of systems that model contours with those that model statistical distributions were included to establish the significance of taking into consideration the contour shapes. The classification accuracy of a system that combined all model parameters (pitch, glottal parameters and vocal tract parameters) was found to be comparable to human classification accuracy.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

This thesis reports research conducted into automatic emotion recognition based on speech with the aim of: (i) developing an understanding of the relationship between speech parameters and emotions and in turn investigating novel features; (ii) understanding and reducing variability in these parameters not related to emotions; (iii) exploring different approaches to modelling emotion specific variations in the parameters; and (iv) investigating the modelling of temporal contours of parameters as opposed to their static distributions.

7.1.1 Investigation of Novel Features

Chapter 3 outlined several novel features proposed for use in AER systems: GFCCs, group delay, FM features, weighted frequency, wavelet scale feature, LPRCCs and fractal dimension. All of them, apart from fractal dimension, were interpreted in terms of the traditional source-filter model of speech production; which led to comparisons, in terms of classification accuracies, of features describing similar components of the speech production model in similar levels of detail. Such comparisons indicated that

- GFCCs and FM features, while usable as features, were not sufficiently superior to the traditional and established MFCCs to warrant a replacement.

- Group delay is characteristic of the formant bandwidth but is not directly representative of other information contained in the magnitude spectrum.
- The EMD based weighted frequency outperformed the other broad measures of the spectral power distribution, including the novel wavelet scale feature.

7.1.2 A Novel Speaker Normalisation Technique

Section 4.1.2 presented a novel technique to reduce the variance in feature that arises due to differences between speakers that are characteristic of them in order to improve the performance of a speaker independent emotion classification system. The proposed method involves the use of cumulative distribution mapping to transform features from each speaker, such that they are all mapped to the same distribution. Since all the features from all emotions from each speaker are mapped onto the same distribution, the relative distributions within the feature space for each speaker are not changed and consequently the emotion specific variability is maintained. However the different feature spaces for each speaker are shifted onto the same region of a new feature space to reduce speaker specific variability.

7.1.3 Investigation of Variability

Section 4.2 and section 4.3 report the investigation of speaker dependent and phoneme dependent variability in the context of emotion recognition. The performances of speaker independent AER systems with different front-ends using the proposed normalisation were compared with the performances of AER systems using the same front-ends but without normalisation since the differences in performance can be attributed to speaker specific variability. The analysis of phonetic variability reported was carried out by performing phone recognition prior to emotion modelling and then comparing the

recognition rates obtained by phoneme specific emotion models with those obtained by phoneme independent emotion models.

These comparisons revealed that MFCCs are very discriminative but are also very characteristic of the speaker. They also indicated that source specific features such as pitch and energy lend themselves more to normalisation than detailed spectral features such as MFCCs and group delay, suggesting that MFCCs may not be the front-end of choice in a speaker independent system. The phonetic analyses indicated that emotions are better conveyed by some phonemes than others. However, speaker specific variability has a larger effect on emotion models.

7.1.4 Investigating Classification Approaches

Investigations into the choice of classifier, the necessity of a pre-classification stage and the question of turn vs. frame based classification are reported in Chapter 5. In section 5.1, the abilities of three static classifiers, namely Gaussian mixture models, probabilistic neural networks and support vector machines, to model emotions were compared in terms of classification accuracy. None of them made explicit use of temporal emotions and all of them modelled patterns based on the distribution of training data (hence the term static classifiers as opposed to dynamic classifiers). The overall classification accuracies of AER systems based on all three were found to be similar. However, the recognition rates of the individual emotions indicated a bias towards ‘boredom’ in the PNN and SVM based systems, suggesting that a Gaussian mixture model based classifier was the preferable.

The investigation into the question of turn based vs. frame based classification approaches, reported in section 5.3, also indicated that the dimensionality of the features played a role. The turn based approach consisted of estimating statistics from the set of

features extracted from all frames in a turn (utterance) and modelling these statistics, while the frame based approach consisted of modelling the distribution of the features directly and hence implicitly modelling the statistical parameters. A comparison of the recognition rates obtained when using both approached suggested that the higher level of abstraction of the turn based approach suited high dimensional features while the frame based approach was preferable for low dimensional features.

7.1.5 Investigating 3-Component Source-Filter Model Parameters

The use of parameters derived from a source-filter speech production model to represent speech and subsequently as features in an AER system is explored in Chapter 6. In the first section, the use of an explicit glottal model in the traditional source-filter model framework was investigated and the glottal parameters were used as features in an AER system. The recognition rates achieved indicated that these glottal parameters were indicative of emotions.

The remainder of the chapter studied the temporal evolution of these model parameters and their use as features. Initially pitch contours were analysed and a linear approximation based contour parameterisation approach, validated by subjective evaluations, adopted for compact representation. The approximate parametric representation was used as features and the classification accuracies obtained were compared with those obtained when only static modelling was performed (i.e., contour shapes were not considered). The comparison indicated that modelling of the temporal variation in the form of contour shape, increased classification accuracy significantly.

Finally, the idea of contour shape modelling was extended to other model parameters and similar comparison of systems that model contour shapes to those that model only

static distributions indicated significant improvements in classification accuracies could be gained by taking temporal variations into account.

The overall accuracies (speaker independent) of the best performing systems that make use of static modelling approach and that models contour shapes are reported in Table 7.1, along with the classification accuracy achieved by humans.

Table 7.1: Summary of overall accuracies

Classification Test	Accuracy
Human – Actual Speech	63.6 %
Automatic – Static modelling	59.0 %
Automatic – Modelling all parameter contours	62.6 %

7.2 Future Work

The research outlined in this thesis involved an investigation of the use of acoustic and prosodic parameters from speech for automatic recognition of emotions. However, these are not the only cues available from speech; and based on what is known about how humans recognise emotions a number of other cues can also be exploited. An exploration of their use in an automatic emotion recognition framework offers a number of possible avenues for future work. It is also important to acknowledge some of the limitations of the studies reported in this thesis. Future work should also involve addressing some if not all of these limitations. Some of the avenues that address the limitations and those that would be natural extensions of the work described in this thesis are listed below.

- All the experiments reported were performed on a single database which was limited in size, contained speech in only one language (English) and consisted of acted emotional speech. This in turn limits the work reported in this thesis to an exploratory investigation that unearthed interesting hypotheses that need to be

validated on other larger databases under different contexts (different languages, elicited emotions as opposed to acted ones, etc.)

- When humans recognise emotions from speech, its taken for granted that they also recognise what is being said. It is also known that the linguistic content plays a role in human emotion recognition. In order for automatic emotion recognition systems to approach human capabilities, it is essential to investigate the relationship between the linguistic content of speech and emotions.
- The performance of various features on a common back end is included in this thesis and is useful for comparison purposes. However, the use of classifiers individually optimised for different features, followed by a fusion of their predictions was not investigated and is an obvious avenue for future research. This is in fact an area of considerable ongoing research around the world.
- The work reported in this thesis indicates that modelling temporal variations of speech parameters leads to an inprovement in system performance. This suggests that an investigation to determine the optimum representation, if any, and suitable modelling approaches of these temporal contours is warranted.
- The human brain simultaneously recognises linguistic and paralinguistic (speaker identity, gender, age, emotion, cognitive load, etc.) information and it is most likely these tasks are integrated, unlike automatic systems which deal with only one of these issues each. Moreover in any automatic classification system, the other categories act as sources of variability and noise, degrading performance. An investigation into the combined modelling of linguistic and paralinguistic information could lead to improvement in all individual recognition tasks.

