Spoken Affect Classification: Algorithms and Experimental Implementation

A thesis presented in partial
fulfilment of the requirements
for the degree of
Master of Science
in Computer Science

at Massey University,
Palmerston North, New Zealand.

Donn Alexander Morrison
2005

Para Consuelo, mi fiel furgoneta

Abstract

Machine-based emotional intelligence is a requirement for natural interaction between humans and computer interfaces and a basic level of accurate emotion perception is needed for computer systems to respond adequately to human emotion. Humans convey emotional information both intentionally and unintentionally via speech patterns. These vocal patterns are perceived and understood by listeners during conversation. This research aims to improve the automatic perception of vocal emotion in two ways. First, we compare two emotional speech data sources: natural, spontaneous emotional speech and acted or portrayed emotional speech. This comparison demonstrates the advantages and disadvantages of both acquisition methods and how these methods affect the end application of vocal emotion recognition. Second, we look at two classification methods which have gone unexplored in this field: stacked generalisation and unweighted vote. We show how these techniques can yield an improvement over traditional classification methods.

Acknowledgements

I would like to thank my supervisor, Dr. Ruili Wang for putting faith in me and allowing me to pursue this degree under scholarship. Without this financial help, it would have been unfeasible. His tireless direction and advanced motivational techniques also helped keep my focus.

My co-supervisors, Dr. Liyanage C. De Silva and Dr. Peter Xu, also lended support throughout the research. Their extensive experience was indispensable at times when I needed support.

I also owe gratitude to Pete Morrison of Mabix International for his unique insight into the research. He provided the data used in this research, and without it, it could not be possible.

The second speech database was provided by Tin Lay Nwe of the National University of Singapore. This database was collected and compiled by her and was graciously provided to aid in this research.

And to my partner Jen, who endured countless late nights, filled me with confidence when I lacked it, fed me when I didn't have time to feed myself, bathed me when...well, you get the picture.

Of course, I thank my family, Bill, Debb, Bugs, Michael, Todd, and Jodybird, for their constant support and love, all the way across the Pacific Ocean.

To my postgraduate friends at Massey University, who helped create a fun and relaxed working environment: Cath, Frank, Matthew, Michael, Stefan, and Yiming, among others. And Francis, for always seeming to be in Australia when I needed it the most.

Last, I would like to thank the developers of the many free and open-source software tools I used during the research. Packages such as LaTeX, Gnuplot, OpenOffice, Graphviz, Vim, Octave, donnrisk, Praat, The Speech Filing System, WEKA, The GIMP, Mozilla Firefox, and most of all Debian GNU/Linux were where I spent most of my time this past year and were instrumental in the development and completion of this work.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

With the ever-increasing importance and reliance on computers in our society comes the unnatural burden of interacting with those systems. This increase in human-computer interaction has, in turn, led to a marked increase in research on modelling such systems against human behaviour in an effort to enable more natural interaction. For this to succeed, these systems must have at least a basic level of *emotional intelligence*.

Emotional intelligence is defined by Salovey *et al.* (2004) as having four branches: the perception of emotion, emotions facilitating thought, understanding emotions, and managing emotions. These will be discussed below, with the exception of emotions facilitating thought, as this assumes the ability to think independently, which current computer systems cannot.

The *perception* of emotion is the ability to recognise emotion in oneself and others. These perceptions generally come from three channels: sight, sound, and language or contextual information present in text or prose. For example, a person may recognise that his or her friend feels distraught by the expression in the face or the tone of the voice. The perception of emotion also covers the recognition of emotion in oneself. An emotionally intelligent being is aware of the emotions expressed in itself at any time.

Following perception, an emotionally intelligent being must be able to *understand* emotions and emotional characteristics in order to correctly process and respond to emotional information. This consists of the knowledge of how emotions relate to one another, what causes them, what follows them, etc. Take, for example, a person who becomes angry at him or herself by missing the bus to work before an important meeting. The ability to determine the causes of this anger (e.g., the bus that is missed) is a critical part of emotional intelligence. An emotionally intelligent being will be aware of emotional changes and their nature.

Emotional understanding is a prerequisite for *managing* emotions. An emotionally intelligent being is one that can be open to all types of emotion, reflect on them, manage them in

oneself, and engage, prolong, or detach from an emotional state in oneself or others (Oatley, 2004). A hypothetical situation may involve a doctor tending to a critically injured relative. The doctor must manage his or her emotions in order to operate in an effective manner.

Humans feel most natural communicating with other humans because the extra information conveyed in their emotional expressions can be recognised, processed, and reflected. This information is conveyed through several modes: facial expressions, vocal properties, bodily gestures, and behaviour. This added information helps people understand each other and interact more naturally and efficiently.

The work in this thesis is dedicated to the *perception* of human emotion from the prosodic properties of speech. In other words, this thesis aims to build a system that can capture and interpret the vocal expression of emotion in humans. More specifically, we seek to improve on traditional emotional speech classification methods using ensemble or multi-classifier system (MCS) approaches. We also aim to examine the differences in perceiving emotion in human speech that is derived from different methods of acquisition. For example, how is the perception of acted emotion different from that of spontaneous or naturally occurring emotion?

1.2 Research motivations and applications

There are wide-ranging applications for emotionally intelligent systems in real-world situations. Taking advantage of the emotional information in speech allows more effective processing of the contextual (language) information and a much more natural interaction between humans and machines. The following are some examples of how emotion recognition can yield improvement in the field of human-computer interaction. Figure 1.1 shows the relationships between vocal emotion recognition and potential application areas.

1.2.1 Health and public safety

Situations in which public safety is a major issue would greatly benefit from real-time automatic affect recognition. For example, such a system could be placed in the cockpits of airliners, oceanliners, and buses, where one or two principal operators control the fate of the vessel. These systems would be used to detect pilot boredom, inattention, or fatigue (Pantic and Rothkrantz, 2003). In private vehicles, detection of anger could reduce incidents of road rage by alerting the driver and trying to make them aware of the situation (Fragopanagos and Taylor, 2005).

Affect recognition could avoid concerns of having observers constantly monitoring or recording in situations where security or safety is of concern. For example, in hospitals, closed-circuit security systems, prisons, etc. (Pantic and Rothkrantz, 2003). These systems could alert personnel to certain situations such as disputes, accidents, riots or fighting.

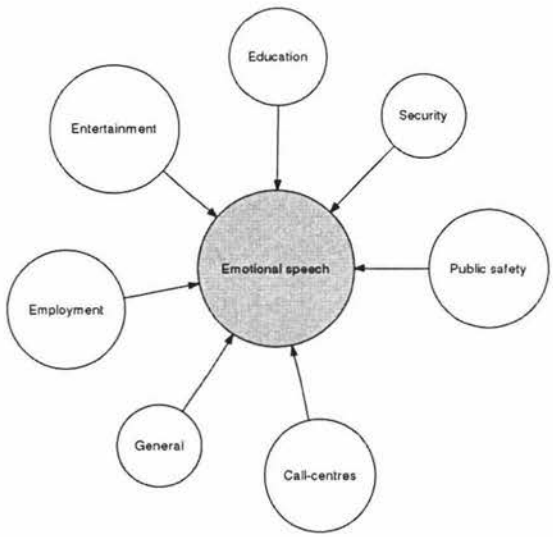


Figure 1.1: Applications of vocal emotion recognition

1.2.2 Education

Perception of human affect is important in areas where subjects are being taught or instructed. Human teachers can recognise student boredom, fatigue, and confusion and are then able to take steps to revive attention levels, or perhaps terminate the instruction if too many students are unable to process effectively.

Emotion and affect recognition from speech would be beneficial in an automated tutoring environment. The system could determine the affective states of the students and depending on how well they appear to be learning, or based on feedback (levels of frustration, confusion, boredom, fatigue, etc.), adjust the rate at which the information is presented to make the learning as efficient as possible (Picard, 1997).

1.2.3 Fraud and crime prevention

Voice profiling is directly related to vocal affect recognition. Voice profiling aims to classify speech samples according to predefined psychological profiles. These profiles can be generated or trained on pathological examples.

The use of voice profiling for fraud detection can be a useful measure to reduce the number of fraudulent insurance claims for insurance companies. The time needed to process claims can be reduced if claims that are potentially fraudulent are eliminated early on in the process. A system could be easily developed that allows claimants to provide information about their claim over the telephone with a disclaimer stating that their voice profile will be analysed for signs of

fraud. If the analysis comes back positive for possible fraud, the customer can be notified of the result and offered an opportunity to retract their claim without penalty. Such a system does have obvious drawbacks, for example people may be discouraged from submitting a valid claim over fears of a false-positive from the voice profile analysis.

Another practical use of voice profiling would be for police and security in interviewing suspects for criminal cases. Suspects could be interviewed and their speech analysed by profiling software that could detect pathological patterns correlating to lying or nervousness. As with the above scenario, however, there are many ethical issues relating to this application and its output would have to be used only as one of many sources of information during interrogation.

1.2.4 Leisure and entertainment

An area ripe for new applications of emotion perception is that of leisure and entertainment. Here, the technology is applied in anecdotal ways. An example is the Sony ERS-7 Aibo Entertainment Robot. This robotic pet dog learns from interaction with its "owner" and can express different emotional states.

Computer video games are the result of billions of dollars of research and development investment aimed at making the player feel like he or she is experiencing reality. Emotion detection and synthesis in these games could greatly improve the gaming experience. Online games such as Everquest where human players interact with other human and computer players can benefit from both emotion recognition and synthesis to enhance the experience. Interaction with computer characters is often unnatural due to the lack of emotional understanding on the part of the computer character. Adding an affective element to these characters would introduce an entire new level to the gaming experience, providing a much more natural environment that would more closely model reality. This can be accomplished by integrating speech and facial expression recognition using cameras and microphones to measure the human player's affect. This affect can then be transmitted to other human or computer players in the game (Nakatsu *et al.*, 1999).

The research of Breazeal and Aryananda (2002) has primarily focused on the integration of a multi-modal emotion classification system in a robot. This robot, named Kismet, responds to caretakers by way of sight and sound. An integrated affective intent classification system allows the basic recognition and modelling of primary emotions. The robot approximately models an infant that responds to affirmation, prohibition, attention and soothing. After more research, this could be extended to a more full set of emotions or affective states allowing the robot to interact naturally with human operators.

1.2.5 Employment

Voice profiling can help streamline the processing of job applicant interviews. By interviewing applicants through an automated telephone system, the responses can be analysed for specific qualities which can be mapped to different positions within the company. For example, if a company is screening applicants for job openings in multiple departments, e.g., sales or customer support, the applicants can be automatically sorted into groups based on how their voice profile fits the target profile for each category. Job positions where an employee is constantly interacting with customers may require specific voice qualities. An applicant with a monotone pitch contour can be screened out automatically, and an applicant with a melodic pitch contour can be placed in a sales category for further inspection. Such a system would not be designed to completely take the place of human interviewers, but can greatly reduce the time requirements for selecting candidates.

1.2.6 Call-centres

Last, we look at applications of emotionally intelligent systems in call-centres. This is the primary focus of the end result of this research. Call-centres often have a difficult task of managing customer disputes. Ineffective resolution of these disputes can often lead to customer discontent, loss of business and in extreme cases, general customer unrest where a large amount of customers move to a competitor. It is therefore important for call-centres to take note of isolated disputes and effectively train service representatives to handle disputes in a way that keeps the customer satisfied (Petrushin, 2000).

Additionally, a team lead or manager may want to inquire on the status of any currently active calls in order to help coach new or inexperienced CSRs. Additionally, a manager can use the information provided by a spoken affect recognition system in several other ways. First, if such a system is deployed with each CSR, then a manager or senior member of staff can preview the emotional states of every caller at once, having an "overview" snapshot in real time. Other uses include the generation of statistics on the number of angry or upset callers each CSR has or whether any CSRs are being angry at the customers. This can lead to action to correct this behaviour or find things that a CSR can improve on and in turn help the call-centre more effectively manage the customer base.

Automated telephone systems are another potential application area that humans find themselves interacting with more and more. These systems have speech recognition units that process user requests through spoken language. A spoken affect recognition system can help process callers according to perceived urgency. If a caller is detected as being angry or confused in the automated system, their call can be switched over to a human operator for assistance. This could be particularly useful for the elderly who can often be disoriented when interacting with

automated telephone systems. Petrushin (2000) built a system to monitor voice-mail messages in a call-centre and prioritise them with respect to emotional content. Such systems can make interaction with automated call-centres more efficient and less daunting.

1.3 Methodology

In this section we present the methodology followed during the development of this thesis. Figure 1.2 shows a flow diagram describing the methodology. Because the research focus is primarily a classification problem, that being the classification of different emotions, the methodology followed is much like any other classification problem. The first step is a review of the literature relevant to the field. Previous research on automatic emotion recognition was surveyed to build a knowledge of the state of the art.

Once a general knowledge of the state of the art was achieved, data had to be collected. Fortunately, a natural speech database was provided through the partner company for this project. A second speech database was collected from a previous study on emotion research (Nwe, 2003). Unlike the natural set, this database used actors and actresses. This provided a way to compare the classification methods on different types of data as well as investigate inherent differences between the two datasets. To gain a ground truth on the natural database, a system was developed to allow human listeners to judge the emotions present in the database.

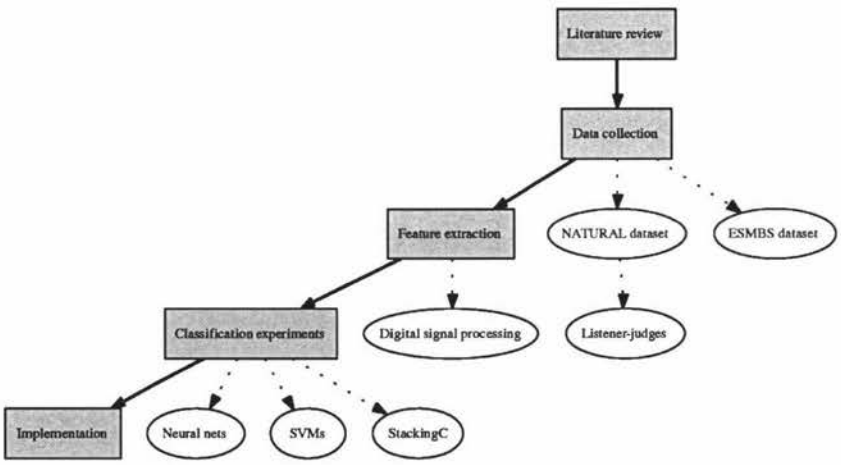


Figure 1.2: A flow diagram of the methodology followed for this thesis.

Next, characteristics of emotional speech from the existing literature were reviewed. Prominent psychologists such as Klaus Scherer who have explored emotion research for many years provide a strong basis for this area. These characteristics were extracted and compiled into feature vectors. These feature vectors describe the most relevant characteristics of emotional

speech. Briefly, these include the fundamental frequency, energy, and formant frequency contours as well as features relating to rhythm such as the rate of speech.

Classification algorithms were then reviewed. As a starting point, artificial neural networks were experimented with, as they have proven quite useful in previous studies. These are subsequently improved upon using support vector machines. Feature selection techniques such as forward selection, genetic search, and principal component analysis were compared to reduce dimensionality in the feature space.

We then tested novel ensemble classification approaches in this field of using stacked generalisation and a simple voting scheme. Stacked generalisation takes as input base-classifier predictions and target classes and attempts to predict when the base-classifiers are incorrect. The voting scheme takes the predicted classes from each base-level classifier and determines the class with the greatest popularity.

The last step was to build an implementation of the theoretical system. This took all previous steps, the algorithms for endpoint detection, feature extraction, the use of the feature selected sets, and classification and brought them together into a single, modular system. This application reads input from a microphone or WAVE file and outputs a prediction based on the recorded speech sample. A modular artificial neural network functions as a plug-in to facilitate efficient replacement. Figure 1.3 shows the data flow for the emotion recognition system.

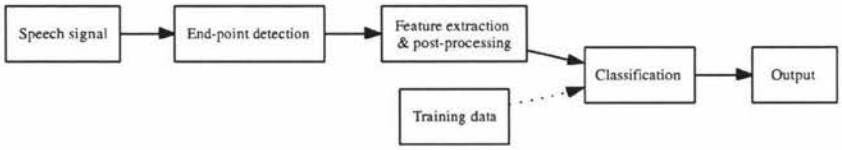


Figure 1.3: A data flow diagram of the real-time emotion recognition system.

1.4 Structure of the thesis

This thesis is organised as follows. In Chapter 2, a brief history of emotion research and theoretical representations of emotion are presented. This chapter also introduces the expression of emotion in humans and lists previous work in automatic spoken emotion recognition. Some areas which require additional attention are defined.

Chapter 3 presents the three data acquisition methods that are applied to vocal emotion research. Next, the two emotional speech datasets used in this research are introduced. The first database is collected from a call-centre and consists of natural interactions between humans. The second database is collected from non-professional actors and actresses. The advantages and disadvantages of each collection method and how it affects the research are discussed in

detail.

Different emotions induce different physiological changes in the body, which in turn directly affect prosodic patterns in speech. Chapter 4 formalises and reviews correlations and characteristics of emotional speech.

Building on Chapter 4, Chapter 5 explores features chosen to describe emotional content contained in speech. These features are taken from previous research and experimental features based on the formant frequencies are investigated.

Chapter 6 introduces several classification algorithms used in this research. These algorithms are compared against each other in an attempt to reveal the most efficient and suitable candidate for use in the system. Feature selection methods are also compared. Next, we introduce two ensemble techniques: stacked generalisation and unweighted vote.

Chapter 7 presents the experimental results based on the classification and feature selection algorithms described in Chapter 6. This chapter also offers an in-depth look at the building of a prototype emotion classification system. The system is developed using existing algorithms and is brought together using C and C++. It functions in real-time and performs automatic classification via a modular artificial neural network.

Finally, Chapter 8 presents a conclusion and directions for future work.

Chapter 2

Foundations and background

2.1 Introduction

This chapter presents the various theoretical arguments that span the field of psychology with respect to the study of emotion and provides the foundation on which the remainder of the thesis rests.

First, a short history and discussion on the two major ideologies of emotion research are presented. These are the evolutionary (supported by the Darwinians) and the culturally or socially learned (supported by the social constructivists). Next, the two prevailing theoretical representations of emotion are introduced. These are the discrete and the dimensional emotion theory. Both have advantages and disadvantages, and these are explained in detail.

Different representations of the discrete emotions are then given. We show how even within the discrete theology there is disagreement over how to organise the emotion classes. Next, we begin to investigate the expression of emotion in humans, starting with the various channels of expression and finishing with an analysis of the expression of emotion through speech. Last, previous research in automatic emotion recognition is introduced, followed by areas which are in need of improvement.