Appendix A

Classification Accuracies for the Features Reported in Chapter 3

A.1 Typical Features used in AER Systems

A.1.1 Mel Frequency Cepstral Coefficients (MFCCs)

Table A-1: Confusion matrix for Speaker **INDEPENDENT** AER system using MFCCs

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**45.3 %**	0 %	24.5 %	0 %	30.2 %
Anger	0 %	**74.7 %**	1.4 %	22.9 %	0 %
Sad	12.2 %	0 %	**37.8 %**	10.8 %	39.2 %
Happy	2.4 %	16.5 %	11.8 %	**56.5 %**	12.9 %
Bored	20.7 %	0 %	29.4 %	15.2 %	**34.8 %**
Overall Accuracy = 49.7 %					

Table A-2: Confusion matrix for Speaker **DEPENDENT** AER system using MFCCs

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**100.0 %**	0 %	0 %	0 %	0 %
Anger	3.6 %	**85.7 %**	0 %	10.7 %	0 %
Sad	3.6 %	3.6 %	**64.3 %**	10.7 %	17.8 %
Happy	3.1 %	12.5 %	6.3 %	**65.6 %**	12.5 %
Bored	8.6 %	0 %	17.1 %	2.9 %	**71.4 %**
Overall Accuracy = 74.8 %					

A.1.2 Formant Frequencies

Table A-3: Confusion matrix for Speaker **INDEPENDENT** AER system using formant information

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**43.4 %**	0 %	17.0 %	1.9 %	37.7 %
Anger	1.4 %	**82.4 %**	0 %	13.5 %	2.7 %
Sad	27.0 %	5.4 %	**37.8 %**	6.8 %	23.0 %
Happy	5.9 %	27.1 %	9.4 %	**32.9 %**	24.7 %
Bored	16.3 %	7.6 %	32.6 %	16.3 %	**27.2 %**
Overall Accuracy = 43.7 %					

Table A-4: Confusion matrix for Speaker **DEPENDENT** AER system using formant information

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**81.3 %**	0 %	0 %	0 %	18.8 %
Anger	3.6 %	**64.3 %**	0 %	28.6 %	3.6 %
Sad	7.1 %	0 %	**64.3 %**	10.7 %	17.8 %
Happy	3.1 %	12.5 %	18.8 %	**53.1 %**	12.5 %
Bored	22.9 %	0 %	17.1 %	17.1 %	**42.8 %**
Overall Accuracy = 58.3 %					

A.1.3 Reflection Coefficients

Table A-5: Confusion matrix for Speaker **INDEPENDENT** AER system using reflection coefficients

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**35.9 %**	0 %	5.7 %	11.3 %	47.2 %
Anger	0 %	**74.3 %**	5.4 %	16.2 %	4.1 %
Sad	6.8 %	4.1 %	**37.8 %**	20.3 %	31.1 %
Happy	8.2 %	14.1 %	4.7 %	**52.9 %**	20.0 %
Bored	5.4 %	4.4 %	28.3 %	20.7 %	**41.3 %**
Overall Accuracy = 48.9 %					

Table A-6: Confusion matrix for Speaker **DEPENDENT** AER system using reflection coefficients

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**93.8 %**	0 %	0 %	0 %	6.3 %
Anger	0 %	**85.7 %**	0 %	14.3 %	0 %
Sad	10.7 %	0 %	**57.1 %**	3.6 %	28.6 %
Happy	3.1 %	6.25 %	3.1 %	**81.3 %**	6.3 %
Bored	11.4 %	0 %	17.1 %	20.0 %	**51.4 %**
Overall Accuracy = 71.2 %					

A.1.4 Pitch (Fundamental Frequency)

Table A-7: Confusion matrix for Speaker **INDEPENDENT** AER system using pitch

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**47.2 %**	0 %	24.5 %	3.8 %	24.5%
Anger	0 %	**75.7 %**	1.4 %	23.0 %	0 %
Sad	31.1 %	1.4 %	**50.0 %**	12.2 %	5.4 %
Happy	0 %	29.4 %	18.8 %	**47.1 %**	4.7 %
Bored	47.8 %	2.2 %	21.7 %	8.7 %	**19.6 %**
Overall Accuracy = 46.6 %					

Table A-8: Confusion matrix for Speaker **DEPENDENT** AER system using pitch

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**93.8 %**	0 %	6.3 %	0 %	0 %
Anger	3.6 %	**57.1 %**	0 %	39.3 %	0 %
Sad	17.9 %	7.1 %	**35.7 %**	21.4 %	17.9 %
Happy	6.3 %	25.0 %	12.5 %	**50.0 %**	6.3 %
Bored	20.0 %	2.9 %	22.9 %	11.4 %	**42.9 %**
Overall Accuracy = 51.8 %					

A.1.5 Intensity (Energy)

Table A-9: Confusion matrix for Speaker **INDEPENDENT** AER system using energy

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**26.4 %**	7.6 %	32.1 %	13.2 %	20.8%
Anger	16.2 %	**60.8 %**	9.5 %	10.8 %	2.7 %
Sad	36.5 %	13.5 %	**29.7 %**	14.9 %	5.4 %
Happy	22.4 %	29.4 %	20.0 %	**23.5 %**	4.7 %
Bored	29.4 %	14.1 %	31.5 %	16.3 %	**8.7 %**
Overall Accuracy = 28.8 %					

Table A-10: Confusion matrix for Speaker **DEPENDENT** AER system using energy

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**62.5 %**	6.3 %	6.3 %	6.3 %	18.8%
Anger	28.6 %	**17.9 %**	17.9 %	28.6 %	7.1 %
Sad	17.9 %	17.9 %	**32.1 %**	14.3 %	17.9 %
Happy	18.8 %	25.0 %	15.6 %	**12.5 %**	28.1 %
Bored	25.7 %	11.4 %	22.9 %	20.0 %	**20.0 %**
Overall Accuracy = 25.2 %					

A.1.6 Energy Slope (Spectral Slope)

Table A-11: Confusion matrix for Speaker **INDEPENDENT** AER system using energy slope

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**60.4 %**	0 %	5.6 %	3.8 %	30.2%
Anger	1.4 %	**78.4 %**	2.7 %	16.2 %	1.4 %
Sad	27.1 %	5.4 %	**24.3 %**	23.0 %	20.3 %
Happy	9.4 %	15.3 %	22.4 %	**41.2 %**	11.8 %
Bored	39.1 %	1.1 %	27.2 %	9.8 %	**22.8 %**
Overall Accuracy = 43.4 %					

Table A-12: Confusion matrix for Speaker **DEPENDENT** AER system using energy slope

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**93.8 %**	0 %	0 %	0 %	6.3%
Anger	3.6 %	**85.7 %**	0 %	10.7 %	0 %
Sad	17.9 %	0 %	**42.9 %**	14.3 %	25.0 %
Happy	6.3 %	15.6 %	15.6 %	**50.0 %**	12.5 %
Bored	20.0 %	2.9 %	25.7 %	8.6 %	**42.9 %**
Overall Accuracy = 59.0 %					

A.1.7 Zero Crossing Rate (ZCR)

Table A-13: Confusion matrix for Speaker **INDEPENDENT** AER system using ZCR

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**47.2 %**	0 %	24.5 %	7.5 %	20.7%
Anger	1.4 %	**81.1 %**	0 %	17.6 %	0 %
Sad	36.5 %	2.7 %	**14.9 %**	14.9 %	31.1 %
Happy	3.5 %	15.3 %	8.2 %	**50.6 %**	22.4 %
Bored	25.0 %	2.2 %	8.7 %	21.7 %	**42.4 %**
Overall Accuracy = 47.1 %					

Table A-14: Confusion matrix for Speaker **DEPENDENT** AER system using ZCR

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**62.5 %**	0 %	18.8 %	6.3 %	12.5%
Anger	0 %	**82.1 %**	7.1 %	3.6 %	7.1 %
Sad	46.4 %	0 %	**35.7 %**	3.6 %	14.3 %
Happy	9.4 %	21.9 %	15.6 %	**40.6 %**	12.5 %
Bored	25.7 %	0 %	37.1 %	11.4 %	**25.7 %**
Overall Accuracy = 46.8 %					

A.1.8 Spectral Centroid

Table A-15: Confusion matrix for Speaker **INDEPENDENT** AER system using spectral centroid

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**60.4 %**	0 %	11.3 %	1.9 %	26.4%
Anger	1.4 %	**70.3 %**	1.4 %	25.7 %	1.4 %
Sad	23.0 %	9.5 %	**10.8 %**	20.3 %	36.5 %
Happy	4.7 %	25.9 %	14.1 %	**42.4 %**	13.0 %
Bored	43.5 %	2.2 %	9.8 %	18.5 %	**26.1 %**
Overall Accuracy = 40.2 %					

Table A-16: Confusion matrix for Speaker **DEPENDENT** AER system using spectral centroid

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**62.5 %**	0 %	6.3 %	0 %	31.3%
Anger	0 %	**67.9 %**	3.6 %	17.8 %	10.7 %
Sad	10.7 %	10.7 %	**28.6 %**	10.7 %	39.3 %
Happy	6.3 %	18.8 %	15.6 %	**46.9 %**	12.5 %
Bored	34.3 %	2.9 %	20.0 %	14.3 %	**28.6 %**
Overall Accuracy = 44.6 %					

A.1.9 Phoneme Rate

Table A-17: Confusion matrix for Speaker **INDEPENDENT** AER system using phoneme rate

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**49.1 %**	1.9 %	9.4 %	15.1 %	24.5%
Anger	25.7 %	**8.1 %**	14.9 %	12.1 %	39.2 %
Sad	39.2 %	12.2 %	**8.1 %**	12.1 %	28.4 %
Happy	37.7 %	9.4 %	14.1 %	**11.8 %**	27.1 %
Bored	38.0 %	19.6 %	4.4 %	5.4 %	**32.6 %**
Overall Accuracy = 20.6 %					

Table A-18: Confusion matrix for Speaker **DEPENDENT** AER system using phoneme rate

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**18.8 %**	25.0 %	3.1 %	18.8 %	6.3%
Anger	14.3 %	**21.4 %**	25.0 %	14.3 %	25.0 %
Sad	7.1 %	21.4 %	**32.1 %**	25.0 %	14.3 %
Happy	9.4 %	12.5 %	28.1 %	**21.9 %**	28.1 %
Bored	8.6 %	11.4 %	40.0 %	20.0 %	**20.0 %**
Overall Accuracy = 23.0 %					

A.2 Novel Features

A.2.1 Gammatone Filter Cepstral Coefficients (GFCCs)

Table A-19: Confusion matrix for Speaker **INDEPENDENT** AER system using GFCCs

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**73.6 %**	0 %	11.3 %	0 %	15.1%
Anger	0 %	**79.7 %**	0 %	18.9 %	1.4 %
Sad	17.6 %	0 %	**33.8 %**	12.2 %	36.5 %
Happy	1.2 %	20.0 %	16.5 %	**47.1 %**	15.3 %
Bored	19.6 %	0 %	21.7 %	7.6 %	**51.1 %**
Overall Accuracy = 55.6 %					

Table A-20: Confusion matrix for Speaker **DEPENDENT** AER system using GFCCs

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**93.8 %**	0 %	0 %	0 %	6.3%
Anger	0 %	**85.7 %**	3.6 %	7.1 %	3.6 %
Sad	0 %	0 %	**71.4 %**	7.1 %	21.4 %
Happy	0 %	9.4 %	9.4 %	**65.6 %**	15.6 %
Bored	8.6 %	0 %	20.0 %	11.4 %	**60.0 %**
Overall Accuracy = 72.6 %					

A.2.2 LP Model Group Delay

Table A-21: Confusion matrix for Speaker **INDEPENDENT** AER system using Group Delay

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**58.5 %**	0 %	7.6 %	11.3 %	22.6%
Anger	0 %	**59.5 %**	0 %	40.5 %	0 %
Sad	24.3 %	1.4 %	**20.3 %**	28.4 %	25.7 %
Happy	8.2 %	15.3 %	14.1 %	**50.6 %**	11.7 %
Bored	18.5 %	1.1 %	27.2 %	21.7 %	**31.5 %**
Overall Accuracy = 42.9 %					

Table A-22: Confusion matrix for Speaker **DEPENDENT** AER system using Group Delay

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**93.8 %**	0 %	0 %	0 %	6.3%
Anger	0 %	**85.7 %**	3.6 %	7.1 %	3.6 %
Sad	7.1 %	0 %	**57.1 %**	7.1 %	25.0 %
Happy	9.4 %	12.5 %	6.3 %	**71.9 %**	0 %
Bored	11.4 %	2.9 %	20.0 %	11.4 %	**54.3 %**
Overall Accuracy = 69.8 %					

A.2.3 Frequency Modulation

Table A-23: Confusion matrix for Speaker **INDEPENDENT** AER system using FM

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**15.1 %**	0 %	37.7 %	3.8 %	43.4 %
Anger	4.1 %	**85.1 %**	1.4 %	9.5 %	0 %
Sad	12.2 %	0 %	**27.1 %**	17.6 %	43.2 %
Happy	2.4 %	25.9 %	18.8 %	**45.9 %**	7.1 %
Bored	18.5 %	1.1 %	34.8 %	4.4 %	**41.3 %**
Overall Accuracy = 44.4 %					

Table A-24: Confusion matrix for Speaker **DEPENDENT** AER system using FM

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**75.0 %**	0 %	6.3 %	0 %	18.8%
Anger	7.1 %	**67.9 %**	0 %	25.0 %	0 %
Sad	7.1 %	3.6 %	**50.0 %**	10.7 %	28.6 %
Happy	3.1 %	15.6 %	6.3 %	**75.0 %**	0 %
Bored	5.7 %	2.9 %	17.1 %	17.1 %	**57.1 %**
Overall Accuracy = 64.0 %					

Table A-25: Confusion matrix for Speaker **INDEPENDENT** AER system using GFCC + FM

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**56.6 %**	0 %	22.6 %	7.6 %	13.2 %
Anger	0 %	**78.4 %**	0 %	20.3 %	1.4 %
Sad	21.6 %	0 %	**33.8 %**	17.6 %	27.0 %
Happy	2.4 %	30.6 %	14.1 %	**44.7 %**	8.2 %
Bored	26.1 %	0 %	34.8 %	9.8 %	**29.4 %**
Overall Accuracy = 47.1 %					

Table A-26: Confusion matrix for Speaker **DEPENDENT** AER system using GFCC + FM

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**93.8 %**	0 %	0 %	0 %	6.3%
Anger	3.6 %	**78.6 %**	3.6 %	14.3 %	0 %
Sad	0 %	0 %	**71.4 %**	10.7 %	17.9 %
Happy	3.1 %	9.4 %	12.5 %	**75.0 %**	0 %
Bored	0 %	0 %	22.9 %	17.1 %	**60.0 %**
Overall Accuracy = 73.4 %					

A.2.4 EMD based Weighted Frequency (WF)

Table A-27: Confusion matrix for Speaker **INDEPENDENT** AER system using Weighted Frequency

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**58.5 %**	0 %	18.9 %	3.8 %	18.9%
Anger	0 %	**77.0 %**	0 %	20.3 %	2.7 %
Sad	21.6 %	0 %	**24.3 %**	24.3 %	29.7 %
Happy	5.9 %	15.3 %	14.1 %	**57.7 %**	7.1 %
Bored	34.8 %	0 %	18.5 %	19.6 %	**27.2 %**
Overall Accuracy = 47.6 %					

Table A-28: Confusion matrix for Speaker **DEPENDENT** AER system using Weighted Frequency

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**68.7 %**	0 %	6.3 %	0 %	25.0%
Anger	0 %	**82.1 %**	3.6 %	10.7 %	3.6 %
Sad	17.9 %	7.1 %	**32.1 %**	10.7 %	32.1 %
Happy	6.3 %	15.6 %	15.6 %	**56.3 %**	6.25 %
Bored	14.3 %	2.9 %	31.4 %	14.3 %	**37.1 %**
Overall Accuracy = 53.2 %					

A.2.5 Wavelet Scale based Feature

Table A-29: Confusion matrix for Speaker **INDEPENDENT** AER system using Wavelet Scale feature

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**39.6 %**	0 %	50.9 %	0 %	9.4%
Anger	0 %	**81.1 %**	1.4 %	17.6 %	0 %
Sad	55.4 %	0 %	**10.8 %**	21.6 %	12.2 %
Happy	2.4 %	17.7 %	5.9 %	**65.9 %**	8.2 %
Bored	37.0 %	0 %	31.5 %	17.4 %	**14.1 %**
Overall Accuracy = 41.8 %					

Table A-30: Confusion matrix for Speaker **DEPENDENT** AER system using Wavelet Scale feature