2.2 A brief history of emotion research

The study of human emotion has been present in many disciplines for hundreds of years, spanning psychology to linguistics to philosophy (Weigand, 2004). From the psychological perspective, most notable was Darwin's work *The expression of the emotions in man and animals* (Ekman, 1973; Cornelius, 1996; Picard, 1997). His collection of observations and theories sparked much interest in the subject. Darwin developed the theory that the expression of emotion has roots deep within the evolutionary chain (Darwin, 1872/1965). According to Darwin, every emotional expression results from a period when these expressions were used in ways key to

survival. He argued that seemingly involuntary movements such as the squeezing of a fist or clenching of teeth when angered trace back to when these actions meant something more than just an expression of anger; they determined life or death. The key issue here is that if there are biological roots to the expression of emotions, it can be deduced that all humans, regardless of culture, share the same basic emotional constructions.

Since Darwin's time, much has changed in the study of psychology, but his theories have provided a major foundation for further research. Contemporary Darwinians continue to find new evidence that support his theories. For instance, one of the foremost researchers in facial emotion expression, Paul Ekman, studied the recognition of facial expressions in the 1970s and 1980s. He discovered that emotional facial expressions were generally recognised cross-culturally. In one notable study, Ekman presented pictures of emotional expression in human faces to members of an isolated tribe in New Guinea. These people had no previous contact with Western culture. When asked about the photographs depicting facial expression, they chose the correct emotion classes with accuracy significantly higher than chance. Further, when asked to make emotional expressions in their faces according to certain situations, the expressions they made concurred with those expressions seen in Western cultures (Ekman, 1973).

In the 1980's, a new perspective began to formalise. Theorists sharing this view call themselves the "social constructivists." They claimed that emotions were learned through social and cultural means, directly contradicting the Darwinian theorists (Averill, 1980). Social constructivists argue that emotions originate completely by social construction; that is to say that the emotions are learned, and are not the result of evolution.

The different ideologies of emotion can be grouped into two general categories: the biological and evolutionary theories introduced by Darwin and supported by (Ekman, 1973; Damasio *et al.*, 2000; Izard, 1977), and social learning theory, supported by the social constructivists (Averill, 1980; Cornelius, 1996). Ultimately, it is more realistic that there exist a combination of these two ideologies. This view is shared by Elfenbein and Ambady (2002) who conducted a large survey of cross-cultural emotion research and concluded that "certain core components of emotions are universal and likely biological", however, "culture can have an important role in shaping [...] emotional communication" (Elfenbein and Ambady, 2002).

Other psychologists have also argued in support of this. Matsumoto (1989) states that culture and social environment play a major part in the construction of emotional intelligence even though the basic emotions stem from evolutionary origins. Murray and Arnott (1993) suggest that there are significant similarities of basic emotional characteristics "between all people, with cultural differences in some secondary emotions."

2.3 Theoretical representations of emotion

Allowing for aspects from both the Darwinian and social constructivist traditions, we now focus on how emotions are represented. The need arises for a representation that can eventually be mathematically modelled.

The question presents itself: How are we to represent different emotions? "Emotions are indeterminate concepts. They are describable to some extent, but cannot be defined. [...] There are no fixed, strict definitions, only approximations, continua, probabilities, approaches to detertimateness within indeterminateness" (Weigand, 2004). What Weigand is saying is that emotions are difficult to categorise and thus it is difficult to describe and come to an agreement about characteristics of different emotional states. This is essentially one of the long-trodden disagreements in psychology.

There exist several theoretical representations of emotion in the literature, but we will discuss the two most popular. These are the *discrete* and the *dimensional* emotion models (Izard, 1972; Scherer, 2003).

2.3.1 Discrete emotion theory

Discrete emotion theory observes a set of distinct and separate emotional classes which can be distinguished by the different modalities by which they are expressed, (e.g., facial expressions, vocal characteristics, body gestures, etc.). Those that follow this theory include Ekman (1992), Izard (1977), Damasio (1994).

Typically, this set of basic emotions often comprises anywhere from four or five to fourteen distinct emotions (Scherer, 2003). There is some disagreement between psychologists over what this basic set consists of, but generally it consists of six, the so-called prototypical emotions: *anger, happiness, sadness, fear, disgust*, and *surprise*. These basic emotions are often only the primary branches of a tree that contains secondary and tertiary emotion classes. These branches will be discussed in a following section. Figure B.1 in Appendix B shows an example tree containing the primary, secondary, and tertiary emotions defined by Parrot (2001).

The advantages of following this representation are that many previous studies have followed this approach and therefore it is easier to compare results from different studies. Another advantage is that it is much easier to build a mathematical model based on discrete classes, as we will explain below.

2.3.2 Dimensional emotion theory

Dimensional theory essentially allows emotions to be represented in non-discrete ranges along linear axes. Emotional states are represented within a two-dimensional space with one axis representing the positive-negative range and the other representing the degree of arousal. Scherer

(2003) states that sometimes a third dimension is added which defines power or control.

Dimensional emotion theory was first introduced by Spencer (1890) in his model of pleasantness-unpleasantness (P-U) and subsequently extended by Wundt (Wundt, 1896). Wundt added two new dimensions: excitement-quiet and tension-relief. Later, Woodsworth (1938) formalised the model further by likening P-U to acceptance-rejection (Izard, 1972). Woodsworth is generally credited for laying the foundation for further exploration into dimensional emotion theory (Izard, 1972).

Several recent studies on emotion detection use dimensional theory to represent emotional states as a way of avoiding the dependence on discrete emotion groups (Bachorowski, 1999; Zagalo *et al.*, 2004).

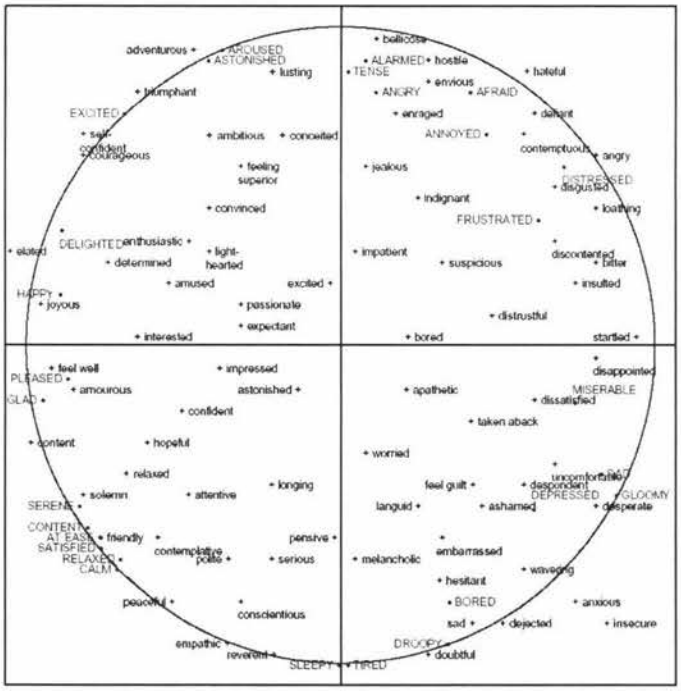


Figure 2.1: The dimensional representation of emotion (from (Scherer, 2001))

The argument for this representation of emotion is that it allows for the modelling of different levels of emotion, which are difficult to describe using discrete theory. It also does not limit the different types of emotion that can occur, as the discrete representation does, because a seemingly infinite number of ranges over the two (or three) dimensions is possible.

2.3.3 Summary of theoretical representations of emotion

Defining types of emotions is difficult. It is relatively easy to infer when a person is angry from the way they act, but there are many different types of anger and of varying levels of intensity. There is hot anger, which can have attributes such as the flailing of the arms and yelling (Darwin, 1872/1965). There is also cold anger, which can have the same undertones but is much more subtle. But where is the line drawn? The same is true for any emotion. This is the problem with categorising emotion. Picard (1997) agrees that there is no clear categorisation of emotions and Wierzbicka (1985) states that "there are countless emotions that can be perceived as distinct and recognisable", which may be true from a psychological perspective, but ultimately we have to distinguish basic emotions which can be more easily identified in mathematical terms.

Therefore, it follows that the dimensional representation, while simplifying theoretical research, only serves as an abstraction with respect to applied theory and that any useful application will only have to draw mappings from this abstraction into some discrete emotion space. It is based on this that we follow the convention of using discrete emotional states rather than a multi-dimensional, non-deterministic representation.

2.4 Defining the emotion classes

Following the choice of discrete emotional theory, the next logical step is to define which classes of emotion we are interested in studying. An important factor here is that we must choose a model that covers a majority of the basic emotions expressed in humans. For example, anger, joy, sadness, etc.

For studies in discrete emotion recognition, where a specific group of emotion classes are compared, a fundamental issue is the choice of which discrete emotion classes to include. There are several different representations presented in the literature. In this section we will describe some of the more popular representations, which one we are using, and why.

Emotion theorists have long disagreed over which, if any, basic emotions exist. Because we are discussing basic emotions, these theorists are those who adhere to the discrete representation of emotion, described in the previous section.

2.4.1 Primary, secondary, and tertiary emotions

Some theorists argue that emotions are best described based on their prominence within the human psyche (Damasio *et al.*, 2000; Danes, 2004). These are described by Damasio as the primary, secondary, and tertiary emotions. The primary emotions are the most visible and easily differentiated: happiness, sadness, fear, anger, surprise, and disgust. Secondary emotions are less obvious externally and include embarrassment, jealousy, guilt. Third, Danes (2004) de-

fines "background emotions" and gives Damasio's example of "well-being or malaise, calm or tension." These are the tertiary emotions and they have a much longer cycle compared to the primary or secondary emotions. Tertiary emotions describe how one may feel for longer periods of time. Nwe (2003) discusses this as well, pointing out that emotions can have a "narrow sense effect," where a person may feel a certain way for only a few seconds, or "broad sense effect," where these feelings can last for months, even years (Nwe, 2003).

A "palette theory" of emotions, whereby secondary and tertiary emotions can be created by mixing the basic or primary emotions has been suggested, although this has not been widely accepted in the psychological community (Murray and Arnott, 1993).

2.4.2 Primitive emotions

Some psychologists have come to an understanding on a set of primitive emotions. These emotions are those which can easily be attributed to the evolutionary theory on emotions (Deigh, 2004). These emotions are seen to be shared between both humans and animals. For example, fear, anger, or delight are primitive emotions.

The difference here is that there is an obvious boundary between an intentional state of mind (cognitive) and a primitive emotion. Deigh (2004) finds fault in the contemporary theories on emotion using the primitive emotions. Cognitive theorists cannot easily account, Deigh says, for the primitive emotions "if one considers the thought content of every intentional state to be a proposition," where the proposition is the representation of that thought in a language. The primitive emotions lack this underlying proposition by short-circuiting the need for this thought.

2.4.3 The basic emotions

Both of the above have some degree of overlap as there are primary emotions which are also shared by the primitive emotions. Furthermore, the primary emotions and the primitive emotions are made up of emotions which are argued by many theorists as *basic* (Plutchik, 1970; Izard, 1977; Ekman, 1992; Oatley, 2004). The basic emotions are defined as those which are generally displayed and recognised cross-culturally. In other words, emotions which are common in humans of different culture.

There is some lack of agreement over exactly which emotions should be categorised as the basic emotions (Scherer, 2003). For example, Plutchik (1970) gives anger, disgust, joy, fear, sadness, surprise, anticipation, and acceptance as basic emotions whereas Oatley (2004) describes only five basic emotions: anger, disgust, anxiety, happiness, and sadness.

Izard (1977) defines the basic emotions as anger, joy, shame, surprise, contempt, disgust, distress, fear, guilt, and interest. These are categorised into different levels, based on duration and the origin of activation. Emotions lasting for a short period of time are said to be brought

on by external events: interest, excitement, surprise, startle, contempt, fear, and disgust. State triggered emotions, which last longer than event triggered emotions, are based on how one is perceived and judged by one's peers. The emotions at this level are happiness, sadness, distress. Izard goes on to describe the third level, which are triggered by the perceptions of a person's own actions and include pride, guilt, and shame. It can be seen that the basic emotions span the first two levels of Izard's model. Ekman (1992) studies the most popular group of basic emotions: anger, disgust, happiness, fear, sadness, and surprise (McGilloway *et al.*, 2000; Nwe *et al.*, 2003b; Lee *et al.*, 2004; Nakatsu *et al.*, 1999).

2.4.4 Summary of emotion classes

Despite some disagreement in psychology circles, it has been widely accepted that there exist a specific set of basic emotions. The most common are *anger, happiness, sadness, disgust, fear,* and *surprise*. A seventh state, *neutrality*, can also be considered a baseline for the other emotions (Ekman, 1992; McGilloway *et al.*, 2000; Nwe *et al.*, 2003b; Lee *et al.*, 2004; Nakatsu *et al.*, 1999).

In this thesis, we take the widely supported theory that there exist a basic set of six emotions, with a seventh neutral state denoting a natural or unaroused emotional state. We will follow Ekman's model of the six basic emotions. These basic six are also common between those groups with wider variation defined by Plutchik, Izard, and Oatley.

2.5 Emotional expression in humans

People express emotion in everyday communication and it is obvious that emotion is a very important part of everyday life (Weigand, 2004). In social situations, emotional expression "plays a regulatory role" in groups of individuals and often leads to more efficient "conflict resolution strategies" (Scheutz and Schermerhorn, 2004). People who are unable to express or recognise emotion via some cognitive defect or injury have difficulty dealing with certain aspects of daily life. Picard (1997) gives us an example of a person who suffered a brain injury and was unable to feel or express emotion. Having this deficiency caused enormous problems in his life. For example, he lacked the ability to learn from poor decisions because there were no emotions or feelings associated with the memories. Also, people who exhibit little or no emotion are often viewed with suspicion or distrust by others (Dellaert *et al.*, 1996), because as emotional beings, humans feel most comfortable communicating with other emotional beings (Picard, 1997). By perceiving, understanding, and managing emotions, humans can interact on a higher level, as more information is transmitted through the interaction.

2.5.1 Channels of expression

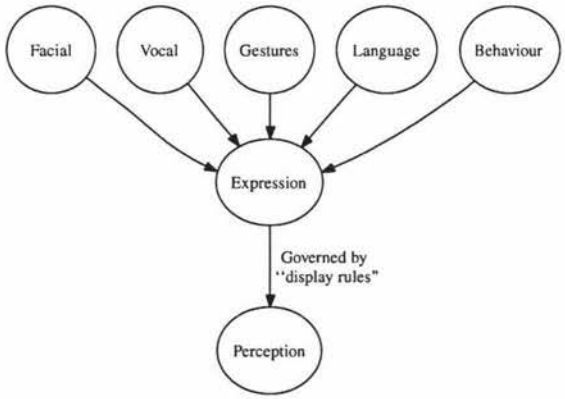


Figure 2.2: Channels of expression and their relation to perception in humans

> "*Preventing others from eating rotten food by emitting disgust signals may be most useful to those eating at the same place as the signaler. In this context, facial expressions of disgust are most adequate. [...] In contrast, there is clear adaptive advantage in being able to warn friends (in fear) or threaten foes (in anger) over large distances, something for which vocal expression is ideally suited.*" (Scherer, 2003)

As the above quote states, the channels through which emotions are most reliably transmitted are the result of a biological process that saw functions such as these, which were at one time important for survival, carried down the evolutionary line. Following this, there are several ways in which emotions can be expressed. The first and most recognisable form is through facial expressions. For example, when a person displays happiness or joy, their eyebrows raise, their eyes widen vertically, and their mouth widens into a smile.

The second is through body gestures. Movements of the head, hands, and bodily posture transmit emotional information to other people. An example of this is someone who is frightened. This person may cower or cover their head or face instinctively. An angry person may throw their arms in the air or move them erratically.

Another expressive channel is simply through language information. This can take many forms including spoken word, written text or other communicative means where a language or symbols of a language are used. When people want to get their emotional state across, they will state it by choosing specific words that describe this. For example, "I am very annoyed at you today."

General behaviour patterns are another channel of emotional expression in humans (Plutchik, 1970). While this may seem to closely resemble expression in body gesture, behavioural expression typically takes on a longer term effect. For instance, a grieving mother may mourn and cry for days or weeks at the loss of a child.

The last channel of emotional expression is through vocal cues. Research in vocal expression of emotion has centred on several properties of speech that separate different emotional states. These include changes in energy (volume), pitch (fundamental frequency and tone), voice quality, and speaking rate.

Pantic and Rothkrantz (2003) talks about affect recognition in humans and how there are multiple simultaneous channels of "modality", for example, sight, touch, hearing. "A mouth expression that might be interpreted as a smile will be seen as a display of sadness if at the same time we can see tears and hear sobbing."

2.5.2 Ekman's display rules

To explain the cross-cultural variation in the expression of emotion, Ekman and Friesen (1969) introduced the theory that there are culture-specific "display rules" that govern and regulate this expression. Essentially, display rules can be thought of as a filter through which emotional expression is passed. Depending on the culture or social context, the underlying emotional expression is modified to be more acceptable to those who perceive it.

In a notable study by Ekman, Japanese and American subjects were shown images depicting graphic content and peaceful content. When alone and recorded secretly, both the Japanese and American subjects displayed uneasy facial expressions during the graphic content and neutral or happy expressions during the peaceful content. However, when an observer was placed in the room, the Japanese subjects forced a smile for the graphic content, while the American subjects displayed the same uneasy or disgusted expression.

In another study, Friedman and Miller-Herringer (1991) examined emotional expression in people playing a game. Victors were more likely to hide their expression of joy if their opponent was present, but when absent, the victors were more likely to display joy uninhibited.

These examples are evidence that display rules do exist, and these social customs shape the way emotions are expressed in humans.

Display rules may present a problem to researchers in the recognition of emotion. If a person is trying to hide their internal state, then their current expression is not valid. However, it can be argued that every expression is a portrayal to some extent, and therefore is an adequate representation of what the portrayer wants others to perceive (Scherer, 2003).

2.5.3 The human speech production aparatus

Before we more closely examine emotional expression in speech, we should first introduce the basic principles of speech production. The sounds that compose speech are created by forcing air from the sub-glottal area over the vocal chords or glottis which causes them to vibrate. Then, depending on the position of the velum, the air is then pushed through either both the nasal and vocal tracts (if the velum is opened) or only the vocal tract (if the velum is closed). It is here that resonant frequencies are created, based on the shapes of these tracts. Finally, the air is pushed out through the nose and mouth where it's properties are again changed depending on the shapes of the openings (Rabiner and Schafer, 1978; Rabiner and Juang, 1993). Figure 2.3 shows an X-ray of the human speech production system and Figure 2.4 shows a simplified schematic model of the human speech production system.

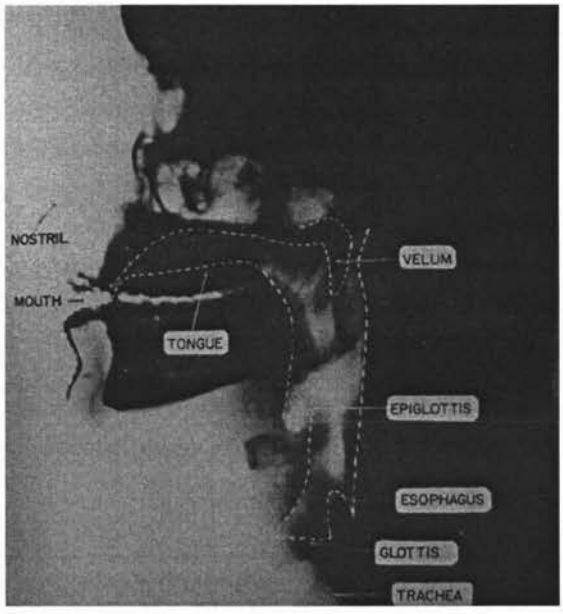


Figure 2.3: A cross-sectional X-ray of the human speech system (from (Flanagan *et al.*, 1970)).

According to Rabiner and Schafer (1978), there are three distinct types of sounds produced by the human speech production system. The first are *voiced* sounds. These are produced at the glottis by pushing air up from the subglottis area over the "relaxed" vocal chords such that "quasi-periodic pulses of air" are sent through the vocal and nasal tracts. This produces many of the common vowel sounds. Rabiner lists these as /U/, /d/, /w/, /i/, and /e/. Next are the *frictave* or *unvoiced* segments. These are created by opening the glottis such that little or no voiced sound is produced, and forming the lips or teeth into a barrier such that when the air passes through this

constriction it creates a "broad-spectrum" sound, with a resulting waveform resembling random noise. Rabiner gives /*S*/ ("sh") as an example of a frictave segment.

Last are the *plosive* speech segments. Plosive segments are created by "making a complete closure" somewhere in the vocal tract. This is usually at the end of the vocal tract, namely the front of the tongue or the lips. /*tS*/ ("ch") is provided by Rabiner as an example of a plosive sound segment.

Together, these three groups of speech sounds make up the large number of phonemes used in all language. Rabiner notes that American English contains 42 unique phonemes.

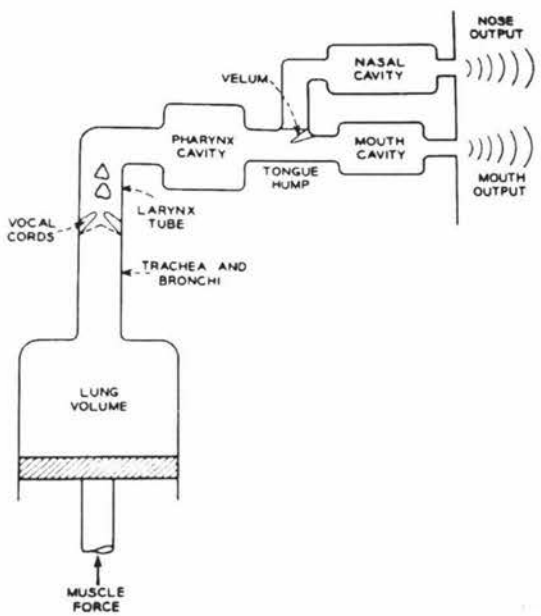


Figure 2.4: A schematic model of the human speech production system (from (Flanagan, 1972)).

During speech production, there are two channels of information conveyed to the listener. The first is the verbal content, containing contextual information encoded in the language of the speaker. The second is the vocal content, which comprises intonation, voice quality, and prosody (Murray and Arnott, 1993).

Scherer (2003) states that there is an "assumption," with regards to the differences of emotion represented in speech, that "emotional arousal [...] is accompanied by physiological changes that will affect respiration, phonation, and articulation in such a way as to produce emotion-specific patterns of acoustic parameters."

2.5.4 Physiological responses to the emotions

During emotional stress, the human body undergoes physiological changes that directly affect function and control. Two regions of the autonomous nervous system, the sympathetic and parasympathetic systems, automatically regulate bodily changes in response to different emotional stimulus (Cornelius, 1996). These two subsystems have a somewhat but not entirely opposite function. For example, the sympathetic system is responsible for constricting the pupils, increasing salivation, slowing the heart, and stimulating tear glands, while the parasympathetic system is responsible for dilating pupils, restricting salivation (and increasing sweating), and increasing the heart rate, but has no effect on tear glands (Cornelius, 1996).

With respect to speech production, Fonagy (1981) deduced that "anger causes spasmodic contraction of the throat, giving a strangled voice along with sudden outbursts, and an angry person would behave in a tense way, perhaps given to sudden violent actions." Lindsley (1951) found that "certain emotional situations, especially startle, conscious attempts at deception, and conflict" cause respiratory fluctuations.

Dryness of the mouth during "excitement, anticipation, fear and anger" is caused by the activation of the sympathetic nervous system (Williams and Stevens, 1972), as is "tremor and disorganization of motor response" during "emotional conflict" (Williams and Stevens, 1972), and hence these emotions have an effect on the various instruments of speech production, "including the larynx, which is directly involved in the control of F0" (Williams and Stevens, 1972). Because the vocal cords are directly responsible for the production of F0 (by controlling the velocity of air), the speech produced is more likely to be affected by these physiological changes here than in the jaw, lips, tongue, etc. (Williams and Stevens, 1972). In addition, Frick (1986) states that an increase in the size of the vocal tract leads to a "lowering of the formant frequencies" associated with anger, and a link has been found between increased loudness in speech and higher throat tension (Murray and Arnott, 1993).

Tartter (1980) discovered that the fundamental and formant frequencies (and in some cases energy and duration) were raised due to smiling. This leads to the assumption that feelings of happiness trigger physiological changes that cause the mouth to smile, directly altering the size and shape of the vocal tract.

Vocal expression of disgust is mainly generated in the chest area (Fonagy and Magdics, 1963). Fonagy (1981) discovered that people expressing disgust experience "spasmodic muscle action" which disturbs vibration in the "true vocal folds" (Murray and Arnott, 1993). Moreover, Williams and Stevens (1972) state that "an increase in respiration rate would presumably result in an increased subglottal pressure during speech, [... yielding] a higher F0 during voiced sounds in speech."

From these findings we can see that emotions with high arousal levels such as fear, anger, and happiness activate responses from the sympathetic nervous system, and those with lower

arousal levels such as sadness and disgust activate responses from the parasympathetic nervous system.

2.6 Review of the research on vocal emotion recognition

Considerable effort has been focused on the study of emotion recognition, most dealing with facial emotion recognition. In this section we will examine past works relating to emotion recognition from speech, however some studies mentioned use bimodal techniques, that is they use both audio and visual cues (speech and facial) to recognise emotion. Of these bimodal studies, we will focus only on the audio aspect.

A common avenue for future work described in several studies has been to improve recognition of emotional speech by looking for more realistic speech data. Generally, it is difficult to acquire candid emotional speech samples due to ethical and moral implications. Therefore in most research databases of speech samples are recorded using actors or students, and although the recognition results are often promising, these rates drop significantly when applied to natural emotional speech.

An example of this reduction in classification are highlighted in two studies from Huber *et al.* (1998, 2000). In the first study, acted speech data was used and a recognition rate of 94% was achieved for anger and neutral by neural network. In the following study, Huber *et al.* (2000) attempted to gather more natural speech data using induction techniques (see Chapter 3 for more on this). In this setting, users were naïve to the methods used to elicit emotions. They believed they were testing a voice automated appointment scheduling system. However, it was secretly controlled by a human and designed to be defective in order to cause annoyance among the subjects. In this study, using the more natural emotional speech, they achieved an average recognition rate of 60% for the same emotion set - a significant difference. This highlights a problem with most research in this area: a lack of naturally occurring data.

2.6.1 Instance-based learners

Dellaert *et al.* (1996) used K-nearest neighbours to detect different emotional speech samples. In this study, a corpus of over 1250 emotional utterances was obtained from five actors. 50 samples were recorded for each emotion: happiness, sadness, anger, fear, neutral. They composed two feature sets: the first with seven basic prosodic features (mean, standard deviation, minimum, maximum, range, speaking rate, and pitch slope) and the second with 17 prosodic features extracted using a smoothing spline approximation of the pitch contour. Dellaert *et al.* (1996) used promising first selection and forward selection feature selection techniques to determine an optimal set vector for each classifier. They compared human performance with several well-known pattern recognition techniques including maximum likelihood Bayes, kernel regression,

and K-nearest neighbours. In their preliminary tests, it was found that an approach using K-nearest neighbours was the most efficient classification technique. Next, to improve the accuracy of the classifier, they employed a technique of majority voting of subspace specialists based on cooperative composition. In this technique, specialist classifiers are added to a master classifier based on how they cooperate with the existing members of the set. It was also found that using the smoothing spline approximation of the pitch contour was useful in gaining more information from the speech signal, but that an advanced technique was needed to manage this larger set of features, hence the use of forward selection. Their final results show that the chosen method displays a significant improvement over other methods discussed in the study. It was determined that the K-nearest neighbours classifier with majority voting and the set of 17 features pruned with forward selection yielded the highest classification rate at 79.5% compared with the set of seven features which were classified correctly 67.5% of the time.

Lee *et al.* (2004) also used a K-nearest neighbours classifier. Approximately 7200 speech samples from real users interacting with software at a call centre were collected. In an attempt to simplify classification, only negative versus non-negative emotions were studied. Using only prosodic features and forward selection to optimise the features vectors, a recognition rate of roughly 80% was achieved.

2.6.2 Artificial neural networks

The following are several studies which have taken advantage of supervised neural networks to classify emotions. In these studies, it is typical to use approximately 70% of the sample data for training, where the network is taught to learn the pattern, and the remaining 30% for testing, where the accuracy of the network is tested.

Huber *et al.* (1998) used a neural network devised of multi-layer perceptrons to classify emotional speech. A database of emotional speech was collected consisting of 2000 utterances of angry and neutral speech from 20 actors. They used feature vectors based on the fundamental frequency such as regression coefficients, pauses, speaking rate and some lexical flags. The neural network was trained using the bootstrap method, which uses different subsets of training samples to improve classification accuracy, and a recognition rate of 94% was achieved.

Later, in a novel approach for gathering more natural speech data, Huber *et al.* (2000) devised a Wizard of Oz (WoZ) scenario where subjects were given a task designed to elicit specific emotional responses such as frustration and anger. They recorded 4684 utterances from 39 non-professional subjects. Using prosodic, lexical and global features, and a neural network classifier similar to that in their previous study they achieved a classification rate of roughly 60% for anger and neutrality. These two studies highlighted the differences between acted emotional speech and more natural emotional speech and the difficulties involved classifying real-world data.

Nakatsu *et al.* (1999) designed a neural network ensemble consisting of eight subnetworks

each trained to recognise one of eight emotions. The eight emotions were anger, sadness, happiness, fear, surprise, disgust, playfulness, and neutrality. The speech corpus, or database, consisted of a large set of samples from 50 male and female subjects who recited phrases for each of the eight emotions. Features extracted in this study were pitch and pitch derivatives (linear prediction coefficients) and intensity. They attained recognition rates between 50-55%.

Petrushin (2000) compared three classifiers in his study: K-nearest neighbours, a neural network and an ensemble of neural networks. The database in this study consisted of 700 short utterances by non-professional actors containing samples for the emotions happiness, anger, sadness, fear, normal. Prosodic features such as speaking rate, formant frequencies (F0, F1, F2) and intensity were identified and the RELIEF-F algorithm was used as a method of feature selection. It was concluded that the neural network ensemble outperformed K-nearest neighbours and a single neural network with a classification rate of 75% for agitation and calm. For a larger set of emotions (the five stated above), an average classification rate of 66% was achieved.

In their research, Sato *et al.* (2001) classified emotional speech using the average of ten simulations of a neural network. To gather emotional speech samples, they recorded 13 actors reciting the Japanese phrase "kimura" totalling 62 utterances in four emotions: neutral, anger, joy, and sadness. Pitch information was extracted by using cepstrum analysis. Utilising prosodic features alone (accent, intonation, pitch structure, stress, tempo, and rhythm among others) a classification rate of 71.5% was attained for two emotions (neutral versus anger).

In an extensive study, Yacoub *et al.* (2003) found that a supervised neural network was a better classifier than a support vector machine (kernel method), 3-nearest neighbour algorithm (clustering method) and a C4.5 decision tree. They compared classification rates on several sets of emotions: anger versus neutral; hot/cold anger versus neutral/sadness; hot anger versus neutral/sadness versus happiness; and neutral, hot anger, cold anger happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, contempt. The speech corpus consisted of 2433 utterances collected from eight actors from the LDC consortium. 37 prosody-based features were considered from several categories: pitch contour, first derivative of the pitch contour, jitter, energy contour, first derivative of the energy contour, shimmer, and durational features. Following Dellaert *et al.* (1996), they used a forward selection technique to prune insignificant features which resulted in two secondary feature sets consisting of 19 and five features. A classification rate of 90% was reported for anger and neutrality, 91% for hot/cold anger versus neutral/sadness, and 8.7% for all 15 emotions (slightly higher than random).

2.6.3 Probabilistic methods

In a study aimed at examining the importance of features from different domains (e.g., prosodic, spectral, contextual), Polzin and Waibel (2000) employed Gaussian mixtures and compared classification rates based on features from different domains. They collected a corpus of over

5000 samples containing angry, sad and neutral speech from feature films starring professional actors. They selected eight prosodic features, 32 spectral features and some verbal features. To reduce the dimension of the spectral features they used a linear discriminant feature selection technique. They concluded that features from the spectral domain had the highest recognition rate at 63.9% compared to prosodic (60.4%).

Devillers *et al.* (2002) designed a recognition system based weighted features and a maximum likelihood estimate. The speech data was collected from natural client-agent conversations in a stock exchange support centre and consisted of 5012 sentences from 100 client-agent conversations containing anger, fear, satisfaction, excuse and neutral emotions. This was one of the first studies which used real-world data to build a classification system. They extracted prosodic information and identifiers based on the context of the speech and achieved a 68% average accuracy rate for classification between five emotions.

Nwe *et al.* (2003b) used five-state hidden Markov models to classify emotional speech samples. The samples consisted of 3150 utterances from the SUSAS (Speech Under Simulated and Actual Stress) database. Five emotions related to stress were studied: neutral, anger, clear, lombard and loud. They used a novel approach by locating prosodic clues to emotion in log frequency power coefficients extracted from the frequency domain in speech samples. These features indicated different levels of intensity in the frequency bands for different emotions and closely model the frequency range of the human ear. An average rate of 84% was achieved for speaker dependent classification.

In a later study, Nwe *et al.* (2003a) collected 720 utterances from 12 non-professional speakers and as in their previous research, they used subband based features from 12 log frequency filters to closely model the frequency range of the human ear. In this study, however, they used four-state hidden Markov models as it was determined that with five and six state models most of the samples remained in only four of the states. This research yielded a 78.1% classification rate for six emotions, 88.8% for two groups of three similar emotions, and 100% accuracy for anger versus sadness. It should be noted that the models trained were speaker dependent.

McGilloway *et al.* (2000) used a linear discriminant method for the classification of five emotions: afraid, happy, neutral, sad, angry. They used collected 197 passages from five non-professional actors reciting emotional speech for each of the five emotions. They extracted 32 features based on prosody. Using the optimally reduced feature vectors, McGilloway *et al.* (2000) compared three recognition algorithms: the linear discriminant method, support vector machine (with linear and Gaussian similarity measures), and vector quantisation. It was determined that use of the linear discriminant method yielded the lowest classification error and proved most useful for their features. The classification rate achieved by this method was approximately 55%.

2.6.4 Decision trees

Ang *et al.* (2002) used a decision tree to recognise emotion from human-computer speech data from the DARPA Communicator project where users made travel arrangements through a computer system. This data was used because it more closely modelled a real-world scenario. Humans interacted with a simulated computer interface that produced errors and caused annoyance in the users. It should be noted, however, that because the users were not making real travel arrangements, the levels of frustration were not as high as they might be in a natural setting. This study differs from others because Ang *et al.* (2002) chose to use language information alongside prosodic features to improve recognition rates. This was motivated by the realisation that much information about the psychological state of the subject can be inferred by the language they use such as negatives (ie no, not, never) and swear words. Features were selected using a method of brute force selection and a decision tree was employed to classify the samples. 80.2% of samples were classified correctly into two groups: anger/frustration versus everything else.

Breazeal and Aryananda (2002) designed an emotion detection system in a robot for natural interaction with human caretakers. A dataset of 266 utterances were collected depicting four intonation styles normally used with infants and pets: prohibition, affirmation, attention, and approval. Sequential forward selection was used to find a subset of features that carried the greatest variance. Gaussian mixture models for each emotion class were trained and achieved an overall classification accuracy of 81.94%.

2.7 Areas for improvement

Databases. One major issue that spans the entire field of emotion recognition, be it facial, vocal, gesture-based, or multi-modal, is the lack of standardised datasets. Because of the difficulties in sharing datasets and the ethical issues involved, most studies collect and use their own data, which leads to diversity of results which are difficult to compare.

In this research, we propose to compare results based on two different datasets, one which has been used in a previous study (Nwe, 2003). These two datasets are different because one has been collected from a call-centre environment where people are interacting and conversing in a natural environment, and the other is collected using non-professional actors and actresses. These databases are introduced in detail in Chapter 3.

Real-world data. Most studies attempt to generate emotion recognition models using emotional data collected from professional and non-professional actors and actresses. This often includes radio actors, audio samples from motion pictures (Polzin and Waibel, 2000), students, colleagues, etc.

As stated in Scherer (2001), these data sources are certainly useful for exploring feature sets, classification schemes, and the like as they yield a controlled and usually noise-free set of

samples on which to base research. However, this data is of almost no use for systems designed for real-world use. Spoken emotion in real-world situations is highly spontaneous, often subtle, and difficult to model using actors and actresses.

Data collected from actors and actresses, whether professional or non-professional, is inherently biased because these actors and actresses are being instructed to give emotional responses. The resulting data often only contains extreme emotions and is usually over- or under-acted. Studies using non-professional actors and actresses claim that doing so alleviates the over-acting by professional actors and actresses. But studies using professional actors and actresses claim that because these individuals are professionals, they will know how to accurately model emotion.

Studies such as (Ang *et al.*, 2002; Huber *et al.*, 2000) attempt to yield more realistic data by presenting naïve users with scenarios which elicit the desired emotional response. While this is a more reliable way to derive data close to the real-world, there are factors such as display rules and the awareness of being involved in a study that can affect the spontinaety of the responses.