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**87.5 %**	0 %	6.3 %	0 %	6.3%
Anger	3.6 %	**71.4 %**	0 %	25.0 %	0 %
Sad	28.6 %	0 %	**53.6 %**	10.7 %	7.1 %
Happy	6.3 %	21.9 %	6.3 %	**53.1 %**	12.5 %
Bored	28.6 %	0 %	34.3 %	8.6 %	**28.6 %**
Overall Accuracy = 54.7 %					

A.2.6 LP Residue Cepstral Coefficients (LPRCC)

Table A-31: Confusion matrix for Speaker **INDEPENDENT** AER system using LPRCCs

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**49.1 %**	0 %	13.2 %	1.9 %	35.8%
Anger	0 %	**77.0 %**	5.4 %	17.6 %	0 %
Sad	17.6 %	1.4 %	**41.9 %**	12.1 %	27.0 %
Happy	4.7 %	18.8 %	22.4 %	**49.4 %**	4.7 %
Bored	39.1 %	1.1 %	19.6 %	4.3 %	**35.9 %**
Overall Accuracy = 50.0 %					

Table A-32: Confusion matrix for Speaker **DEPENDENT** AER system using LPRCCs

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**81.3 %**	0 %	12.5 %	0 %	6.3%
Anger	0 %	**71.4 %**	3.6 %	25.0 %	0 %
Sad	0 %	7.1 %	**57.1 %**	10.7 %	25.0 %
Happy	6.3 %	15.6 %	0 %	**71.9 %**	6.3 %
Bored	5.7 %	0 %	20.0 %	8.6 %	**65.7 %**
Overall Accuracy = 68.4 %					

A.2.7 Fractal Dimension (FD)

Table A-33: Confusion matrix for Speaker **INDEPENDENT** AER system using FD

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**62.3 %**	0 %	13.2 %	1.9 %	22.6%
Anger	1.4 %	**81.1 %**	2.7 %	14.9 %	0 %
Sad	18.9 %	0 %	**43.3 %**	27.0 %	10.8 %
Happy	5.9 %	20.0 %	22.4 %	**42.4 %**	9.4 %
Bored	40.2 %	1.1 %	28.3 %	15.2 %	**15.2 %**
Overall Accuracy = 46.3 %					

Table A-34: Confusion matrix for Speaker **DEPENDENT** AER system using FD

	Neutral	**Anger**	**Sad**	**Happy**	**Bored**
Neutral	**43.8 %**	0 %	31.3 %	0 %	25.0%
Anger	0 %	**64.3 %**	7.1 %	21.4 %	7.1 %
Sad	10.7 %	0 %	**42.9 %**	39.3 %	7.1 %
Happy	9.4 %	18.8 %	15.6 %	**40.6 %**	15.6 %
Bored	40.0 %	0 %	22.9 %	17.1 %	**20.0 %**
Overall Accuracy = 41.0 %					

Appendix B

Empirical Mode Decomposition

B.1.1 Analytic Representation - Instantaneous Frequency

Writing a signal as an analytic signal allows for a definition of instantaneous frequency as the time derivative of the phase of the complex valued analytic signal.

$$z(t) = x(t) + i\mathcal{H}\{x(t)\} \tag{B.1}$$

where, $x(t)$ is the real valued signal and $\mathcal{H}\{\cdot\}$ is the Hilbert transform operator and $z(t)$ is the analytic signal.

$$\phi(t) = \arctan\left(\frac{\mathcal{H}\{x(t)\}}{x(t)}\right) \tag{B.2}$$

$$\theta(t) = \frac{d\phi(t)}{dt} \tag{B.3}$$

$$a(t) = \sqrt{x^2(t) + \mathcal{H}\{x(t)\}^2} \tag{B.4}$$

where, $\phi(t)$ is the instantaneous phase, $\theta(t)$ is the instantaneous frequency and $a(t)$ is the instantaneous amplitude.

B.1.2 Decomposition

The necessary conditions for a meaningful definition of instantaneous frequency based on the analytic representation of the signal are that the signal is symmetric with respect to the local zero mean, and has the same number of extrema and zero crossings (Huang et al. 1998). Functions satisfying these conditions are referred to as intrinsic mode functions (IMF) in (Huang et al. 1998). The empirical mode decomposition (EMD) enables any signal to be written as a sum of a few intrinsic mode functions and in some cases a

monotonic residue that represents the overall trend of the signal. The empirical mode decomposition process begins by extracting the first intrinsic mode function, which consists of oscillations on the smallest scale, locally by a sifting process. This IMF is then subtracted from the signal and the process is iterated until all possible intrinsic mode functions have been extracted and only a monotonic residue is left (Figure B.1).

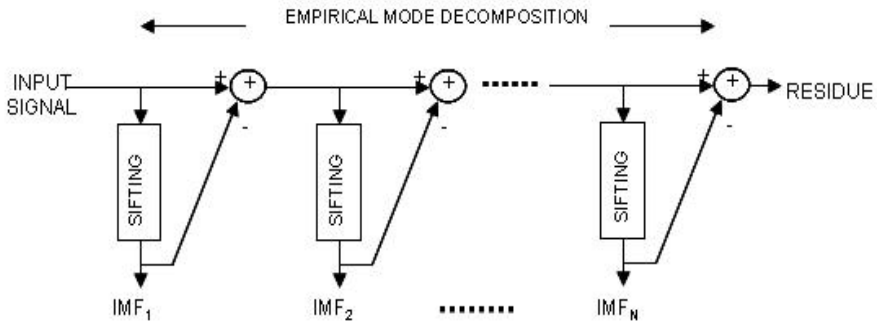

Figure B.1: Overview of Empirical Mode Decomposition

B.1.3 EMD and Speech

Due to the present lack of a mathematical framework for the EMD, there are limitations to the study of its properties. However, its application to speech signals may be investigated empirically to a certain extent.

Typically, speech signals sampled at 22050 Hz contain between 14 and 19 IMFs and the instantaneous amplitudes and frequencies derived from all these intrinsic mode functions (together with the residue) contain all the information present in the signal. In informal experiments it was observed that for speech signals in general approximately the first five modes (IMFs) contained most of the perceptually significant information.

An investigation of this observation, was conducted by measuring the PESQ scores of speech reconstructed from the *M* most significant modes together with mean IMF frequencies and mean IMF energies from over 9 min of 22050 Hz sampled speech. Results from this experiment, together with informal listening tests, showed that speech reconstructed from the first five modes (IMFs) was of sufficiently high quality for

classification tasks. Thus, only the first five modes were used in all the work reported in this thesis. Average PESQ scores obtained for speech signals reconstructed with different number of IMFs, and the mean instantaneous frequency of each intrinsic mode function are shown in Figure B.2.