We will show that using real-world data is difficult yet necessary because of the subtle and spontaneous nature of the emotions present in these situations.

Improvements on classification. Because most studies in emotion recognition come from disciplines which focus more on the psychological aspects rather than the mathematical, there is little or no focus on applying recent developments in machine learning and data mining. Generally, a particular study will somewhat arbitrarily compare two or three classification algorithms. More advanced studies will actually try to improve the classification system by using techniques common in machine learning and data mining, such as bagging, boosting, feature selection, or hybrid approaches.

We put the focus of our research on the actual classification scheme used. Much work has gone into comparing the performance of many different types of classification methods as well as ensemble schemes.

Real-time processing. Although several previous studies mention the real-time processing and classification of spoken affect (Petrushin, 1999; Breazeal and Aryananda, 2002), none go into any detail on how this processing is done. In this thesis, the real-time processing aspect of the research is discussed, with an implementation in software that gives a concrete example.

2.8 Summary

In this chapter we formed the foundations and introduced the background of research on the expression of emotion in humans. The two main theoretical representations of emotion were described, and we explained why we have chosen discrete emotion theory to simplify the mappings between the psychological state and the perceived expression.

Emotional classes were then explored, and after evaluating several prevalent veins, we chose the six basic emotions: anger, fear, happiness, surprise, sadness, and disgust. A seventh state neutral was also added as a baseline emotional state. These particular classes of emotion are the most heavily studied and therefore comparison between studies is simplified.

Emotional expression in humans was then summarised. The different channels of expression were introduced: facial, vocal, behavioural, contextual, and gestural. We then began to show how humans show emotion through vocal expression. Ekman's display rules are also introduced, which have an effect on how the emotion is perceived. These display rules also show that there is some cultural relativity in the expression of emotion.

Following was a detailed summary of previous research in automatic emotion recognition from speech signals. Studies were introduced by way of classification method.

In the next chapter we introduce the different data acquisition methods prevalent in the literature on vocal emotion research. We then introduce our databases of emotional speech, how they were collected, validated for emotional content, and processed in preparation for feature extraction.

Chapter 3

Emotional Speech Data Acquisition

3.1 Introduction

The previous chapter laid the foundation for the rest of this thesis. We explored the different theoretical representations of emotion as well as the language used to describe the different emotion classes of discrete emotion theory. Despite a slight lack of agreement in the field, we have rested on the well explored six basic emotions: anger, fear, disgust, sadness, happiness, and surprise. Neutrality is used as a baseline emotional or unemotional state.

Automatic vocal emotion recognition rests heavily on emotional speech data. The acquired data should be representative of that in the real world in order to gain insight into which features are useful in determining patterns of emotional speech.

In this chapter three different methods of emotional speech data acquisition are presented. The advantages and disadvantages of each are also discussed. Next, we introduce the datasets we have acquired. In this research, we utilise data from two of the three acquisition methods. This allows a comparison to be made which highlights the differences in speaking patterns between the different methods.

3.2 Emotional speech acquisition

There are three methods of data acquisition in emotion research (Scherer, 2003). The first is natural expression, where data is collected from a real-world situation where users are not under obvious observation and are free to express emotions naturally, as they would in an everyday situation. The second is induced emotional expression, where naïve users are presented with scenarios that induce the required emotional response. Last, speech acquisition using simulated or portrayed emotional expression makes use of professional or non-professional actors and actresses. Subjects are instructed to produce emotional expressions for various emotion classes, with sometimes varying degrees of intensity or arousal.

3.2.1 Natural expression

For emotion research with applications in real-world situations, an obvious advantage would be to use speech samples recorded during naturally occurring scenarios. Research is further advantaged if these samples can be obtained from a similar setting to the intended application. This is due to several reasons. The first is that the distribution of expected emotional speech should remain the same, as long as the setting or focus of the setting does not change significantly. The second is that the speech samples will be of similar recording quality, provided similar recording equipment is used. Another reason is the sampling of speakers should remain the same. Areas where numerous speakers interact will have a much higher variation in vocal expression, whereas systems with a small number of regular speakers will have less variation, and might be better suited to a speaker-dependent model.

Apart from the seemingly obvious advantages, this type of acquisition has several drawbacks. First, as reported in other studies (Nwe, 2003; Batliner *et al.*, 2003), there are ethical and privacy issues involved with recording people without their knowledge or permission.

Second, there is a lack of control in recording natural speech (Johnstone, 1996; Scherer, 2003). The researcher does not have the power to govern which emotions are to be expressed, and so is often forced to rely on the natural distribution of the resulting dataset.

For studies trying to gain a better picture of the vocal emotion encoding and decoding process, a lack of equal distribution of emotional classes would severely affect the resulting research outcome (Lee and Narayanan, 2005). As stated previously, this effect is reduced when the research aims to provide some application to the same area from which the dataset was recorded.

When the researcher lacks control in the recording environment, there is difficulty in judging the underlying emotional expression (Johnstone, 1996). This can be mostly alleviated by the use of objective listeners classifying the samples (Devillers *et al.*, 2002).

Other problems arise due to background noise and recording quality. Often there is speaker talk-over, where two or more speakers are talking at the same time. Background noise such as cars, computers or appliances, cats meowing, children playing musical instruments, etc., contribute to degradation of the recording quality. Speaker proximity to the recording device also has a negative effect on the collection of the full utterance, as noted by (Scherer, 2003).

Some studies have argued that it is impractical and difficult to acquire natural emotional speech recordings because of technical difficulties in allowing such recordings to take place (Batliner *et al.*, 2003). However, nowadays, with the increased presence of government and privately controlled video and audio recording devices (usually for security purposes), as well as systems put in place to aid employee training (call-centres, supermarkets), the public's exposure to these devices is steadily increasing. Now it is more common than not to have telephone calls to call-centres or emergency services recorded and monitored and often saved for indefinite periods of time. The usefulness of this data is almost infinite, and with research teaming up

with private enterprise, the exposure of this data to the research community is becoming more common. However, this is where ethics and morality play a key role. These recordings are usually disclosed to the public based on the premise that the intended use is for a certain purpose (security, training, etc.), and when this data is used for other purposes that ethical barriers are broken.

3.2.2 Induced expression

A popular method for capturing a more tightly regulated set of vocal emotion is through the use of induction techniques (Huber *et al.*, 2000; Ang *et al.*, 2002). In the field of automatic emotion recognition from speech, this is more commonly known as the *Wizard-of-Oz* (or simply *WOZ*) method, alluring to the idea that someone is specifically altering the scenario from behind the scenes. Here, participants are given tasks which are craftily designed to elicit certain emotional responses. This method has the advantage of giving researchers a high degree of control over the timing and nature of the emotional response.

In (Huber *et al.*, 2000), naïve users were presented with a malfunctioning speech activated appointment system. When the system performs as expected, the subjects exhibit neutral or low-arousal affect. However, when the system is made to malfunction, frustration is caused in the users, which results in vocal responses with higher arousal.

Other scenarios, such as altering components of computer games during play, have also been carried out (Johnstone, 1996; Johnstone *et al.*, 2001). The use of positive or negative images, videos, or other stimuli have been used to measure the effects of emotion in the voice (Karlsson *et al.*, 2000). In extreme situations, Stemmler *et al.* (2001) have resorted to causing anger in their participants through "arrogant and offensive behaviour of the experimenter" (Scherer, 2003). However, Murray and Arnott (1993) point out that this practice is rarely used due to the ethical implications.

There are drawbacks to these methods. For instance, these situations usually only provide weak or low amounts of emotional arousal (Scherer, 2003). This may be because subjects know they are participating in a study and may inhibit their emotions by using display rules[1].

3.2.3 Simulated expression

The use of simulated or portrayed emotional expression is probably the most common technique for collecting emotional speech samples. Quite simply, this method involves actors and actresses reciting normally neutral text with different emotional properties. Most automatic emotion recognition systems explored in the past have used datasets of portrayed emotional expression (Lee *et al.*, 2004; Yu *et al.*, 2001; Dellaert *et al.*, 1996; Polzin and Waibel, 2000).

[1] See Section 2.5 for more on Ekman's display rules.

Samples collected in this manner are often relatively similar with respect to the level of arousal and experimenters are afforded a high degree of control over the types of emotion expressed. This is useful for research that involves comparing characteristics of many types of discrete emotions.

However, this method is not spared controversy; there is often argument within the research community about whether or not portrayed expression is a useful model for naturally occurring expression (Scherer, 2003). This is due to the fact that actors and actresses may overact or neglect important patterns in emotional speech (Scherer, 1986).

This argument aside, data collected in this way does at least offer a basis for comparing different vocal characteristics of emotion (Scherer, 2003). Moreover, since all "spontaneous" emotional expression is in some sense an act or portrayal of an internal state or appraisal, it can be argued that simulated emotional expression is perfectly suited to the study of vocal emotion. Scherer (2003) brings up a good point when he says that because acted emotion is generally well recognised by human listeners, the assumption can be made that these portrayals are generally accurate renditions of natural expression.

Another debate is whether the use of professional actors and actresses is better than non-professional actors and actresses. Some psychologists believe that professionals have a greater idea of how to represent different emotions more naturally than non-professionals, and thus are a more reliable source (Scherer, 2003). However, other studies argue that professional actors and actresses are subject to over-acting (Nwe *et al.*, 2003a), and non-professional actors and actresses should be used for that same reason.

3.2.4 Summary of acquistion methods

Each of the above methods for gathering vocal expressions of emotion have advantages and disadvantages. Each method, when used alone, has the potential to undermine the soundness of a particular piece of research. Scherer (2003) remarks that it is better to combine and compare data acquired from natural, inducted, and portrayed sources to determine which correlations can be derived from the overlap. In fact, Scherer makes a plea for new studies doing just this: "Unfortunately, so far there has been no study in which a systematic attempt has been made to compare portrayed and naturally occuring vocal emotions. This would seem to be one of the highest priorities in the field."

For this reason, two different sources of emotional expression were collected for this thesis: natural and portrayed (simulated). These two datasets are introduced in the following section.

Previously, (Batliner *et al.*, 2000) compared two different data acquisition methods: portrayed and induced. The findings of this research were that the portrayed emotions were recognised with greater accuracy than the induced emotions.

3.3 Databases of emotional speech

The two datasets used in this study are presented below. The first dataset is taken from a natural scenario, as described in Section 3.2.1. The second dataset is acquired from actors and actresses portraying emotional speech, as described in Section 3.2.3. Table 3.1 summarises the resulting datasets.

Table 3.1: Summary of the datasets used in this study. The NATURAL dataset is collected from a call-centre and the ESMBS dataset is obtained from a previous study and consists of utterances by non-professional actors and actresses.

Name	Description	Emotion classes	Speakers	Samples
NATURAL	Collected from a call-centre for an electricity company	2 (anger, neutral)	11	388
ESMBS	Collected from Burmese and Mandarin non-professional actors	6 (anger, happiness, sadness, disgust, fear, surprise)	12	720

3.3.1 Natural data collected from a call-centre

The first database used in this research was provided by a call-centre that handled customer inquiries for several electricity companies. Customers call and speak directly to a customer service representative (CSR). The customers query or provide information about their accounts, billing information, address, payment methods, etc. Often, customers have a dispute to resolve with the company and subsequently, emotions are expressed.

The average length of a conversation between a customer and CSR was 3 minutes and 40 seconds. The median call length was 2 minutes and 38 seconds. The longest call duration was 34 minutes and 3 seconds, and the shortest recordings were around 1 second, but involved no audible speech and were probably the result of some technical error during the recording process.

The difficulty of a human manually classifying the samples in the database and the apparent quantifiable lack of common emotive states such as happiness, sadness, surprise, leads the researcher to consider only two emotive states: anger and neutrality. This is not necessarily a detriment to the research, as a call-centre is most interested in distinguishing between the satisfied (neutral) customers and the unsatisfied (angry) customers. Table 3.2 shows the distributions for respective emotion classes.

Because of the low distributions of happiness, sadness, fear, disgust, and surprise, it can be assumed that the probability of these occurring in the call-centre are quite low, and because of this it is safe to consider only anger and neutral emotional states. Similarly, (Devillers *et al.*, 2002) also used data from a customer service centre. This study also found low emotion dis-

Table 3.2: Distribution of perceived speaker affect from natural corpus (NATURAL)

Number of conversations (%)	Emotion class
93.3	Neutral
3.1	Anger
1.8	Happiness
0.1	Sadness
0.0	Surprise, Fear, Disgust

tribution and subsequently retained two of the basic emotion classes, anger and fear, because the probabilities of other emotions in that context were very low. Ang *et al.* (2002) used induction methods for collecting emotional speech data and observed a high amount (84%) of neutral samples, followed by a low amount (8%) of annoyance. Due to this they limited their study to include only annoyance and frustration versus everything else.

Telephone-quality recording is 8000 Hz, and as such frequencies greater than 4000 Hz cannot be accurately transmitted and are lost. Sampling theory states that due to the *Nyquist frequency*,

$$1/T > 2F_N \tag{3.1}$$

where $1/T$ is the sampling rate and F_N is the Nyquist frequency, the highest frequency that can be accurately sampled is at most half that of the sampling rate[2] (Rabiner and Schafer, 1978). Therefore, frequencies over 4000 Hz are lost through telephone communications. As will be covered in Chapters 4 and 5, the highest frequencies required are those relating to the third formant frequency (F3), which is typically less than 3000 Hz, and hence telephone quality speech is suitable for the estimation of F3.

Pantic and Rothkrantz (2003) have stated that it is important to have a wide range of speakers: sex, age, smoking pattern, social background. This dataset is extremely good in this regard. Although only a small number of speakers from the entire database were selected, the statistical sampling of these speakers is very large because the source of the data is a call-centre in which nearly everyone in society must at some time or another speak to on the telephone. Whether it is signing up for a new account, resolving a dispute over a certain issue, or simply calling with a bi-monthly metre reading. People of all sexes, ages, and social backgrounds must interact at this nexus.

[2]Aliasing occurs when the sampling rate is not high enough, and high frequencies can be mistaken for lower frequencies.

Format and content of NATURAL dataset

The original database consisted of approximately 1500 conversation files. Each conversation involves at least two speakers, with sometimes more speakers temporarily interjecting. The files are sorted into directories based on the customer service representative (CSR). For example, a CSR named John Smith may have fifteen conversation files, each involving himself and at least one other speaker.

Each conversation is formatted in MPEG Layer III at 32 kilobits per second. The sampling rate is 22050 Hz with 16 bit quantisation and one channel (mono).

Conversion of NATURAL dataset

From the 600 conversation files listened to, 6 conversations were selected as neutral (many existed so the selection was purely arbitrary) and 5 conversations were selected as angry. Each conversation file was converted from its original format of MPEG Layer III to pulse code modulation (PCM) WAVE format to simplify the segmentation process (described below). The sampling rate, quantisation and channel did not change during this process. The tool used in this conversion was the GNU command line program "mpg123."

Segmentation of NATURAL dataset

Conversation files are too long and contain multiple speakers which renders the database is unusable. A more appropriate granularity is single-speaker phrase-level utterances. In other words, we need to segment each conversation file into a series of phrase-level utterances each containing only one speaker. For this process, the popular speech tool "Praat" was utilised[3]. With the entire conversation loaded in the Praat sound editor, utterances were selected by listening, 'cut' out, and saved as new PCM WAVE files of the same format (22050 Hz, 16 bit, 1 channel). Table 3.3 shows several transcriptions from samples in the database.

Table 3.3: Sample utterances from the NATURAL database.

Utterance	Affective state
This is insane.	Anger/frustration
It's on the account twice.	Anger/frustration
What do you mean?	Neutrality
This is under action - we will do something about it.	Anger/frustration
Right, we want to arrange to have our power put on.	Neutrality

[3]Praat can be downloaded from `http://www.praat.org/`.

In the end, utterances were selected from a total of 11 speakers, 2 male and 9 female. The maximum utterance duration is 13.60 seconds, the minimum is 0.52 seconds, and the mean duration is 3.25 seconds.

Validation of NATURAL dataset

Initially, the dataset comprised 190 angry utterances and 201 neutral utterances totalling 391. However, to ensure that the manual classifications were objective, nine listener-judges were instructed to classify the entire dataset (see Appendix A for full details on the system and setup of the listener-judges). After the results of the listener-judges were available, the final dataset comprised 155 angry utterances and 233 neutral utterances. In total there were 388 utterances (three utterances were labelled as ties and were subsequently discarded).

3.3.2 Simulated data from the ESMBS database

The ESMBS database (Emotional Speech of Mandarin and Burmese Speakers) was collected for a doctoral thesis by Tin Lay Nwe (Nwe, 2003). This dataset was collected to study emotional effects on vocal parameters, as in this thesis.

The dataset was collected from a set of 12 non-professional actors and actresses. Six Mandarin and six Burmese speakers were used, with each of these six consisting of three men and three women. Each speaker recorded ten different utterances for each of the six emotions. In total, for the 12 speakers, there were 720 emotional utterances.

The main goal of the collection of this dataset was to study intra-speaker variations caused by emotions. A secondary goal was described as studying emotion classification in a text-independent and speaker-dependent manner.

Format and content of ESMBS dataset

The emotion set represented by this dataset are the six prototypical emotions most often studied in this field: anger, disgust, fear, joy, sadness, and surprise. The mean length of the samples in the dataset was 1.50 seconds. All speech samples were recorded with 16 bits per sample at 22050 Hz and stored in PCM WAVE format. The content of each utterance was a phrase or sentence which also contained emotion from one of the above six.

Validation of ESMBS dataset

Four listeners were used to judge the emotional content of each utterance. Table 3.4 lists the accuracy rates for the human evaluators by emotion category. These listeners could not understand the language of the respective speakers, so vocal characteristics and not contextual information was the only information used for classification.

Table 3.4: Human classification performance by emotion categories (from (Nwe, 2003))

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Mean
Burmese (α_B)	98.3	63.3	45	53.3	85	65	68.3(mean x_B)
Mandarin (α_M)	96.7	55	45	41.7	91.7	48.3	63.1(mean x_M)
Mean	97.5	59.15	45	47.5	88.35	56.65	65.7

Average classification by human evaluation was found to be 65.7% (68.3% for Burmese and 63.1% for Mandarin). This human accuracy coincides with previous research in this field (Dellaert *et al.*, 1996; Petrushin, 2000; Polzin and Waibel, 2000), as well as previous research on cross-cultural emotion recognition from speech (Scherer, 2003; Elfenbein and Ambady, 2002) which typically describe human classification rates between 55% and 70%. Nwe (2003) concluded that vocal characteristics of emotion must be shared between at least these two differing cultures, and as described in the previous chapter, vocal expression of emotion shares foundations across different cultures. These hypotheses are backed up by psychologists who have studied cross-cultural expression of emotion (Ekman, 1992; Izard, 1977; Scherer, 2003).

From Table 3.4, it can be seen that anger and sadness are highly recognisable emotions. This is supported by Scherer (2003), who also notes that disgust is normally recognised by humans with the lowest accuracy (slightly above chance).

3.4 Summary

In this chapter, the two databases used in this research were introduced. The database NATURAL was collected in a natural environment (call-centre) in order to build an automatic system which will be applied in this same environment. Therefore, the source of this data is highly relevant. Due to the distribution of emtional data found in this environment, only relevant emotion classes were considered for study: anger versus neutral. The probabilities that fear, sadness, joy, disgust, and surprise will happen in this environment are extremely low, and therefore can be ignored. Human listener-judges were used to validate this dataset in order to determine a ground truth and to reduce subjectivity in the manual classification.

The ESMBS database was collected from a previous study in vocal emotion research. This data was collected from non-professional actors and actresses reciting emotional utterances under direction of experimenters. This data will be used to compare both feature extraction and classification techniques with the natural data. Emotion classes anger, joy, fear, sadness, disgust, and surprise are considered. As with the NATURAL database, human classification accuracy was provided, but in this case it was used to compare against automatic classification, rather than to establish a ground truth.

These databases were derived using different acquisition methodologies in order to compare differences between them and discover shared characteristics found in any overlap. This method follows recommendations by Scherer (2003) to better understand the differences between these methods. Although a third method, induced expression, was mentioned, it was not feasible during the course of this study to acquire this data, so only the remaining two were considered. This can be left as a future work.

In the next chapter, we look at some of the findings from the literature on the vocal expression of emotion and describe the correlations between emotional states and the human speech production system and the most relevant characteristics are examined.

Chapter 4

Acoustic Correlates to Emotional States

4.1 Introduction

This chapter focuses on the properties of speech that correlate to the different emotional states described in the previous chapter. As stated, the emotions with which we are concerned are the basic and primary emotions which have received so much attention in the literature. These are *anger, fear, happiness, disgust, sadness, surprise*, and the baseline state *neutral*[1]. Questions as to whether such properties exist are easily dismissed: "without such distinguishable acoustic patterns for different emotions, the nature of the underlying speaker state could not be communicated reliably" (Scherer, 2003).

Research in this field dates back as early as the start of the 20th century (Scherer, 2003). Studies by Skinner (1935), Fairbanks and Pronovost (1939), and Cowan (1936) are often noted in the literature as pioneers investigating the properties of emotional speech. The surge in research in this area was attributed by Scherer (2003) as being caused by the "rapid dissemination of the telephone and [...] radio." These and other psychologists of the time found that the vocal expression of emotion carries significant differences for each of the basic emotions. These differences were theorised and later discovered to reside in the vocal or prosodic content.

Murray and Arnott (1993) suggest that during the speech process, two different channels of information are conveyed. These are the verbal content (meaning the words which are expressed in the language used and relating to linguistics) and vocal or prosodic[2] content (comprising of intonation, volume, voice quality, and rate of speech and relating to phonetics)[3].

[1]The *neutral* state is an unemotional state that is used in the literature mainly as a reference to which other emotional states are compared (Cowan, 1936; Murray and Arnott, 1993; Breazeal and Aryananda, 2002).

[2]Also sometimes referred to as *extralinguistic features* (Mahl, 1963).

[3]Prosodic content also contains vocal signatures that humans take advantage of for differentiating and recognising

4.2 Prosody-based features

Prosodic parameters have been found to represent the majority of emotional content in verbal communication (Murray and Arnott, 1993; Scherer, 2003). Of these, fundamental frequency (pitch), energy, and speaking rate are widely observed to be the most significant characteristics (Batliner *et al.*, 2003; Lee *et al.*, 2004; Dellaert *et al.*, 1996; Ang *et al.*, 2002; Huber *et al.*, 2000; McGilloway *et al.*, 2000; Polzin and Waibel, 2000; Nwe *et al.*, 2003b; Petrushin, 1999). The specific correlations between the basic emotions and these prosodic features are discussed below.

4.2.1 Fundamental frequency and emotional speech

The fundamental frequency (F0), often referred to as the pitch, is one of the most important features for determining emotion in speech (Dellaert *et al.*, 1996; Huber *et al.*, 1998; Nakatsu *et al.*, 1999; Polzin and Waibel, 2000; McGilloway *et al.*, 2000; Petrushin, 2000; Ang *et al.*, 2002; Devillers *et al.*, 2002; Nwe *et al.*, 2003b; Yacoub *et al.*, 2003; Lee *et al.*, 2004). The fundamental frequency is defined as the lowest frequency at which the speech signal repeats itself (O'Shaughnessy, 2000).

The F0 contour has been shown to vary depending on the emotional state being expressed. Cowan (1936) discovered that neutral or unemotional speech has a much narrower pitch range than that of emotional speech, and found that as the emotional intensity is increased, the frequency and duration of pauses and stops normally found during neutral speech are decreased (Murray and Arnott, 1993).

More specifically, angry speech typically has a high median, wide range, wide mean inflection range, and a high rate of change (Fairbanks and Pronovost, 1939). Williams and Stevens (1972) discovered vowels of angry speech to have the highest F0, and Fonagy (1978) found that angry speech exhibits a sudden rise of F0 in stressed syllables and the F0 contour has an "angular" curve. Frick (1986) postulated that frustration, which has similar but less extreme physiological causes as anger, has a higher fundamental frequency than neutral speech. Scherer (1996) describes anger as having "an increase in mean pitch and mean intensity." Downward slopes are also noted on the pitch contour. Breazeal and Aryananda (2002) have shown that prohibitions or warnings directed at infants are spoken with low pitch and high intensity in "staccato pitch contours."

Cowan (1936) and Fonagy and Magdics (1963) found that happiness expressed in speech, like anger, has an increased pitch mean and pitch range.

Fear was discovered to have a high pitch median, wide range, medium inflection range, and a moderate rate of change (variation) (Fairbanks and Pronovost, 1939; Williams and Stevens,

different speakers.

1972), and increased pitch level is also apparent (Fonagy, 1978). Conversely to fear exhibiting a wide range, there are reports that fear instead has a narrow F0 range (Fonagy and Magdics, 1963).

Contrasting these more excited emotions are sadness and disgust which typically have lower physiological activation levels. Sadness is shown to yield lower pitch mean and narrow range (Skinner, 1935; Davitz, 1964; Fonagy, 1981; Oster and Risberg, 1986; Johnson *et al.*, 1986). Fairbanks and Pronovost (1939) report that disgust generally has a low pitch median, wide range, lower inflectional range, lower rate of pitch change during inflection. As with fear, there are contrasting findings with Fonagy and Magdics (1963) noting disgust having a narrow pitch range.

Figure 4.1 shows the pitch contours of two example utterances from the NATURAL dataset. It can be seen that the angry sample has downward slopes, concurring with Scherer (1996), and a greater range. The neutral sample has a monotonous contour with a shallow range.

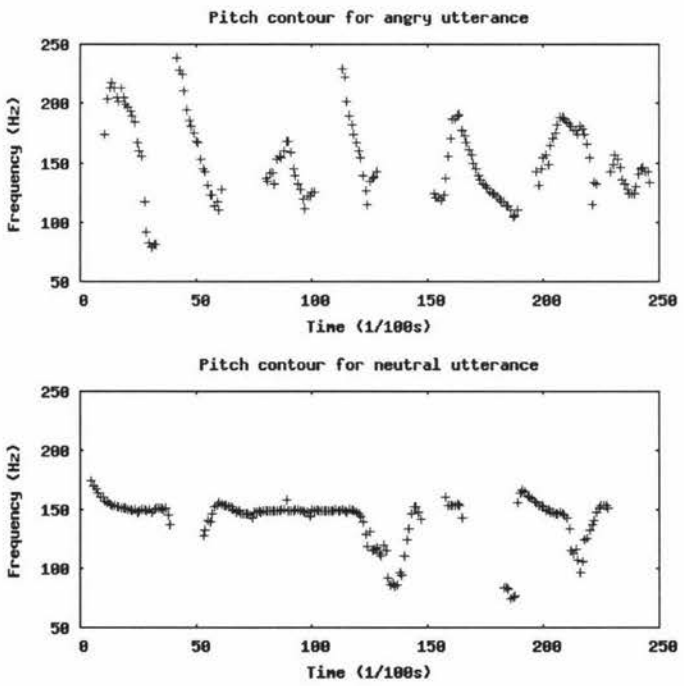


Figure 4.1: Example pitch contours for anger and neutral utterances from the NATURAL dataset. The contour for angry speech typically has a much wider range, while neutral speech is narrow and monotonous.

4.2.2 Formant frequencies and emotional speech

The resonant frequencies produced in the vocal tract are referred to as formant frequencies or formants (Rabiner and Schafer, 1978). Although some studies in automatic recognition have looked at the first two formant frequencies (F1 and F2) (Petrushin, 2000; Lee *et al.*, 2004), the formants have not been extensively researched.

Williams and Stevens (1972) found that anger produced vowels "with a more open vocal tract" and from that inferred that the first formant frequency would have a greater mean than that of neutral speech. It was also noticed that the amplitudes of F2 and F3 were higher with respect to that of F1 for anger and fear compared with neutral speech. Neutral speech typically displays a "uniform formant structure and glottal vibration patterns," contrasting the "irregular" formant contours of fear, sadness, and anger.

Scherer (2003) lists predictions of the formant frequencies along with several emotion classes. For happiness, he notes that the F1 mean is decreased while the F1 bandwidth is increased. For anger, fear, and sadness, the F1 mean is increased while the F1 bandwidth is decreased. For the F2 mean, Scherer lists it as decreased for sadness, anger, fear, disgust. Realising the shortcomings of research in this area, we hope to deduce our own findings with respect to the correlations between the formant frequencies and the speaker's emotional state.

4.2.3 The use of energy as an emotional marker

Energy, often referred to as the volume or intensity of the speech, is also known to contain valuable information (Huber *et al.*, 1998; Nakatsu *et al.*, 1999; Polzin and Waibel, 2000; McGilloway *et al.*, 2000). The intensity contour provides information that can be used to differentiate sets of emotions.

In their research, Fonagy (1981) found that angry speech had a noticeably increased energy envelope. Happiness showed similar characteristics, as reported by Davitz (1964); Skinner (1935). Sadness was associated with decreased intensity (Fonagy, 1981; Davitz, 1964) and disgust had reduced loudness (Fonagy and Magdics, 1963). Scherer (2003) notes that in fear, joy, and anger there is an increase in high frequency energy, whereas sadness has a decrease in high frequency energy.

These characteristics follow with what is expected of the emotional state. Those with high activation levels such as anger, surprise, and happiness generally have a higher intensity, while fear, sadness, and disgust have lower intensity (Nwe, 2003).

4.2.4 Rhythm-based characteristics

Properties of rhythm-based characteristics include pauses between voiced sounds, lengths of voiced segments, and rate of speech (articulation). The rate of speech is usually calculated by

measuring the number of syllables per second.

Speaking rate has been used in previous research (Dellaert *et al.*, 1996; Huber *et al.*, 1998; Petrushin, 2000; Ang *et al.*, 2002). It has been noted that fear, disgust, anger, and happiness often have a higher speaking rate, while surprise has a normal tempo and sadness a reduced articulation rate (Nwe, 2003).

Some studies use features based on the number of voiced and unvoiced frames (Dellaert *et al.*, 1996; Ang *et al.*, 2002) or features such as the average duration between voiced frames. As mentioned in Chapter 2, voiced segments consist of vowel phonemes such as "ah" or /i/ ("e") and unvoiced frames are those such as glottal stops and frictave sounds such as "ch" and "s". Unvoiced frames typically have a high number of zero crossings, meaning the signal crosses the horizontal axis more frequently than it does for voiced frames, and can resemble a random signal (Rabiner and Schafer, 1978).

Fairbanks and Hoaglin (1941) and Fonagy (1981) found that anger has an increased speech rate, and "pauses forming 32% of total speaking time." Happiness has been shown to have anywhere from a slower tempo (Oster and Risberg, 1986), to a "regular" rate (Davitz, 1964), to even an increased rate (Fonagy, 1981). For sadness, on the other hand, it has been generally agreed that the tempo is slower (Skinner, 1935; Davitz, 1964; Fonagy, 1981; Oster and Risberg, 1986; Johnson *et al.*, 1986) and the speech contains "irregular pauses" (Davitz, 1964).

Williams and Stevens (1972) stated that fear exhibited a reduced speech rate, while Fairbanks and Hoaglin (1941) contrast this by noting a high speech rate, and "pauses forming 31%." Disgust has a very low speech rate, increased pause length, with pauses typically comprising 33% of speaking time (Fairbanks and Hoaglin, 1941). The correlations mentioned above are summarised in Table 4.1.

Table 4.1: Speech correlations of the basic emotions.

	F0 mean	**F0 range**	**Energy**	**Speaking rate**	**Formants**
Anger	increased	wider	increased	high	F1 mean increased; F2 mean higher or lower, F3 mean higher
Happiness	increased	wider	increased	high	F1 mean decreased; F1 bandwidth increased
Sadness	decreased	narrower	decreased	low	F1 mean increased; F1 bandwidth decreased; F2 mean lower
Surprise	normal or increased	wider	-	normal	-
Disgust	decreased	wider or narrower	decreased or normal	higher	F1 mean increased; F1 bandwidth decreased; F2 mean lower
Fear	increased or decreased	wider or narrower	normal	high or low	F1 mean increased; F1 bandwidth decreased; F2 mean lower

4.3 Summary

This chapter examined acoustic correlates to emotional speech. It was found that the emotions with high physiological activation such as anger, happiness, surprise, and sometimes fear generally have an increase in pitch mean and range, an increase in energy (with anger having more energy in high frequencies), and increased an articulation rate. Those with lower activation levels such as sadness, disgust, and neutral, have narrower F0 range, lower F0 mean, and generally a flatter pitch contour. Energy levels for these are also lower, and the speaking rate is decreased.

There are some disagreements in the literature, for example the speaking rate of happiness is shown to be varied and fear reported to have both a wide and narrow pitch range. There are reasons as to why these disagreements may occur. First, as discussed in Chapter 3, the methods used for data collection vary between studies. Some use actors to acquire emotional data, while others rely on data induced from naïve participants, and more rarely data is acquired from natural settings. This can explain why some properties of vocal emotion are different. Second, as discussed in Chapter 2, there are different levels of the basic emotions. For example frustration, often classed simply as anger, can differ significantly from hot anger, where a person is extremely agitated.

Barring these disagreements, the correlations presented allow for the construction of models which can appropriately differentiate between the basic emotions. These models will be introduced in Chapter 6. The next logical step is the development of algorithms which can extract the aforementioned features from the speech samples in the dataset in preparation for classification. The methods and algorithms utilised for feature extraction are presented in the following chapter.

Chapter 5

Feature Extraction

5.1 Introduction

In the previous chapter, characteristics of the basic emotional states were reviewed. It was found that the prosody-based features consisting of fundamental frequency, energy, and rhythm carry the majority of emotional information during vocal communication. The formant frequencies were also covered, however research on these with respect to emotion is sparse.

This chapter aims to compile an exhaustive list of the acoustic correlates such that they can be input to classification models which will ultimately validate or invalidate these choices of features. We will first introduce features used in previous research and subsequently justify our choice of features based on this knowledge.

This chapter also goes into detail on the extraction methods used for each group of features. We will show that certain methods for pitch tracking are inadequate for the type of data we have acquired, and utilise a more advanced method for generating a cleaner pitch contour.

5.2 Features used in past works

The features used in past research are generally similar between studies. Researchers focus mainly on the fundamental frequency and energy related features, but some studies have explored other techniques. Dellaert *et al.* (1996) composed two feature sets: the first with seven basic prosodic features (F0 mean, standard deviation, minimum, maximum, range, slope, and speaking rate) and the second with 17 prosodic features extracted using a smoothing spline approximation of the pitch contour. Prosodic features such as speaking rate, pitch, formant frequencies (F1, F2) and intensity were utilised in a study by Petrushin (2000). Sato *et al.* (2001) employed prosodic features (accent, intonation, pitch structure, stress, tempo, and rhythm among others). Pitch information was extracted by using cepstrum analysis. In another study, 37 prosody-based features were considered from several categories: pitch contour, first deriva-

tive of the pitch contour, jitter, energy contour, first derivative of the energy contour, shimmer, and durational features (Yacoub *et al.*, 2003). Nwe *et al.* (2003b) used a novel approach by locating prosodic clues to emotion in log frequency power coefficients extracted from the frequency domain in speech samples. These features indicated different levels of intensity in the frequency bands for different emotions and closely model the frequency range of the human ear. McGilloway *et al.* (2000) used 32 prosodic features in 11 groups: relating to tune, spectral, intensity contour, intensity at local extrema in the intensity contour, magnitude of rises or falls in intensity contour, pitch of points in F0 contour, pitch at local extrema in F0 contour, magnitude of rises in F0 contour, duration of rises and falls in intensity contour, durations of level sections in intensity contour, durations of features in F0 contour.

5.3 Chosen features

Based on the acoustic correlates described in the previous section and the literature relating to automatic emotion detection from speech, we selected features based on four prosodic groups: the *fundamental frequency, energy, rhythm*, and the *formant frequencies*.

The fundamental frequency, energy, and formant frequencies are represented as contours. From these contours, we selected seven statistics: the *mean, minimum, maximum, standard deviation, value at the first voiced segment, value at the last voiced segment*, and the *range*.

For the rhythm-based features, we selected three: *speaking (articulation) rate, average length of unvoiced segments (pause)*, and *average length of voiced segments*.

In total, we selected 38 prosodic features which are used as a starting point for describing the variation between the basic emotions. These are listed in Table 5.1.

5.4 Extraction methods

With the exception of those relating to rhythm, all features were calculated over the voiced segments of the sample. In general, a frame can be flagged as unvoiced if it has no value for the fundamental frequency (Rabiner and Juang, 1993). In other words, unvoiced frames contain little or no periodicity and are viewed as modelling random noise.

Tools from the Speech Filing System (SFS)[1] were used to extract and calculate statistics on each contour-based feature group. Figure 5.1 shows the data flow and relationships between the different extraction algorithms.

[1]The Speech Filing System can be downloaded from `http://www.phon.ucl.ac.uk/resource/sfs/`.

Table 5.1: 38 prosodic features selected for input into classification algorithms. Features are divided into six groups: fundamental frequency (F0), first three formant frequencies (F1, F2, F3), short-time energy, and rhythm.