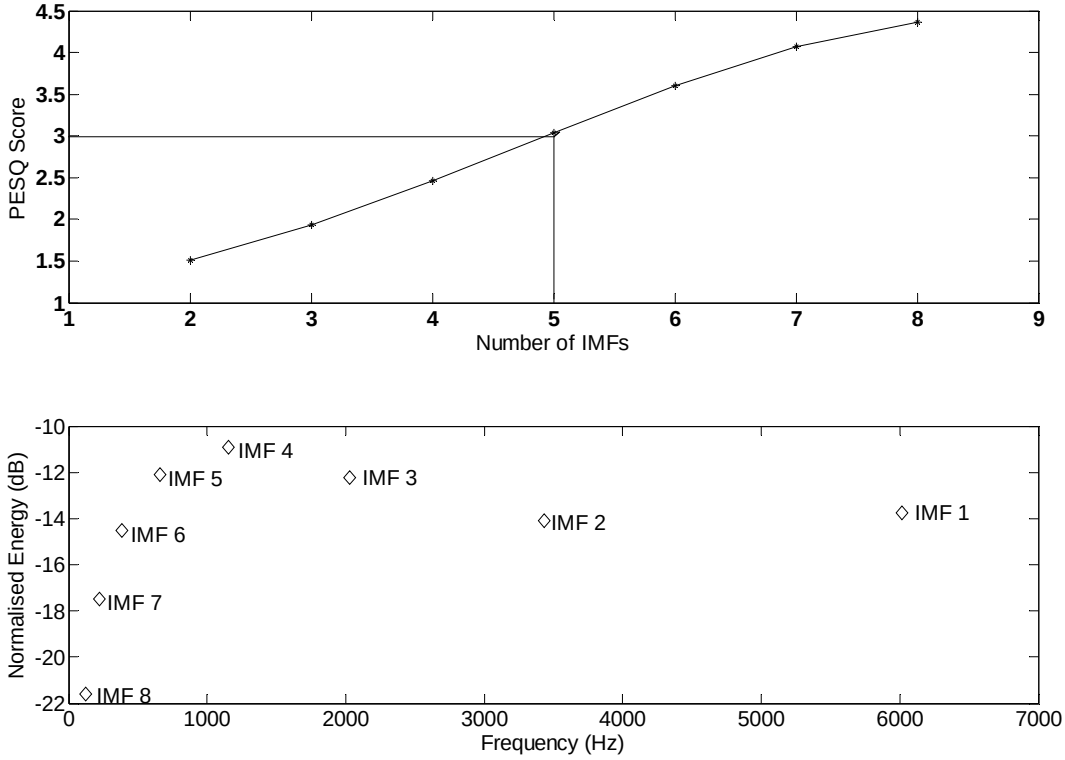

Figure B.2: (a) Average PESQ scores for reconstructed speech using different number of IMFs; (b)Mean Instantaneous frequency and mean energy for first 8 IMFs

Appendix C

Pre-Classification: Self Organising Maps

Pre-classification can identify and separate large clusters prior to emotion recognition within these clusters. An example of a two dimensional feature space is depicted in Figure C.1, showing the distribution of feature vectors. In the space depicted, it can be seen that apart from emotion specific clustering, other clusters are present as well, making emotion classification a complex problem. Identifying these clusters and performing a pre-classification to separate them divides the feature space and simplifies the emotion classification problem within each division (Figure C.2). The example depicted here is exaggerated in its simplicity to illustrate the idea, and in a realistic scenario the reduction in complexity may not be very significant.

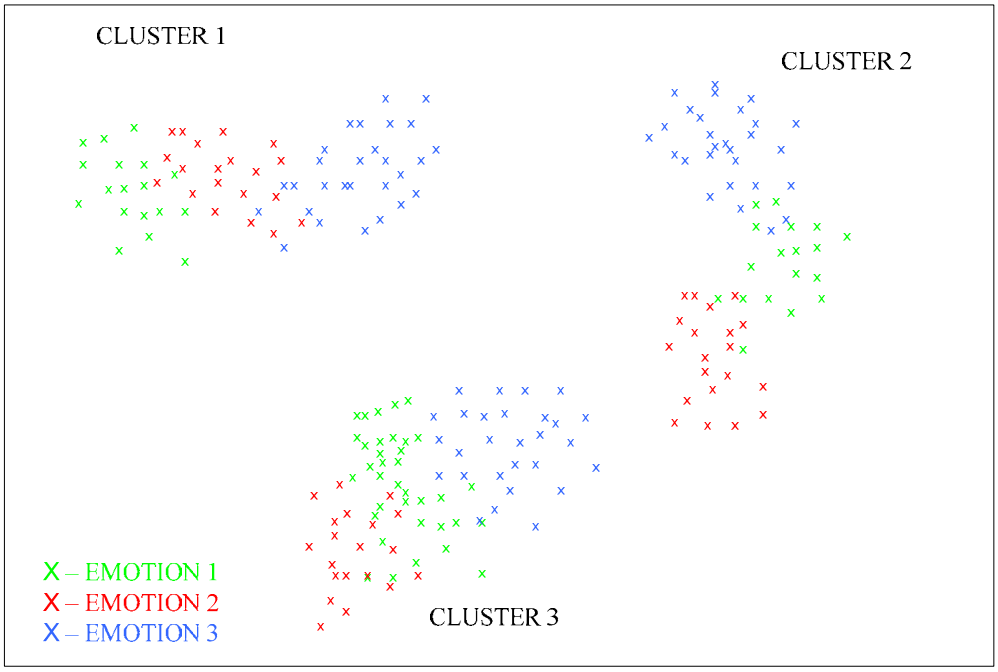

Figure C.1: A synthetic example of a 2-dimensional feature space (not necessarily representative of real data)

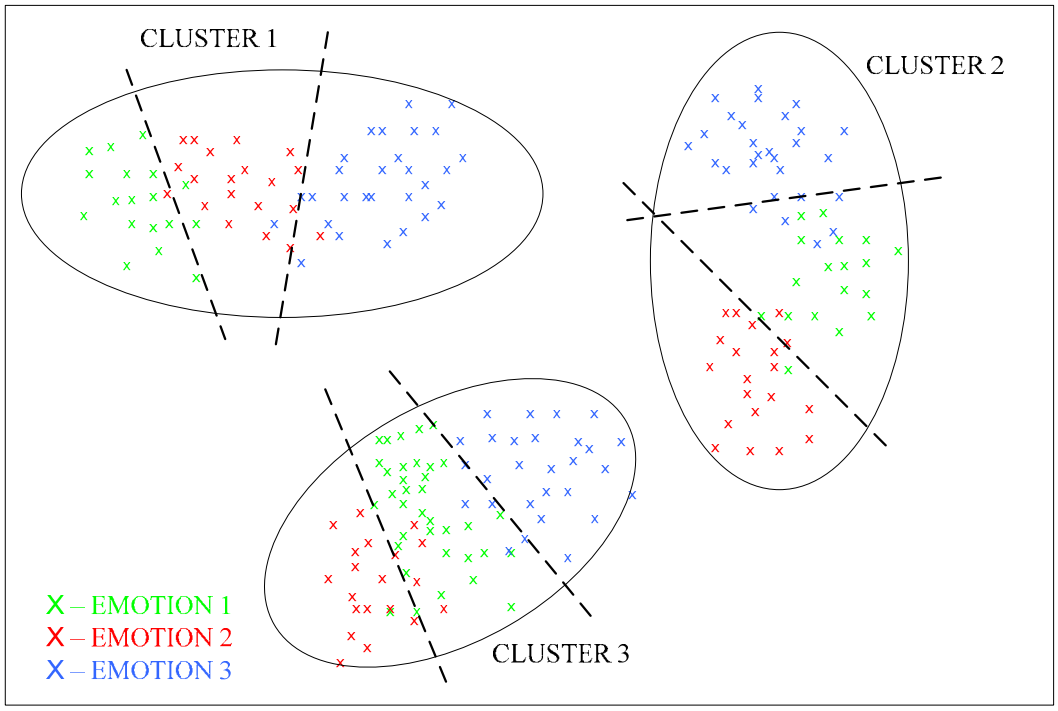

Figure C.2: The 2-D feature space after pre-classification

Multi-layer Kohonen maps have been shown to be useful in clustering applications involving low-level acoustic features with high dimensionality (Wang et al. 2007). In this case, a 3-layer Kohonen map was used to identify clusters from the training data and pre-classify all feature vectors as belonging to one of these clusters. The first layer was a 100 by 50 array of neurons and the second layer was a 40 by 20 array of neurons. The dimensions of the third and uppermost layer were chosen based on the number of clusters required. The inputs to the first layer were the feature vectors themselves while the inputs to the successive layers are the neuron weights (outputs) of the previous layer.

Cluster specific and cluster independent GMM based emotion models were trained in a manner similar to the experiment reported in section 4.3. A comparison of cluster specific and cluster independent classification accuracies should reveal any advantage that could potentially be gained from this type of pre-classification. Since the number of clusters must be fixed prior to training the self organising map (SOM), and since the optimum number was not known, two comparisons were performed with a different

number of clusters each time. In the first case, 18 clusters were chosen (preliminary analyses indicated that choosing more resulted in empty clusters) and the second case used 6 clusters. These AER systems used a front-end comprising of pitch, energy and weighted frequency and the classification tests were carried out in a speaker independent fashion, utilising the speaker normalisation technique proposed in section 4.1.2. Sevenfold cross-validation was used again as in all speaker independent experiments, with a new SOM trained with the training data in each of the 7 repetitions. Both cluster specific and overall accuracies are reported and since feature vectors estimated from different frames of the same utterance (turn) can belong to different clusters, only frame level classification accuracies are reported in this section (once again similar to section 4.3). Table C-1and Table C-2 report the classification accuracies obtained using an 18 cluster and a 6 cluster SOM based pre-classifier.