Number	Description
1	F0 mean
2	F0 minimum
3	F0 maximum
4	F0 standard deviation
5	first F0 value
6	last F0 value
7	F0 range
8	F1 mean
9	F1 minimum
10	F1 maximum
11	F1 standard deviation
12	first F1 value
13	last F1 value
14	F1 range
15	F2 mean
16	F2 minimum
17	F2 maximum
18	F2 standard deviation
19	first F2 value
20	last F2 value
21	F2 range
22	F3 mean
23	F3 minimum
24	F3 maximum
25	F3 standard deviation
26	first F3 value
27	last F3 value
28	F3 range
29	energy mean
30	energy minimum
31	energy maximum
32	energy standard deviation
33	first energy value
34	last energy value
35	energy range
36	speaking rate
37	average length of unvoiced segments
38	average length of voiced segments

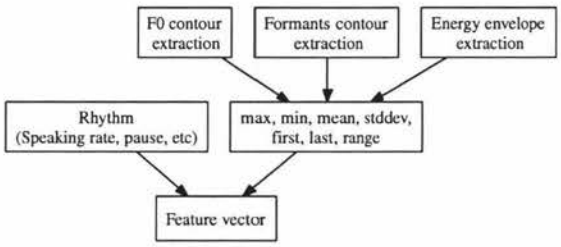


Figure 5.1: Data flow diagram for the feature extraction process

5.4.1 Methods for pitch tracking

The determination of a fundamental frequency contour is called pitch tracking. With pitch tracking the goal is to estimate the fundamental frequency for an utterance over a series of frames. These frames can overlap, but it is not necessary. The resulting points in a pitch contour are represented in Hz (samples/second).

There are two reasons for pitch tracking. The first is that for a specific frame, a decision can be made whether that frame is voiced or unvoiced. The second benefit is that if a frame is found to be voiced, then for this frame we can calculate the fundamental frequency.

Experiments are performed using some of the data to determine an optimal pitch estimation method. Two methods are considered and are discussed below.

Autocorrelation

The autocorrelation method is considered first. The autocorrelation method is a useful time-domain technique for determining several different characteristics of the speech signal. One of the side-effects is that "the autocorrelation function of a periodic signal is [itself] periodic with the same period" (Rabiner and Schafer, 1978). Thus, it can be used to calculate the fundamental frequency for a given window of speech.

One of the caveats of using the autocorrelation function is the selection of an optimal window size. Rabiner and Schafer (1978) state that we are bound by two constraints: on the one hand, we must make the window large enough such that at least two periods of the signal are captured, and on the other hand, we want the window small enough such that there is high enough resolution to model higher frequencies.

Forty Hz is a lower bound for a male speaker with a low voice, and a woman or child with a high voice can often reach frequencies of around 500 Hz. Rabiner and Schafer (1978) suggest that the window size for the autocorrelation function may be dynamically selected depending on the expected pitch, however, in practice this is not usually desired. Instead, a middle-ground value of somewhere between 10-30 milliseconds is typically used. Nwe (2003) used a frame size

of 30 ms updated every 20 ms. In other words, the consecutive frames overlap each other by 10 ms. Nwe (2003) also employed centre-clipping as a spectral flattening technique, which yields a less complex autocorrelation function, as recommended by (Rabiner and Schafer, 1978).

The autocorrelation method is defined as

$$\Phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m-k) \tag{5.1}$$

Typically the autocorrelation method is used in conjunction with a Hamming window:

$$w_H(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{5.2}$$

Windowing functions like the Hamming window help reduce spectral leakage while maintaining the periodicity information by tapering the edges of the frame gradually towards zero.

In our initial experiments, the autocorrelation method worked well for noise-free speech, but when applied to the NATURAL dataset, which contained significant distortion due to noise, the performance was less than adequate.

The Robust Algorithm for Pitch Tracking

The second algorithm is the Robust Algorithm for Pitch Tracking (RAPT). This algorithm uses the cross correlation function to identify pitch candidates and then attempts to select the "best fit" at each frame by dynamic programming. One of the benefits of using the cross correlation function is that it does not suffer the windowing dilemma of the autocorrelation function while maintaining resolution for high pitch values and the ability to detect low pitch values (Rabiner and Schafer, 1978).

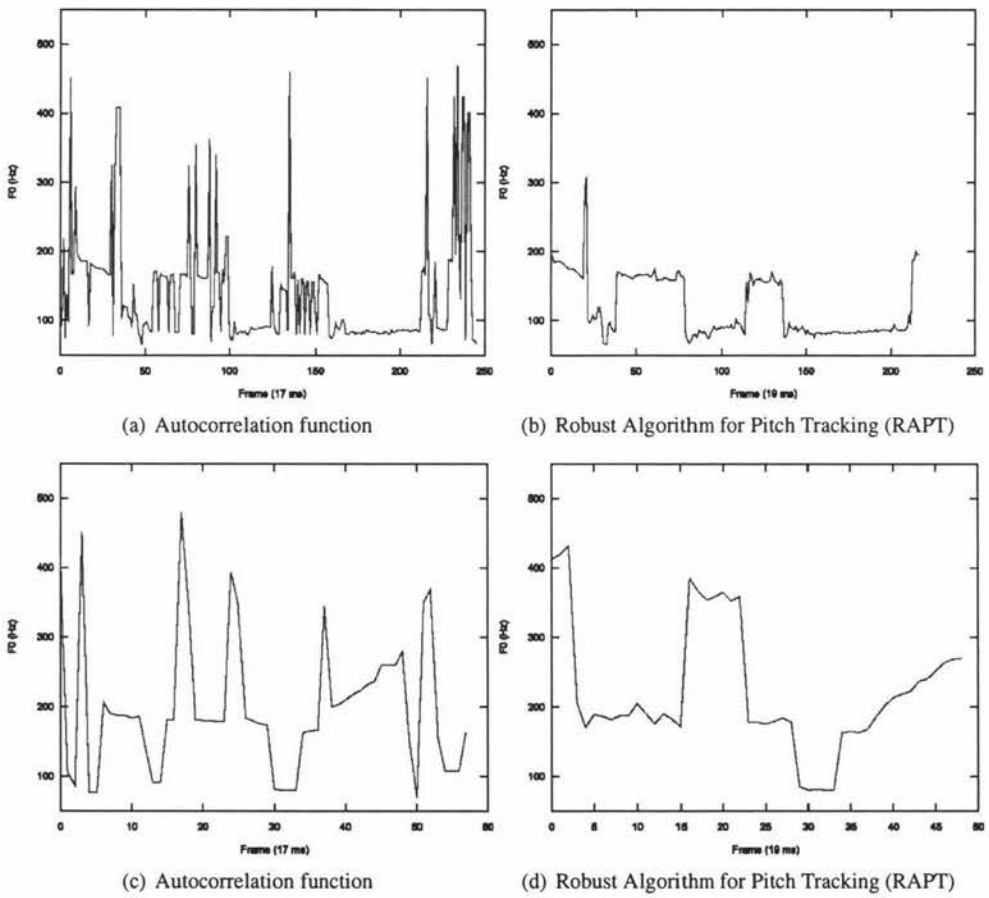
The RAPT is performed as follows. First, the signal is downsampled to roughly 2000 Hz (for decreased computational requirements). Using the downsampled signal, a set of F0 candidates are computed for each 7.5 ms frame using the cross correlation function

$$R_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k), \quad 0 \leq k \leq K \tag{5.3}$$

These candidates are ranked according to their frame-relative properties: amplitude, location, and relation to each of the candidates in the previous frame. For each peak candidate, a "fine-grained" cross correlation is computed, giving increased resolution for the signal area near the peak. The F0 contour is then generated by dynamic programming which selects the optimal candidate for the current trajectory based on the ranks calculated previously. For full detail on the RAPT, the reader is directed to (Talkin, 1995).

Figures 5.2 (a) and (c) show example pitch contours calculated using the autocorrelation

method. It can be observed that the contours suffer from frequency- doubling and halving and are littered with spurious points. Figures 5.2 (b) and (d) show example pitch contours calculated using the RAPT algorithm described above. The contours are much smoother as a result of the optimum epoch detection by dynamic programming after candidates are chosen. This is extremely important, as F0 is one of the most important features for emotion recognition, and statistics based on it must be as accurate as possible.



(a) Autocorrelation function

(b) Robust Algorithm for Pitch Tracking (RAPT)

(c) Autocorrelation function

(d) Robust Algorithm for Pitch Tracking (RAPT)

Figure 5.2: Comparison between the autocorrelation method and RAPT for pitch tracking for two sample utterances from the NATURAL dataset. (a) and (b) show the differences for the first sample, and (c) and (d) show the differences for the second sample.

5.4.2 Formant frequencies

The first three formant frequencies are extracted using linear predictive coding (LPC) and dynamic programming to select optimal candidates based on their scores in relation to previous candidates, similar to the RAPT. Linear predictive analysis is a technique for aproximating the speech signal as a "linear combination of past speech samples" (Rabiner and Schafer, 1978). LPC exploits the periodic nature of the speech signal and the LPC coefficients are estimated by minimising the mean squared error. The minimum mean squared error is defined as the difference between the actual and predicted speech samples:

$$E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k) \tag{5.4}$$

Figure 5.3 shows a block diagram of generating the LPC coefficients from a speech signal and the reconstruction of the signal using the same coefficients. The signal is passed through the inverse impulse response filter $H(z)$, which is defined as

$$H(z) = 1/A(z) \tag{5.5}$$

where

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}. \tag{5.6}$$

(a) Extraction of the LPC parameters by passing the speech signal through the inverse impulse response filter $H(z)$.

(b) Reconstruction of speech by passing the LPC parameters through the impulse response filter $A(z)$.

Figure 5.3: Block diagrams depicting the (a) extraction of and (b) reconstruction using the linear prediction coding coefficients.

First, the normalised autocorrelation coefficients are computed using Eq. (5.1). As shown in Rabiner and Schafer (1978), Eq. (5.1) can be written as

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \quad i = 1, 2, \ldots, p \tag{5.7}$$

where these p equations can be solved for the predictor coefficients $\{\alpha_k\}$. Subsequently, Eq. (5.7) can be shown to be

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \quad 1 \le i \le p \tag{5.8}$$

and these equations can be represented in matrix form:

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \dots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \dots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \dots & R_n(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \dots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \dots \\ \dots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \dots \\ \dots \\ R_n(p) \end{bmatrix} \tag{5.9}$$

The matrix has Toeplitz properties and can be solved for the p coefficients, α, by using Durbin's recursive solution (Eqs. (5.10) to (5.15)[2]).

$$E^{(0)} = R(0) \tag{5.10}$$

$$k_i = \left(R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \right) / E^{(i-1)}, \quad 1 \le i \le p \tag{5.11}$$

$$\alpha_i^{(i)} = k_i \tag{5.12}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \quad 1 \le j \le i-1 \tag{5.13}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{5.14}$$

$$\alpha_j = \alpha_j^{(p)}, \quad 1 \le j \le p. \tag{5.15}$$

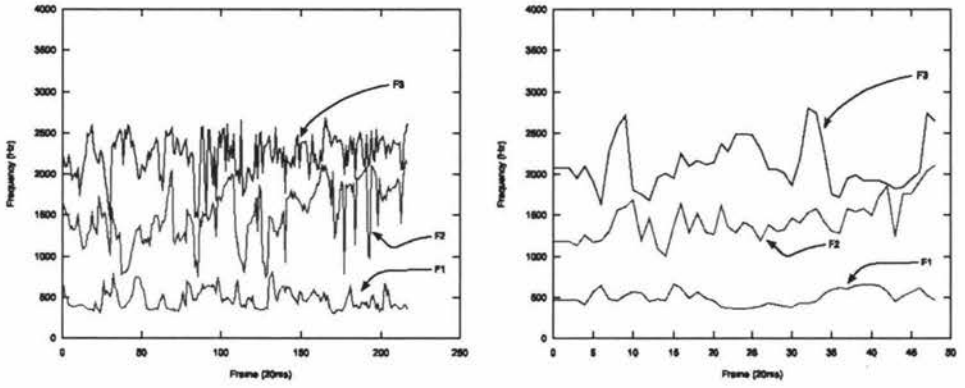
The formant candidate locations ($z_k, k = 1, 2, \dots, p$) are solved in the impulse response of the filter

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = \prod_{k=1}^{p} (1 - z_k z^{-1}). \tag{5.16}$$

The candidates are then ranked according to their relative location, bandwidth, and relation to the previous formant candidates. The best candidates are selected for each formant using dynamic programming similar to that used for the RAPT (Rabiner and Schafer, 1978).

[2]From (Rabiner and Schafer, 1978).

Nwe (2003) also used this technique for the extraction of formant frequencies, although the results of her experiments indicate that the first two formant frequencies (F1 and F2) carry little emotional correlation. Figure 5.4 shows the first three formant frequencies for example utterances from the NATURAL dataset.



(a) Sample 1 formant frequencies (b) Sample 2 formant frequencies

Figure 5.4: Two sample formant frequency contours calculated using a 20 ms window on example utterances from the NATURAL dataset. The first formant (F1) has the lowest frequency, followed by F2, followed by F3 with the highest frequency.

5.4.3 Short-time energy

The energy of a signal is frequently referred to as the amplitude, volume, or intensity. In a nutshell, the energy describes the loudness of the signal, usually in decibels (dBs).

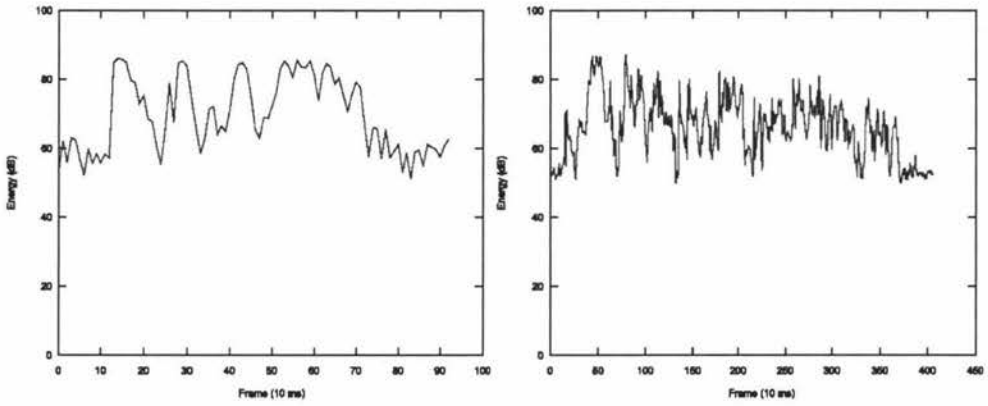
In order to utilise the energy information, it must first be represented as a contour or envelope. The envelope consists of the magnitude of the signal calculated over a frame or window in order to average or smooth the contour. Like frames for the autocorrelation function, the energy frame size should be long enough to smooth the contour appropriately but short enough to retain the fast energy changes which are common in speech signals (Rabiner and Schafer, 1978).

Rabiner and Schafer (1978) suggest a frame size of 10-20 ms would be adequate. Nwe (2003) used a frame size of 5 ms. In this research, we will use a frame size of 10 ms, as this is the default value for the "envelope" tool from the SFS package and from initial evaluations was deemed appropriate for our purposes.

Short-time energy is computed for a window size N using the following equation.

$$E_n = 10 \log_{10} \left(\sum_{n=0}^{N} x(n)^2 \right) \tag{5.17}$$

where $x(n)$ is the signal sample at position n in the window. The resulting energy envelope is represented in decibels. The use of short-time energy is also a convenient way of determining the endpoints of speech, and this will be discussed in more detail in Chapter 7. Figure 5.5 shows the resulting energy envelopes for example samples from the NATURAL dataset.



(a) Sample 1 energy envelope (b) Sample 2 energy envelope

Figure 5.5: Two sample energy envelopes calculated using a 10 ms window on example utterances from the NATURAL dataset.

5.4.4 Rhythm-based statistics

The rhythm-based statistics are all based on the voiced and unvoiced segment durations. The rate of speech (articulation) is estimated by counting the number of syllables, which is roughly equal to the number of voiced-to-unvoiced and unvoiced-to-voiced transitions (hereafter referred to only as voiced-unvoiced transitions) during the utterance. Recall from Section 5.4 that a segment (one or more consecutive frames) is deemed to be voiced if it is periodic, in other words if it has a value greater than zero for the fundamental frequency. A segment is unvoiced if it is aperiodic, or has no fundamental frequency.

The number of voiced-unvoiced transitions for the entire utterance, $Vunv$, is defined as:

$$Vunv = \sum_{i=0}^{m} |sgn[x(i)] - sgn[x(i-1)]| \tag{5.18}$$

where

$$sgn[x(k)] = \begin{cases} 1 & 50 \geq x(k) \geq 550 \\ -1 & \text{otherwise} \end{cases}$$

where 50 and 550 are the minimum and maximum allowable values for F0, respectively.

Thus, the rate of speech, r_{speech}, is defined by simply extending Eq. (5.18) by dividing by the utterance duration, D_u:

$$r_{speech} = \frac{Vunv}{D_u} \tag{5.19}$$

where D_u is represented in seconds and begins at the first voiced segment and ends at the last voiced segment in the utterance.

The average length of the unvoiced segments, or pause, is calculated as:

$$AvD_{unvoiced} = \frac{\sum_{n=0}^{N} x_{unvoiced}(n)}{Vunv/2} \tag{5.20}$$

where $x_{unvoiced}(n)$ is the unvoiced frame at position n of N in the segment. Similarly, the average length of voiced segments is defined as:

$$AvD_{voiced} = \frac{\sum_{n=0}^{N} x_{voiced}(n)}{Vunv/2} \tag{5.21}$$

5.5 Summary

In this chapter we introduced and explained the feature groups and statistics that were chosen for representing the characteristics of emotional speech. Briefly, these were the fundamental and first three formant frequency contours, energy contour, speaking rate, average voiced, and average unvoiced durations.

We also described the algorithms used for extracting the contours and other statistics. For the fundamental frequency, two methods were compared: the autocorrelation function and the Robust Algorithm for Pitch Tracking. The RAPT is significantly better at producing smooth pitch contours. For the formant frequencies, linear predictive coding and dynamic programming are used to calculate the optimal candidates for F1, F2, and F3. The energy envelope is calculated by finding the absolute value for each frame in the signal. The speaking rate is estimated by counting the number of voiced-unvoiced and unvoiced-voiced transitions.

The statistics calculated for each contour are the mean, minimum, maximum, standard deviation, value at the first voiced segment, value at the last voiced segment, and the range. Finally, these statistics are combined into a 38-element feature vector. This feature vector is used in the following chapter for classification and feature selection.

In the next chapter, classification and feature selection methods are presented. These methods are trained using the feature vector generated from the methods in this chapter.

Chapter 6

Classification

6.1 Introduction

In the previous chapter we described the features we selected to describe the majority of variance between the different emotional states. In addition, it was illustrated how these features are extracted from speech signals and subsequently represented as a feature vector.

This chapter presents an overview of the classification algorithms selected. The feature vectors compiled in the previous chapter are used to train the classification algorithms and test their generalisation accuracy. Furthermore, in order to reduce the dimensionality of the feature vectors and in turn reduce the complexity of training the classification algorithms, we describe how feature selection techniques are applied to the dataset.

We also introduce the ensemble methods stacked generalisation (specifically StackingC) and an unweighted voting scheme. These ensemble classifiers are built by combining the predictions of several traditional classification algorithms.

6.2 Traditional classification approaches

In past studies, many different classification schemes have been applied to vocal emotion recognition. A review of these is presented in Chapter 2. It is typical for any applied classification research to determine optimal generalisation accuracy by comparing the results of several classification algorithms on the same dataset. Further optimisation can be achieved by employing data reduction or feature selection techniques. These techniques are covered in Section 6.5.

During the selection of an optimal classification method, it is standard practice to perform the evaluation over a series of classifications, each time using a different portion of the dataset for training the classifier. This is called *cross-validation*, and involves dividing the data into subsets or *folds*. Normally, the number of folds is set to 10, and thus is called 10-fold cross-validation. Moreover, these individual cross-validations should run multiple times using a differ-

ent random seed from which the dataset is partitioned, which accounts for the variance between different partitions of the dataset. This process is called *stratification* (Witten and Frank, 2000). Unless otherwise noted, all classification evaluations in this thesis use stratified 10-fold cross-validation. We revisit cross-validation in more detail in Section 6.4.

Classification was performed using WEKA (Waikato Environment for Knowledge Analysis). WEKA is a data mining workbench that allows comparison between many different machine learning algorithms. In addition, it also has functionality for feature selection, data pre-processing, and data visualisation.

The selection of base classifiers was done by evaluating several algorithms over the NATURAL dataset and selecting the top performers. Table 6.1 shows the classification accuracies for the algorithms initially selected. In order to retain some degree of simplicity, only the top five algorithms are retained. As can be seen, the top performers are the support vector machine (SVM) with the radial basis function (RBF) kernel, the random forest, the multi-layer perceptron (artificial neural network), K^* instance-based classifier, and K-nearest neighbour with $K = 5$. For the SVM, the use of the RBF kernel showed a significant improvement over the use of the polynomial kernel. The selected algorithms are discussed in detail in the following sections.

Table 6.1: Initial ranking of base classification algorithms on the NATURAL dataset.

Algorithm	Accuracy (%)
SVM (RBF)	76.93
KNN (K=5)	75.85
Multi-layer Perceptron	74.25
Random Forest	71.98
K^*	70.67
Naive Bayes	69.56
SVM (polynomial)	69.50
C4.5 Decision tree	67.47
Random Tree	60.05

6.2.1 Support vector machines

Support vector machines (SVMs) are a machine learning algorithm introduced by Vapnik (1995). They are based on the statistical learning theory of structural risk management (SRM) which aims to limit the empirical risk on the training data and on the capacity of the decision function. Support vector machines are built by mapping the training patterns into a higher dimensional feature space where the points can be separated using a hyperplane. The separating hyperplane is determined by a kernel (see Eqs. (6.5) and (6.6)). Figure 6.1 shows an example mapping from

input space to feature space via the mapping function Φ.

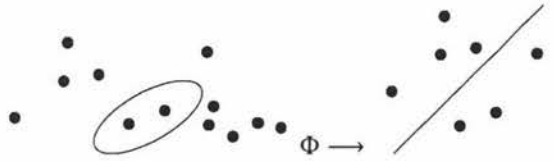


Figure 6.1: Example support vector mapping from input space to feature space.

The mapping function Φ is defined as

$$\Phi : R^N \rightarrow F \tag{6.1}$$

where R^N is the N-dimensional training data and F is the resulting feature space. Training data comprising l examples is defined as $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)$ where $\mathbf{x}_i$ is the input space (feature vector) and y_i is the class label. The support vectors define a margin with maximum distance to the separating hyperplane. The optimal hyperplane is selected from a vector of candidate hyperplanes, $\mathbf{w}$, by minimising

$$\|\mathbf{w}\|^2 + \gamma \cdot \sum_{i=1}^{l} \xi_i \tag{6.2}$$

which is subject to

$$\xi_i \geq 0, y_i \cdot ((x_i \cdot w) + b) \geq +(1 - \xi_i), \quad \text{for } i = 1, ..., l \tag{6.3}$$

(Schölkopf *et al.*, 1996), which becomes

$$\Psi(\mathbf{w}, \xi, \xi^*) = \frac{1}{2}(\mathbf{w}^T \mathbf{w}) + C \sum_{i=1}^{l} \xi_i \tag{6.4}$$

In WEKA, SVMs are implemented as the sequential minimal optimisation (SMO) algorithm (Platt, 1998). There are two kernels available: polynomial,

$$K(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^T \mathbf{y} + r)^d, \quad \gamma > 0 \tag{6.5}$$

and radial basis function (RBF),

$$K(\mathbf{x}, \mathbf{y}) = exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2), \quad \gamma > 0. \tag{6.6}$$

Optimal values for the width of the RBF function, γ, and the cost parameter C, can be found

by performing a grid search on the training data (Chang and Lin, 2001). For our experiments, a grid search of the training data yielded optimal values $\gamma = 0.7$ and $C = 8.0$.

Figure 6.2 shows an example support vector classifier. The support vectors are derived from the training data such that the distance between them and the separating hyperplane is maximised.

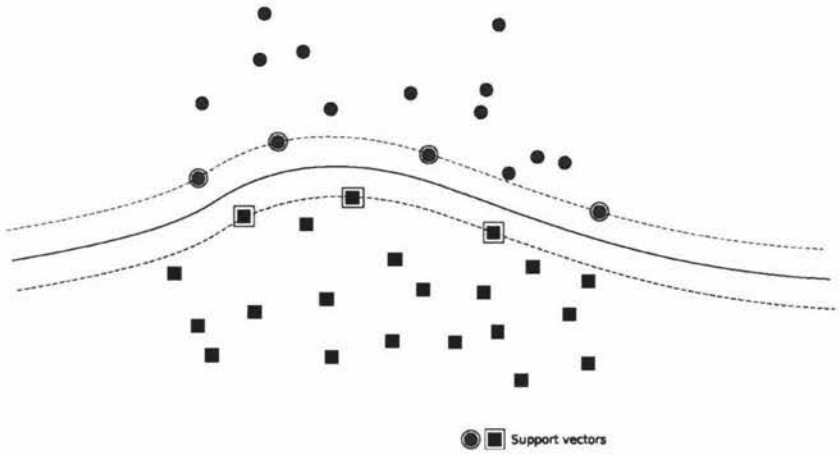


Figure 6.2: An example support vector machine using the radial basis function. The support vectors are represented by the outlined shapes and constitute a maximum margin from the decision surface (solid line).

6.2.2 Random Forests

Random forests, invented by Breiman (2001), consist of ensembles of tree predictors. These tree ensembles are a method of growing a "forest" of decision trees by selecting features for each node randomly and independently of every other tree but with the same distribution. When a random forest has been grown, classification requires that the predictions of each tree are combined by voting to determine the overall prediction.

The algorithm is described as follows: for each tree k, a random vector Θ_k is generated, with the same distribution of, yet independent of all previous random vectors $\Theta_1, ...\Theta_{k-1}$, from which a tree predictor $h(\mathbf{x}, \Theta_k)$ is grown using the training data and Θ_k, where $\mathbf{x}$ is an input vector. When a sufficient forest size has been attained, a vote is performed to decide the most popular prediction.

Breiman (2001) states that if we let $h_1(\mathbf{x}), h_2(\mathbf{x}), ..., h_k(\mathbf{x})$ be an ensemble of classification trees with random training vector $Y, \mathbf{X}$, then, with I as the indicator function, the margin is defined as

$$mg(\mathbf{X}, Y) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j) \tag{6.7}$$

and "measures the extent to which the average [(av)] number of votes at $\mathbf{X}, Y$" for the correct class exceed that for any other class (Breiman, 2001). The generalisation error of a random forest is determined by

$$PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X}, Y) < 0) \tag{6.8}$$

where $P_{\mathbf{X},Y}$ is the probability over the $\mathbf{X}, Y$ feature space (Breiman, 2001).

It is claimed that overfitting by random forests is avoided because by the use of the Strong Law of Large Numbers, they always converge. For the proof of convergence and greater detail about random forests, the reader is directed to (Breiman, 2001).

Although decision trees have been used previously for vocal emotion classification, ensembles of trees in random forests have not. In this thesis, random forests are employed as base classifiers in our ensemble classification system.

6.2.3 Artificial neural networks

Artificial neural networks, specifically multi-layer perceptrons (MLPs), have proved useful for research in emotion recognition from speech (Huber *et al.*, 1998; Nakatsu *et al.*, 1999; Huber *et al.*, 2000; Petrushin, 2000). A MLP consists of a set of layers containing nodes. The layers of a MLP are linked by weighted synaptic connections between nodes. Every network has an input and output layer, but networks can also have one or more hidden layers which are located between the input and output layers. Figure 6.3 shows an example one-hidden layer network.

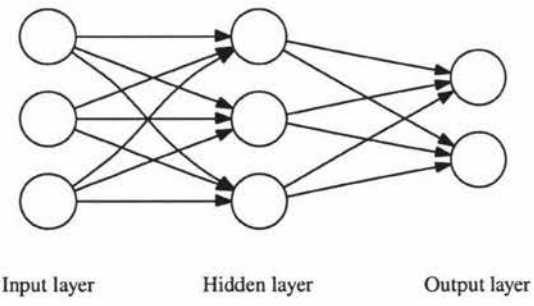


Figure 6.3: An example one-hidden layer artificial neural network architecture. Circles represent the nodes in each layer. The input layer contains nodes which correspond to each feature in the input vector. The output layer contains nodes that carry the result of the propagation of information throughout the network.

There are many types of ANNs, but they can all be divided among two groups: supervised and unsupervised. Supervised networks require feature vectors to be accompanied by the intended classification in what is called the learning phase. During this phase the network "learns" how patterns should be classified by adjusting the weights at each of the nodes depending on the target classification during a process called "error back-propagation" (Haykin, 1999). The rate at which these weights are updated is called the learning rate. Haykin (1999) states that the selection of the learning rate is an important factor: if the learning rate is too fast, the network will be unstable, with the synaptic weights being updated too much, resulting in a coarse error surface. If the learning rate is too slow, the resulting network will be stable with a smooth error surface, but the training time may be too great. In unsupervised networks, there is no learning phase and the network only groups similar features, sometimes called regression or clustering, by way of some kernel method. All supervised and unsupervised networks can be categorised into three general neural network topologies: single/multi-layer perceptrons, self-organising (Kohonen) networks, and feedback or recurrent networks (Hopfield) (Rabiner and Juang, 1993). In the WEKA toolkit, ANNs are implemented as the multi-layer perceptron.

The total error energy for the neurons in the output layer is defined as

$$\xi(n) = \frac{1}{2} \sum_{j \in C} e_j^2(n) \tag{6.9}$$

where C is the set of all neurons in the output layer and $e_j(n) = d_j(n) - y_j(n)$ is the error signal for an individual neuron j.

The output signal $y_k(n)$ is defined as

$$y_k(n) = \sum_{l=0}^{\infty} w^{l+1} x_j(n-l) \tag{6.10}$$

where $x_j(n)$ is the input signal and the activation function associated with each neuron j is

$$v_j(n) = \sum_{i=0}^{m} w_{ji}(n) y_i(n) \tag{6.11}$$

where m is the number of inputs to j and w_{ji} is the synaptic weight (Haykin, 1999).

As can be seen from the hidden layer experiment in Table 6.2, the MLP displays slightly better performance with a hidden layer containing 60 nodes. It can also be seen that the accuracy is not significantly affected with more or less hidden nodes. An early stopping criteria based on a validation set consisting of 10% of the training set is used in all classification experiments involving the MLP. This ensures that the training process stops when the mean-square error (MSE) (see Eq. (6.12)) begins to increase on the validation set and reduces overfitting (Haykin, 1999). The learning rate was set to 0.2 which is the default setting in WEKA.

Table 6.2: Results for the selection of the number of nodes in the hidden layer of the multi-layer perceptron

# of nodes	Accuracy (%)
20	74.15
30	73.58
40	73.56
50	74.18
60	74.25

The mean-square error of a trained network for M validation examples $(\mathbf{y}_i, \hat{x}_i)$ is defined as

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(\hat{x}_i - x_i)^2 \tag{6.12}$$

where $\hat{x}_i$ is the target class value and x_i is the predicted class value for input vector $\mathbf{y}_i$.

6.2.4 K^* instance-based classifier

K^* is an instance-based learning algorithm based on the work of Cleary and Trigg (1995). It uses a similarity function to classify test cases based on training cases which have a high similarity. In this way, it is much like the K-nearest neighbour method (described below), however, the distance measure used by K^* is based on entropy (Cleary and Trigg, 1995).

A probability function, P^* is defined "as the probability of all paths from instance a to instance b" (Cleary and Trigg, 1995):

$$P^*(b|a) = \sum_{\tilde{t}\in P:\tilde{t}(a)=b} p(\tilde{t}) \tag{6.13}$$

and satisfies

$$\sum_b P^*(b|a) = 1 \tag{6.14}$$

and

$$0 \leq P^*(b|a) \leq 1. \tag{6.15}$$

It follows that K^* is defined as

$$K^*(b|a) = -\log_2 P^*(b|a). \tag{6.16}$$

The number of instances included in the probability distribution is defined as

$$n_0 \leq \frac{(\sum_b P^*(b|a))^2}{\sum_b P^*(b|a)^2} \leq N \tag{6.17}$$

where N is the number of training cases, n_0 is the number of training cases "at the smallest distance from a," and b is the "blending parameter" which has a default value of 20% and determines the number of neighbours considered significant to a (Cleary and Trigg, 1995). In our experiments, we left $b = 20\%$.

Category prediction is calculated by summing the probabilities from the current instance a to each instance that is a member of C:

$$P^*(C|a) = \sum_{b \in C} P^*(b|a) \tag{6.18}$$

Further detail on K^* can be found in the paper by Cleary and Trigg (1995).

6.2.5 K-nearest neighbours

K-nearest neighbours is another instance-based classification method introduced by Cover and Hart (1967). This algorithm has proved popular with vocal emotion recognition (Dellaert *et al.*, 1996; Petrushin, 1999; Yacoub *et al.*, 2003) due to its relative simplicity and performance comparable to other methods.

As with the K^* algorithm, the assumption for instance-based classifiers is that new instances will have the same class as pre-classified instances if they are close in feature space. For the K-nearest neighbour classifier, the nearest K neighbours of the current instance are retrieved (from some database of training instances) and the target class which the majority share is used as the class for the current instance (Cleary and Trigg, 1995). In our experiments, setting $K = 5$ performed best on the NATURAL dataset. More information can be found in Aha and Kibler (1991).

6.3 Ensemble classification methods

Several methods for creating ensembles of classifiers have been introduced in the literature of machine learning, pattern recognition, data mining, and artificial intelligence. These include weighted and unweighted voting, boosting, bagging (booststrap aggregation), AdaBoost, and stacked generalisation among others.

Ensembles of classifiers generally combine several base classification schemes into a larger meta classifier. For ensemble classifiers to improve over the best performing base classifier, they must comprise accurate base classifiers. However, the base classifiers must also have high disagreement between one another in order to maintain diversity (Dietterich, 2002). For example, if a voting scheme is made up of several highly accurate base classifiers that cast the same prediction, then there is little improvement over simply using one of the base classifiers. The

complexity involved in building the meta classifier must be outweighed by the improvement in classification accuracy.

6.3.1 Unweighted vote

Voting takes two forms: weighted and unweighted. With weighted voting, a set of classifiers, which may be of different or similar class, output predictions as class probabilities. Each classifier casts a vote coinciding with the predicted class, but classifiers which have higher or lower class probabilities have their votes weighted such that they contribute more or less to the vote.

With unweighted voting, the predictions of the base-level classifiers are summed for each class and the class with the highest number of votes determines the prediction for the ensemble (Shipp and Kuncheva, 2002).

The voting scheme we use is built by combining the aforementioned base classifiers: SVM with RBF kernel, random forest, K^* instance-based learner, KNN with $K = 5$, and multi-layer perceptron. Under this ensemble scheme, each classifier is trained with the same data. To measure performance, a test set is presented to each base classifier. The class predictions from each base classifier are then counted and the target class with the most votes is selected as the final prediction.

6.3.2 Stacked generalisation

Stacked generalisation, or stacking, is an approach to combining predictions from multiple classifiers. Introduced by Wolpert (1992), this method takes the predicted target classes of several different (or similar) base or level-0 classifiers and uses those to train a meta-learner or level-1 classifier. The meta-learner, typically a series (one for each target class) of linear models such as multi-response linear regression (MLR), uses the level-0 predictions and the target classes to determine which classifiers are correct or incorrect and generates a higher level prediction based on this.

StackingC, introduced by Seewald (2002b), is an improvement over the original algorithm. It works by using only class probability distributions which are associated with the target class during training and testing. This has the effect of reducing the dimensionality of the meta-learning phase. The learning process is substantially faster by the encoding of meta-data in the level-1 feature space and uses prediction probabilities rather than actual target classes, which improves performance on multi-classed data. These prediction probabilities carry confidence information, which, when combined with a multi-response linear regression meta-learner, offers a modest improvement in multi-class problems.