Table C-1: Cluster accuracies for a system using an 18 cluster pre-classifier (P+E+WF)

Clusters	**Accuracy**		**Clusters**	**Accuracy**	
	Cluster Independent GMMs	**Cluster Specific GMMs**		**Cluster Independent GMMs**	**Cluster Specific GMMs**
1	45.1 %	40.1 %	*10*	40.6 %	24.3 %
2	41.3 %	40.1 %	*11*	32.7 %	24.9 %
3	42.0 %	39.1 %	*12*	44.9 %	37.1 %
4	42.3 %	38.4 %	*13*	47.4 %	47.0 %
5	39.1 %	32.3 %	*14*	48.7 %	41.3 %
6	45.2 %	37.2 %	*15*	48.8 %	41.6 %
7	44.2 %	40.8 %	*16*	47.1 %	35.5 %
8	37.5 %	39.3 %	*17*	37.0 %	29.9 %
9	35.6 %	26.8 %	*18*	42.8 %	42.1 %
			Overall	**43.8 %**	**38.2 %**

Table C-2: Cluster accuracies for a system using a 6 cluster pre-classifier (P+E+WF)

Clusters	Accuracy	
	Cluster Independent GMMs	Cluster Specific GMMs
1	38.5 %	39.7 %
2	36.5 %	32.2 %
3	45.2 %	40.2 %
4	41.0 %	35.3 %
5	43.3 %	36.0 %
6	53.8 %	49.4 %
Overall	**43.8 %**	**40.1 %**

These results indicate that pre-classification offers no advantage to the AER system and in fact it adversely affects overall system performance. However, it can be argued that spectrally richer features would exhibit more definite patterns in the feature space and could take better advantage of the pre-classifier. In order to ascertain if this was indeed that case, the comparisons were repeated with an AER system using a MFCC based front-end and the results are reported in Table C-3 and Table C-4.

Table C-3: Cluster accuracies for a system using an 18 cluster pre-classifier (MFCC)

Clusters	Accuracy		Clusters	Accuracy	
	Cluster Independent GMMs	Cluster Specific GMMs		Cluster Independent GMMs	Cluster Specific GMMs
1	37.2 %	32.7 %	*10*	35.4 %	31.5 %
2	35.2 %	38.5 %	*11*	29.7 %	34.3 %
3	25.4 %	27.4 %	*12*	49.0 %	47.7 %
4	30.9 %	30.0 %	*13*	39.3 %	35.3 %
5	35.1 %	30.0 %	*14*	40.5 %	37.4 %
6	37.9 %	39.0 %	*15*	44.2 %	35.6 %
7	36.9 %	32.4 %	*16*	39.5 %	39.6 %
8	30.7 %	26.8 %	*17*	40.6 %	42.2 %
9	38.4 %	37.3 %	*18*	37.0 %	34.3 %
			Overall	**37.4 %**	**35.3 %**

Table C-4: Cluster accuracies for a system using a 6 cluster pre-classifier (MFCC)

Clusters	Accuracy	
	Cluster Independent GMMs	Cluster Specific GMMs
1	46.8 %	46.2 %
2	35.2 %	32.5 %
3	28.8 %	32.0 %
4	35.8 %	35.2 %
5	35.8 %	34.5 %
6	33.9 %	34.2 %
Overall	**37.4 %**	**37.3 %**

The classification accuracies obtained when using a MFCC based front-end also indicate that pre-classification based on unsupervised clustering offers no advantage to the AER system in terms of overall classification accuracy.

The overall frame level classification accuracies however weight all clusters equally. If the ultimate aim is to classify each utterance (turn), a few frames classified with high likelihood values could outweigh a larger number of frames classified with low likelihood values resulting in the utterance being classified correctly even though a majority of the frames are classified incorrectly (or vice versa). In order to take this into account, turn level classification accuracies were determined by making turn level decisions based on the sum of the likelihoods of the emotions for each frame in the turn (as described in section 3.5). The accuracies thus obtained for all the above mentioned conditions are reported in Table C-5.

Table C-5: Summary of overall accuracies (UTTERANCE/TURN level)

	P + E + WF		MFCC	
	18 Clusters	*6 Clusters*	*18 Clusters*	*6 Clusters*
Cluster Independent GMMs	56.4 %	56.4 %	48.9 %	48.9 %
Cluster Dependent GMMs	53.7 %	54.5 %	49.7 %	51.3 %

The turn level classification accuracies indicate that the cluster specific modelling of emotions provides no benefit when using the low dimensional pitch, energy and weighted frequency based front-end, while exhibiting a small increase in performance for the MFCC based front-end. This may be explained by the larger amount of information contained in the MFCCs as opposed to the other front-end. However, the increase in performance for the MFCC based system is too small to draw any definite conclusions. Moreover, the performance gain is insignificant compared to the loss due to the higher dimensionality and speaker specific information; which contribute to higher variability and consequently, degradation of AER performance.

References

Aertsen, AMHJ & Johannesma, PIM (1980), 'Spectro-temporal receptive fields of auditory neurons in the grassfrog', *Biological Cybernetics*, vol. 38, no. 4, pp. 223-234.

Alku, P (**1991**), 'Glottal Wave Analysis With Pitch Synchronous Iterative Adaptive Inverse Filtering', in *EUROSPEECH-1991*, pp. 1081-1084.

Allen, F, Ambikairajah, E & Epps, J (**2006**), 'Warped Magnitude and Phase-Based Features for Language Identification', in *Acoustics, Speech and Signal Processing, 2006. IEEE International Conference on*, vol. 1, pp. 201-204.

Allen, R & Mills, D (2004), *Signal analysis: Time, Frequency, Scale, and Structure*, IEEE Press, Piscataway, N.J.

Ang, J, Dhillon, R, Krupski, A, Shriberg, E & Stolcke, A (**2002**), 'Prosody-based automatic detection of annoyance and frustration in human-computer dialog', in *Seventh International Conference on Spoken Language Processing*, pp. 2037-2040.

Arnold, MB (1960), *Emotion and personality: (Vol. 1). Psychological aspects*, Columbia University Press, New York.

Averill, J (1975), 'A semantic atlas of emotional concepts', *JSAS catalog of selected documents in psychology*, vol. 5, p. 330.

Banse, R & Scherer, K (1996), 'Acoustic profiles in vocal emotion expression', *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614-636.

Barra, R, Montero, JM, Macias-Guarasa, J, D'Haro, LF, San-Segundo, R & Cordoba, R (**2006**), 'Prosodic and Segmental Rubrics in Emotion Identification', in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I-I.

Baum, LE & Petrie, T (1966), 'Statistical Inference for Probabilistic Functions of Finite State Markov Chains', *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554-1563.

Baum, LE, Petrie, T, Soules, G & Weiss, N (1970), 'A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains', *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164-171.

Borchert, M & Dusterhoft, A (**2005**), 'Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments', in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, pp. 147-151.

Boucouvalas, A & Zhe, X (**2002**), 'Text-to-emotion engine for real time internet communication', in *International Symposium on CSNDSP*, pp. 164-168.

Bower, G (1981), 'Mood and memory', *American Psychologist*, vol. 36, no. 2, pp. 129-148.

Burkhardt, F & Sendlmeier, W (**2000**), 'Verification of acoustical correlates of emotional speech using formant-synthesis', in *SpeechEmotion-2000*, pp. 151-156.

Cabral, J, Renals, S, Richmond, K & Yamagishi, J (**2007**), 'Towards an improved modeling of the glottal source in statistical parametric speech synthesis', in *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany.

Cabral, J, Renals, S, Richmond, K & Yamagishi, J (**2008**), 'Glottal Spectral Separation for Parametric Speech Synthesis', in *INTERSPEECH-2008*, pp. 1829-1832.

Childers, DG & Lee, CK (1991), 'Vocal quality factors: Analysis, synthesis, and perception', *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394-2410.

Cohen, L (1995), *Time-frequency analysis: theory and applications*, Prentice-Hall, Englewood Cliffs, N.J.

Cowie, R & Cornelius, RR (2003), 'Describing the emotional states that are expressed in speech', *Speech Communication*, vol. 40, no. 1-2, pp. 5-32.

Cowie, R, Douglas-cowie, E, Apolloni, B, Taylor, J, Romano, A & Fellenz, W (1999), 'What a neural net needs to know about emotion words', *N. Mastorakis (ed.): Computational Intelligence and Applications. Word Scientific Engineering Society*, pp. 109-114.

Cowie, R, Douglas-Cowie, E, Tsapatsoulis, N, Votsis, G, Kollias, S, Fellenz, W & Taylor, JG (2001), 'Emotion recognition in human-computer interaction', *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32-80.

d'Alessandro, C & Doval, B (**2003**), 'Voice quality modification for emotional speech synthesis', in *EUROSPEECH-2003*, pp. 1653-1656.

Damasio, AR (2000), 'A second chance for emotion', in R Lane, L Nadel & G Ahern (eds), *Cognitive neuroscience of emotion*, Oxford University Press, USA, pp. 12-23.

Darwin, C (1872), *The Expressions of Emotions in Man and Animals*, John Murray, London.

De Gelder, B (2000), 'Recognizing emotions by ear and by eye', in R Lane, L Nadel & G Ahern (eds), *Cognitive neuroscience of emotion*, Oxford University Press, USA, pp. 84-105.

de la Torre, A, Segura, JC, Benitez, C, Peinado, AM & Rubio, AJ (**2002**), 'Non-linear transformations of the feature space for robust speech recognition', in *Acoustics, Speech,*

and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on, vol. 1, pp. I-401-I-404 vol.401.

Dellaert, F, Polzin, T & Waibel, A (**1996**), 'Recognizing emotion in speech', in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1970-1973 vol.1973.

Dempster, AP, Laird, NM & Rubin, DB (1977), 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38.

Douglas-Cowie, E, Campbell, N, Cowie, R & Roach, P (2003), 'Emotional speech: Towards a new generation of databases', *Speech Communication*, vol. 40, pp. 33-60.

Doval, B, d'Alessandro, C & Henrich, N (2006), 'The Spectrum of Glottal Flow Models', *Acta acustica united with acustica*, vol. 92, no. 6, pp. 1026-1046.

Ekman, P (1992a), 'An argument for basic emotions', *Cognition & Emotion*, vol. 6, no. 3, pp. 169 - 200.

Ekman, P (1992b), 'Facial Expressions of Emotion: New Findings, New Questions', *Psychological Science*, vol. 3, no. 1, pp. 34-38.

Fant, G (1960), *Acoustic Theory of Speech Production*, Mouton, The Hague.

Fant, G, Liljencrants, J & Lin, Q (1985), 'A four-parameter model of glottal flow', *STL-QPSR*, vol. 4, no. 4, pp. 1-13.

Fisher, R (1936), 'The Use of multiple Measurements in taxonomic Problems', *Annals of Eugenics*, vol. 7, pp. 179-188.

Flanagan, JL (1960), 'Models for Approximating Basilar Membrane Displacement', *The Journal of the Acoustical Society of America*, vol. 32, no. 7, pp. 937-937.

Fónagy, I (1981), 'Emotions, voice and music', *Research aspects on singing*, vol. 33, pp. 51–79.

France, DJ, Shiavi, RG, Silverman, S, Silverman, M & Wilkes, M (2000), 'Acoustical properties of speech as indicators of depression and suicidal risk', *Biomedical Engineering, IEEE Transactions on*, vol. 47, no. 7, pp. 829-837.

Frick, R (1985), 'Communicating emotion: the role of prosodic features', *Psychological bulletin.*, vol. 97, no. 3, pp. 412-429.

Frohlich, M, Michaelis, D & Strube, HW (2001), 'SIM---simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals', *The Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 479-488.

Hanson, HM (1995), 'Glottal characteristics of female speakers', Harvard University.

Harrington, J & Cassidy, S (1999), *Techniques in speech acoustics*, Kluwer Academic Pub.

Honda, M (2003), *Human Speech Production Mechanisms*,NTT Technical Review.

Huang, NE, Shen, Z, Long, SR, Wu, MC, Shih, HH, Zheng, Q, Yen, N-C, Tung, CC & Liu, HH (1998), 'The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis', *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903-995.

Huang, R & Ma, C (**2006**), 'Toward A Speaker-Independent Real-Time Affect Detection System', in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 1204-1207.

Hui-Ling, L & Smith, JO, III (**1999**), 'Joint estimation of vocal tract filter and glottal source waveform via convex optimization', in *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, pp. 79-82.

Joachims, T (2003), *SVM^{light}: Support Vector Machine*, <http://svmlight.joachims.org/>.

Johnstone, T & Scherer, K (2000), 'Vocal Communication of Emotion', in M Lewis & J Haviland (eds), *Handbook of Emotions*, second edn, Guilford, New York, pp. 220-235.

Katsiamis, A, Drakakis, E & Lyon, R (2007), 'Practical Gammatone-Like Filters for Auditory Processing', *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 4.

Kawahara, H, Masuda-Katsuse, I & de Cheveigné, A (1999), 'Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds', *Speech Communication*, vol. 27, no. 3-4, pp. 187-207.

Klatt, DH & Klatt, LC (1990), 'Analysis, synthesis, and perception of voice quality variations among female and male talkers', *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820-857.

Kohonen, T (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69.

Kohonen, T (1997), *Self-Organizing Maps*, 2nd edn, vol. 30, Series in Information Sciences, Springer, Heidelberg.

Kohonen, T (1998), 'The self-organizing map', *Neurocomputing*, vol. 21, no. 1-3, pp. 1-6.

Kramer, E (1963), 'Judgment of personal characteristics and emotions from nonverbal properties of speech', *Psychological Bulletin*, vol. 60, p. 408.

Kwon, O, Chan, K, Hao, J & Lee, T (**2003**), 'Emotion recognition by speech signals', in *INTERSPEECH-2003*, pp. 125-128.

Lang, KJ, Waibel, AH & Hinton, GE (1990), 'A time-delay neural network architecture for isolated word recognition', *Neural Networks*, vol. 3, no. 1, pp. 23-43.

Lazarus, R (1991), *Emotion and adaptation*, Oxford University Press, New York.

Lee, C, Narayanan, S & Pieraccini, R (**2002**), 'Combining acoustic and language information for emotion recognition', in *Seventh International Conference on Spoken Language Processing*, pp. 873-876.

Lee, CM & Narayanan, SS (2005), 'Toward detecting emotions in spoken dialogs', *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293-303.

Liberman, M, Davis, K, Grossman, M, Martey, N & Bell, J (2002), *Emotional Prosody Speech and Transcripts*, Linguistic Data Consortium, University of Pennsylvania, PA, USA, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28

Litman, D (**2003**), 'Recognizing emotions from student speech in tutoring dialogues', in *Automatic Speech Recognition and Understanding*, pp. 25-30.

Lugger, M & Yang, B (**2007**), 'An incremental analysis of different feature groups in speaker independent emotion recognition', in *ICPhS*.

Makhoul, J (1975), 'Linear prediction: A tutorial review', *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561-580.

Mandelbrot, B (1983), *The fractal geometry of nature*, W. H Freeman and Company, New York.

Maragos, P (**1991**), 'Fractal aspects of speech signals: dimension and interpolation', in *Acoustics, Speech, and Signal Processing, ICASSP 1991, International Conference on*, vol. 1, pp. 417-420.

Maragos, P & Fan-Kon, S (1993), 'Measuring the Fractal Dimension of Signals: Morphological Covers and Iterative Optimization', *Signal Processing, IEEE Transactions on*, vol. 41, no. 1, pp. 108-121.

Maragos, P & Potamianos, A (1999), 'Fractal dimensions of speech sounds: Computation and application to automatic speech recognition', *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1925-1932.

McAulay, R & Quatieri, T (1986), 'Speech analysis/Synthesis based on a sinusoidal representation', *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744-754.

Mertens, P & d'Alessandro, C (**1995**), 'Pitch contour stylization using a tonal perception model', in *Int. Congr. Phonetic Sciences*, vol. 13, pp. 228-231.

Mowrer, O (1960), *Learning theory and behavior*, Wiley, New York.

Mozziconacci, S & Hermes, D (**1999**), 'Role of intonation patterns in conveying emotion in speech', in *14th International Conference of Phonetic Sciences*, pp. 2001-2004.

Mozziconacci, S & Hermes, D (**2000**), 'Expression of emotion and attitude through temporal speech variations', in *Sixth International Conference on Spoken Language Processing*, vol. 2, pp. 373-378.

Murray, IR & Arnott, JL (1993), 'Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion', *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097-1108.

Naylor, PA, Anastasis, K, Jon, G & Mike, B (2007), 'Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm', *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 34-43.

Ohala, J (1983), 'Cross-language use of pitch: an ethological view', *Phonetica*, vol. 40, no. 1, p. 1.

Osgood, C, Suci, G & Tannenbaum, P (1957), *The Measure of Meaning*, Urbana: University of Illinois Press.

Pantic, M & Rothkrantz, LJM (2003), 'Toward an affect-sensitive multimodal human-computer interaction', *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390.

Pearlmutter, BA (1995), 'Gradient calculations for dynamic recurrent neural networks: a survey', *Neural Networks, IEEE Transactions on*, vol. 6, no. 5, pp. 1212-1228.

Pelecanos, J & Sridharan, S (**2001**), 'Feature warping for robust speaker verification', in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, pp. 213-218.

Petrushin, V (**1999**), 'Emotion in speech: Recognition and application to call centers', in *Conference on Artificial Neural Networks in Engineering*, vol. 1, pp. 7-10.

Plutchik, R (1994), *The Psychology and Biology of Emotion*, HarperCollins College Div, New York.

Pollack, I, Rubenstein, H & Horowitz, A (1960), 'Communication of verbal modes of expression', *Lang. Speech*, vol. 3, pp. 121-130.

Rabiner, L & Juang, B (1986), 'An introduction to hidden Markov models', *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4-16.

Rabiner, L & Juang, B (1993), *Fundamentals of speech recognition*, Prentice Hall, New Jersey.

Rabiner, L & Schafer, R (1978), *Digital processing of speech signals*, Prentice-hall Englewood Cliffs, NJ.

Ravuri, S & Ellis, DPW (**2008**), 'Stylization of pitch with syllable-based linear segments', in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3985-3988.

Reynolds, DA & Rose, RC (1995), 'Robust text-independent speaker identification using Gaussian mixture speaker models', *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72-83.

Riegelsberger, EL & Krishnamurthy, AK (**1993**), 'Glottal source estimation: Methods of applying the LF-model to inverse filtering', in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, pp. 542-545 vol.542.

Rilling, G, Flandrin, P & Gonçalvès, P (**2003**), 'On empirical mode decomposition and its algorithms', in *IEEE EURASIP Workshop on Nonlinear Signal and Image Processing*, Italy.

Rosenberg, AE (1971), 'Effect of Glottal Pulse Shape on the Quality of Natural Vowels', *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583-590.

Rosenblatt, F (1958), 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological review*, vol. 65, no. 6, pp. 386-408.

Rui, S, Moore, E & Torres, JF (**2009**), 'Investigating glottal parameters for differentiating emotional categories with similar prosodics', in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4509-4512.

Russell, JA & Mehrabian, A (1977), 'Evidence for a three-factor theory of emotions', *Journal of Research in Personality*, vol. 11, no. 3, pp. 273-294.

Scherer, K (1982), 'Emotion as a process: Function, origin ond regulation', *Social Science Information*, vol. 21, pp. 555-570.

Scherer, K (1984), 'On the nature and function of emotion: A component process approach', in KR Scherer & P Ekman (eds), *Approaches to emotion*, Lawrence Erlbaum Associates, Inc., New Jersey, pp. 293-317.

Scherer, K (1986), 'Vocal affect expression: a review and a model for future research', *Psychological bulletin.*, vol. 99, no. 2, pp. 143-165.

Scherer, KR (2003), 'Vocal communication of emotion: A review of research paradigms', *Speech Communication*, vol. 40, no. 1-2, pp. 227-256.

Scherer, KR, Banse, R & Wallbott, HG (2001), 'Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures', *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp. 76-92.

Schröder, M (2003), 'Experimental study of affect bursts', *Speech Communication*, vol. 40, no. 1-2, pp. 99-116.

Schröder, M (2004), 'Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (Ph. D thesis)', Saarland University.

Schuller, B, Batliner, A, Steidl, S & Seppi, D (**2009**), 'Emotion recognition from speech: Putting ASR in the loop', in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4585-4588.

Schuller, B, Rigoll, G & Lang, M (**2003**), 'Hidden Markov model-based speech emotion recognition', in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 2, pp. II-1-4 vol.2.

Schuller, B, Seppi, D, Batliner, A, Maier, A & Steidl, S (**2007**), 'Towards More Reality in the Recognition of Emotional Speech', in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV-941-IV-944.

Schwarz, P, Matejka, P & Cernocky, J (**2006**), 'Hierarchical Structures of Neural Networks for Phoneme Recognition', in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I-I.

Specht, DF (1967), 'Generation of Polynomial Discriminant Functions for Pattern Recognition', *Electronic Computers, IEEE Transactions on*, vol. EC-16, no. 3, pp. 308-319.

Specht, DF (**1988**), 'Probabilistic neural networks for classification, mapping, or associative memory', in *Neural Networks, 1988., IEEE International Conference on*, pp. 525-532 vol.521.

Talkin, D (1995), 'A robust algorithm for pitch tracking (RAPT)', in W Kleijn & K Paliwal (eds), *Speech coding and synthesis*, Elsevier, New York, pp. 495-518.

Thiruvaran, T, Ambikairajah, E & Epps, J (2008), 'Extraction of FM components from speech signals using all-pole model', *Electronics Letters*, vol. 44, no. 6, pp. 449-450.

Tolkmitt, FJ & Scherer, KR (1986), 'Effect of experimentally induced stress on vocal parameters', *Journal of Experimental Psychology: Human Perception and Performance*, vol. 12, no. 3, pp. 302-313.

Tompkins, S (1962), *Affect Imagery Consciousness-Volume I the Positive Affects: The Positive Affects*, Springer Publishing Company, New York.

Vapnik, V (2000), *The nature of statistical learning theory*, Springer Verlag.

Veldhuis, R (1998), 'A computationally efficient alternative for the Liljencrants--Fant model and its perceptual evaluation', *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 566-571.

Ververidis, D & Kotropoulos, C (2006), 'Emotional speech recognition: Resources, features, and methods', *Speech Communication*, vol. 48, no. 9, pp. 1162-1181.

Vetterli, M & Herley, C (1992), 'Wavelets and filter banks: theory and design', *Signal Processing, IEEE Transactions on*, vol. 40, no. 9, pp. 2207-2232.

Vidrascu, L & Devillers, L (**2007**), 'Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features', in *Paraling2007*.

Vincent, D, Rosec, O & Chonavel, T (**2005**), 'Estimation of LF glottal source parameters based on an ARX model', in *INTERSPEECH-2005*, pp. 333-336.

Vlasenko, B, Schuller, B, Wendemuth, A & Rigoll, G (2007), 'Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing', in *Affective Computing and Intelligent Interaction*, Springer Berlin, pp. 139-147.

Wang, D & Narayanan, S (**2005**), 'Piecewise linear stylization of pitch via wavelet analysis', in *INTERSPEECH-2005*, pp. 3277-3280.

Wang, L, Ambikairajah, E & Choi, E (**2007**), 'Multi-layer Kohonen self-organizing feature map for language identification', in *INTERSPEECH-2007*, pp. 174–177.

Yacoub, S, Simske, S, Lin, X & Burns, J (**2003**), 'Recognition of emotions in interactive voice response systems', in *Eighth European conference on speech communication and technology*, pp. 729-732.

Young, S, Odell, J, Ollason, D, Valtchev, V & Woodland, P (1995), *The HTK book*, Cambridge University.