A comparison of the meta training sets used for Stacking and StackingC is shown in Figure 6.4. Figure 6.4 (a) shows the original training set with the associated target class. Figure 6.4 (b)

Attrs	*Class*
$AttrVec_1$	a
$AttrVec_2$	b
$AttrVec_3$	b
$AttrVec_4$	c
⋮	⋮
$AttrVec_n$	a

(a) original training set

a	b	c
0.90	0.05	0.05
0.15	0.70	0.15
0.10	0.80	0.10
0.20	0.20	0.60
⋮	⋮	⋮
0.80	0.10	0.10

(b) class probability distribution

$Classifier_1$			$Classifier_2$				$Classifier_N$			
a	b	c	a	b	c		a	b	c	class = a?
$P_{1,a1}$	$P_{1,b1}$	$P_{1,c1}$	$P_{2,a1}$	$P_{2,b1}$	$P_{2,c1}$	⋯	$P_{N,a1}$	$P_{N,b1}$	$P_{N,c1}$	1
$P_{1,a2}$	$P_{1,b2}$	$P_{1,c2}$	$P_{2,a2}$	$P_{2,b2}$	$P_{2,c2}$	⋯	$P_{N,a2}$	$P_{N,b2}$	$P_{N,c2}$	0
$P_{1,a3}$	$P_{1,b3}$	$P_{1,c3}$	$P_{2,a3}$	$P_{2,b3}$	$P_{2,c3}$	⋯	$P_{N,a3}$	$P_{N,b3}$	$P_{N,c3}$	0
$P_{1,a4}$	$P_{1,b4}$	$P_{1,c4}$	$P_{2,a4}$	$P_{2,b4}$	$P_{2,c4}$	⋯	$P_{N,a4}$	$P_{N,b4}$	$P_{N,c4}$	0
	⋮			⋮				⋮		⋮
$P_{1,an}$	$P_{1,bn}$	$P_{1,cn}$	$P_{2,an}$	$P_{2,bn}$	$P_{2,cn}$	⋯	$P_{N,an}$	$P_{N,bn}$	$P_{N,cn}$	1

(c) meta training set for class a, Stacking with MLR

$Classifier_1$	$Classifier_2$		$Classifier_N$	
a	a		a	class = a?
$P_{1,a1}$	$P_{2,a1}$	⋯	$P_{N,a1}$	1
$P_{1,a2}$	$P_{2,a2}$	⋯	$P_{N,a2}$	0
$P_{1,a3}$	$P_{2,a3}$	⋯	$P_{N,a3}$	0
$P_{1,a4}$	$P_{2,a4}$	⋯	$P_{N,a4}$	0
⋮	⋮		⋮	⋮
$P_{1,an}$	$P_{2,an}$	⋯	$P_{N,an}$	1

(d) meta training set for class a, StackingC with MLR

Figure 6.4: Illustration of Stacking and StackingC on a three-class dataset (a, b, c) with n training examples and N base classifiers. $P_{i,jk}$ denotes the class prediction from classifier i for class j on example k (from (Seewald, 2002b)).

lists the probability distribution for each example. Figure 6.4 (c) shows the meta training set for Stacking. It can be seen that the training set includes the probablilities for each class whereas for StackingC, only the probabilities for the target class are used (see Figure 6.4 (d)).

In this thesis, StackingC is designed with the same five algorithms described above: SVM with RBF kernel, random forest, K^* instance-based learner, KNN with $K = 5$, and multi-layer perceptron. The model is built by training each classifier individually on the same training set. The multi-response linear regression classifier is trained on the output predictions of the base classifiers. Performance is then measured by presenting the model with examples from the test set. The base classifiers output predictions which are then used by the MLR classifier to determine whether each base classifier will predict correctly.

6.4 Stratified cross-validation

V-fold cross-validation is a process that helps estimate classifier performance when the size of the dataset is limited. With this scheme, the database is divided into v equally sized subsets. With one subset left aside, the remaining are used to train the model. The model is then tested on the subset which was left out. This is repeated until every subset has been left out. Figure 6.5 shows an example cross-validation where the dataset has been partitioned into four subsets.

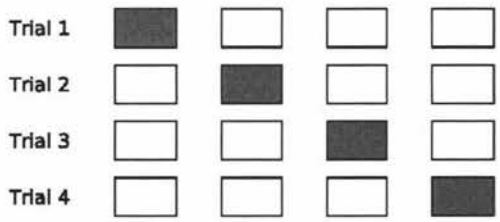


Figure 6.5: An illustration of a cross-validation example where the dataset has been partitioned into four sets. The dark rectangle represents the partition used as the test set, and the white rectangles represent the training sets (from (Haykin, 1999)).

Because of the substantial variance in an individual v-fold cross-validation, each classifier model is evaluated over ten stratified ten-fold cross-validation (v = 10) runs (Witten and Frank, 2000). Stratified cross-validation is the division of the dataset into randomly generated training and testing sets. For each stratified ten-fold cross-validation, the dataset is divided randomly into ten subsets. Nine subsets are used to train the model, while the remaining 10th subset is used to evaluate the classification error. This process is repeated ten times, with each iteration having a different test set held out of the training process. Stratification simply means that the training and test sets are sampled randomly from the dataset and aims to ensure that each class is uniformly represented in each training and test set. The ten-fold cross-validation is repeated ten times in order to reduce the statistical variance introduced by evaluating each classifier using only one ten-fold cross-validation. Ten-fold cross-validation is generally observed to yield a decent estimation for classification error (Witten and Frank, 2000). For example, classification accuracy of each learning algorithm is evaluated as described in Figure 6.6.

```
1  for i=1..10 do
2          partition data into 10 sets using random seed i
3          for j=1..10 do
4                  build classifier model
5                  train classifier model on training set
6                  test classification accuracy using test set
7          end
8          save result
9  end
10 save result
```

Figure 6.6: Psuedocode describing ten × ten-fold cross-validation.

6.5 Feature selection

Often there are irrelevant and redundant features that can reduce classification accuracy. Feature selection is a process which attempts to find a subset of the feature space which accounts for the majority of the variance between the target classes. Therefore, the feature space must be searched for a subset that does not contain these detrimental features. As the feature space is searched, features are ranked according to the performance they add to the set. The search algorithm and subsequent ranking method are important factors to be considered during this

process. For small sets of features, the best search algorithm is to simply use a brute force approach. Here, every possible subset of the original feature set is exhaustively tested. The subset which shows the highest classification accuracy is chosen.

While ideal for small feature spaces, the brute force approach is computationally prohibitive on datasets with a large number of features. For such sets there must be a traversal of subsets which are expected to yield better results. However, this can never be a substitute for an exhaustive search over the feature space, and therefore the assurance that the chosen subset is the optimal one is decreased. In this thesis, three feature selection methods were used: forward selection, principal components analysis, and a genetic search algorithm. These are described in detail below.

6.5.1 Principal components analysis

Principal components analysis (PCA) is a statistical method for performing data reduction on a given dataset. PCA re-aligns the coordinate system for a dataset such that the majority of the variance is accounted for by drawing an axis through the regression line. The resulting axis is called the first principal component. Iteratively, the remaining variance is accounted for by drawing subsequent axes through the remaining lines of regression, and as a result, each principal component is orthogonal to every others.

The first principal component is defined as:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} E\{(\mathbf{w}^T\mathbf{x})^2\} \tag{6.19}$$

where $\mathbf{x}$ is the dataset. Subsequent components can be calculated by:

$$\hat{\mathbf{x}}_{k-1} = \mathbf{x} - \sum_{i=1}^{k-1} \mathbf{w}_i\mathbf{w}_i^T\mathbf{x} \tag{6.20}$$

where $\mathbf{x}_{k-1}$ is the resulting dataset from the last iteration. It follows that the kth principal component is:

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} E\left\{\left(\mathbf{w}^T\hat{\mathbf{x}}_{k-1}\right)^2\right\} \tag{6.21}$$

The number of principal components retained depends on the amount of variance accounted for by each component.

6.5.2 Forward selection

Forward selection is a stepwise search algorithm that iteratively adds a new feature to an existing set based on its performance with the given set. Some type of classification algorithm must be

used at each stage as a subset evaluator to measure the performance of each subset. The process ends when there is no performance improvement and the resulting set comprises those features which perform well together; all irrelevant features are discarded.

However, the algorithm is greedy; the drawback being that each newly added feature may render existing features insignificant. Moreover, members which have been discarded previously may perform well with a later set. Figure 6.7 shows a simple forward selection algorithm.

```
1 for i=1..num_features do
2         for j=1..num_features do
3                 tenatively add feature i to set
4                 test set performance with classifier
5                 remove feature i from set
6         end
7         permanently add best feature to set
8 end
```

Figure 6.7: Psuedocode describing the forward selection algorithm.

6.5.3 Genetic search

Feature selection by genetic search has been popular in recent research (Dieterle, 2003; Emmanouilidis *et al.*, 1999; Vafaie and De Jong, 1992). A genetic search of the feature space mimicks biological evolution by "mutating" chromosomes (feature sets). Genes (individual features) make up the chromosomes which are initially randomly turned on or off (set to "0" = off or "1" = on).

Beginning with an initial population of randomly generated chromosomes, each chromosome is passed through a fitness function (for example, a classification model is generated and tested with the current chromosome) which ranks each member of the current generation. Those chromosomes with the greatest fitness are "selected" and mated, with a mutation probability that introduces or removes one or more genes. When a stopping criteria has been met, such as a maximum number of generations, the process stops and ideally an optimal feature set is produced. A data flow diagram of this process is show in Figure 6.8. A full description of genetic algorithms with examples can be found in Goldberg (1989).

In our experiments, the fitness function for determining the rank of each chromosome is the SVM with RBF kernel described in Section 6.2.1. The population size was set to 20 chromosomes and the stopping criteria was a maximum of 20 generations. The probability of a mutation occurring during mating was 0.033 and the crossover probability during mating was 0.6.

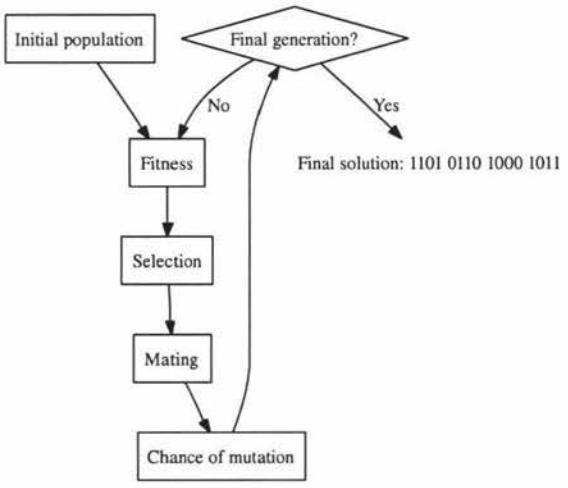


Figure 6.8: Data flow diagram describing the process of genetic search over a feature space (adapted from (Dieterle, 2003)).

6.6 Summary

In this chapter we presented an overview of the classification algorithms selected for vocal emotion recognition. Based on the performance with the NATURAL dataset, five top performers were chosen: a support vector machine utilising the radial basis function kernel, a multi-layer perceptron network with one hidden layer, a random forest of decision trees, the K^* instance-based learner, and a K-nearest neighbours instance-based learner with $K = 5$.

We then introduced two ensemble classification techniques: unweighted vote and stacked generalisation. Both techniques are built using the base classifiers described previously. The unweighted vote combines the predictions of each of the five classifiers into a vote which determines the most popular target class. Stacked generalisation uses a multi-response linear regression (MLR) model to learn the predictions of the five base classifiers associated with each target class and tries to predict when these classifiers are incorrect.

The machine learning technique of stratified cross-validation was used for all classification experiments and allows a more accurate representation of classifier predictions for smaller datasets.

Feature reduction methods were also employed to reduce the complexity of the learning process and the dimensionality of the feature sets. Redundant and irrelevant features are shown to cause a significant decrease in classification accuracy, and therefore it is important that steps are taken to identify and prune these features from the dataset.

In the next chapter, we compare the results between the base classifiers and the ensemble

methods and show results for the feature selection experiments. We also compare the classification rates between the NATURAL dataset and the ESMBS dataset. Finally, we introduce a prototype implementation of the system developed in this thesis.

Chapter 7

Experimental Results and Prototype Implementation

7.1 Introduction

The previous chapter presented traditional classification algorithms and feature selection techniques. We also introduced the ensemble classification methods StackingC and unweighted vote.

In this chapter we compare the results of the classification algorithms discussed in the previous chapter. We first give results on the base classifiers using the original NATURAL and ESMBS datasets. Next, we compare the performance of the ensemble classifiers. Experiments with feature selection are then presented and comparisons are shown between the base and ensemble classifiers.

After the presentation of the experimental results, this chapter then provides an overview of the development and implementation of a prototype emotion recognition system. Building on the theoretical foundations introduced in Chapter 2, the NATURAL dataset introduced in Chapter 3, and the feature set derived in Chapter 5 we develop a simple prototype which is able to extract emotional profiles from human speech.

7.2 Experimental results

In this section we compare the classification accuracies on the NATURAL and ESMBS datasets using the classification methods introduced in the previous chapter. The methodology for acquiring the results below is as follows. Using the full NATURAL and ESMBS datasets, we perform feature selection using the three methods described in Chapter 6. First, the principal components are calculated for both datasets. Next, the other two feature selection methods, for-

ward selection and genetic search, are employed using the SVM with RBF kernel as the subset evaluator. The choice is made to use the SVM with RBF kernel for the subset evaluator because it was the highest performer in the initial base classifier selection experiment and can be relied upon to describe the most relevant feature set. Next, we build the classifier ensembles and perform classification experiments on these using the original datasets as well as the subsets produced from feature selection.

Unless otherwise stated, all classification accuracies given are the result of 10×10-fold stratified cross-validations over the datasets with a significance of 0.05.

7.2.1 Performance of base classifiers

In Tables 7.1 through 7.3 we list the confusion matrices for the different base classifiers on the NATURAL and ESMBS datasets. It can be seen that for the NATURAL dataset, the classification accuracy for neutral is much higher than that of anger. This is due to the unbalanced data (155 anger/233 neutral) in this set.

For the ESMBS dataset, it is easily seen that anger and sadness are classified with high accuracy (generally > 90%), where other emotions such as happiness and fear have much lower accuracies (sometimes < 50%). This is inter-class confusion is also common in human listeners (Scherer, 2003; Nwe, 2003). Emotion classes which oppose each other such as anger and sadness are much more easily separated than those classes with similar characteristics such as happiness and surprise.

For both datasets, the SVM with RBF kernel shows the highest performance. The random forest is the second best for the ESMBS dataset, but is outperformed by the KNN on the NATURAL dataset.

These results show that the ESMBS dataset, while having 6 classes, is almost as accurately classified as the NATURAL dataset, which only has 2 classes. Randomly classifying the ESMBS dataset would show an average rate of about 16.67% whereas a randomly classifying the NATURAL dataset would show an average rate of 50%.

This highlights the difficulties involved in using data collected from natural environments. The emotion represented is subtle and highly varied due to the uncontrolled nature of the method. Even two class problems such as this show quite low classification accuracies. Conversely, the results from the ESMBS dataset show several important points. First, the classification accuracies are very similar to that from human listeners, as we saw in Chapter 3. Second, due to the high classification rates (which are much greater than chance, as mentioned above), we can see that the methods (features used, extraction methods, and classification algorithms) followed in this research are sound. Therefore, we can be confident in the results from the NATURAL dataset.

Table 7.1: Confusion matrices for the support vector machine with RBF kernel on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**67.94**	32.06
Neutral	17.08	**82.92**

Correctly classified (%): **76.93**

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**91.24**	0.00	0.00	1.86	0.00	6.90
Disgust	0.08	**67.36**	15.27	11.78	1.63	3.88
Fear	0.23	16.67	**59.69**	13.64	4.81	4.96
Happiness	1.71	17.67	16.12	**49.61**	1.55	13.33
Sadness	0.00	3.10	4.03	1.24	**91.63**	0.00
Surprise	7.21	3.02	5.27	12.95	0.00	**71.55**

Correctly classified (%): **71.85**

Table 7.2: Confusion matrices for the multi-layer perceptron on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**67.16**	32.84
Neutral	22.19	**77.81**

Correctly classified (%): **73.56**

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**90.54**	0.00	0.08	2.02	0.00	7.36
Disgust	0.00	**56.59**	20.31	14.88	2.48	5.74
Fear	0.31	19.30	**46.67**	16.12	8.45	9.15
Happiness	5.35	14.19	18.06	**42.40**	1.86	18.14
Sadness	0.00	3.41	4.81	1.63	**90.16**	0.00
Surprise	9.61	4.81	6.28	13.41	0.00	**65.89**

Correctly classified (%): **65.37**

Table 7.3: Confusion matrices for the K-nearest neighbour classifier (with $K = 5$) on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**64.26**	35.74
Neutral	16.44	**83.56**

Correctly classified (%): **75.85**

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**93.41**	0.00	0.39	1.40	0.00	4.81
Disgust	0.93	**60.23**	20.00	12.40	1.47	4.96
Fear	0.08	23.33	**53.18**	12.17	4.81	6.43
Happiness	7.29	27.44	18.76	**31.78**	2.25	12.48
Sadness	0.00	7.21	9.84	2.87	**79.92**	0.16
Surprise	20.39	9.15	8.29	9.69	0.00	**52.48**

Correctly classified (%): **61.83**

Table 7.4: Confusion matrices for the K^* instance-based learner on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**62.90**	37.10
Neutral	24.16	**75.84**
Correctly classified (%): **70.67**		

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**91.32**	0.00	0.78	0.85	0.00	7.05
Disgust	0.70	**57.75**	20.47	11.63	4.57	4.88
Fear	1.55	25.81	**45.12**	16.36	3.64	7.52
Happiness	5.43	18.91	22.71	**37.21**	1.01	14.73
Sadness	0.00	5.04	14.57	3.64	**76.12**	0.62
Surprise	12.64	4.65	9.22	18.99	0.00	**54.50**
Correctly classified (%): **60.34**						

Table 7.5: Confusion matrices for the random forest on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**66.58**	33.42
Neutral	22.62	**77.38**
Correctly classified (%): **73.07**		

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**94.88**	0.00	0.54	0.39	0.00	4.19
Disgust	0.08	**65.81**	19.38	11.63	0.62	2.48
Fear	1.40	23.88	**47.60**	11.71	7.13	8.29
Happiness	5.12	22.79	18.68	**36.51**	0.54	16.36
Sadness	0.00	4.50	3.57	0.78	**91.01**	0.16
Surprise	6.98	7.60	4.65	12.40	0.00	**68.37**
Correctly classified (%): **67.36**						

7.2.2 Performance of ensemble classifiers

Here we present the performance statistics for the StackingC and vote ensembles. These classifiers are built from the five classifiers previously listed. For stacked generalisation, the final prediction is based on a meta-classifier which is trained on the class probabilities and targets for each training example. Stacked generalisation attempts to predict when the base classifiers will be incorrect. The unweighted vote simply sums the predictions of each class from the base classifiers and picks the most popular class.

The confusion matrices for StackingC are presented in Table 7.6. It can be seen that StackingC offers an improvement over the base classifiers for both the NATURAL and ESMBS datasets. On the ESMBS dataset, anger and sadness are both highly accurate, while the accuracies for happiness and fear are much lower. Like the results on the base classifiers presented above, these predictions are in line with the accuracies for human classification.

Table 7.7 shows the results for the unweighted voting scheme. The vote shows a slight improvement over StackingC. Interestingly, for the NATURAL dataset, anger is predicted more accurately for the vote while neutral speech is predicted less accurately. For the ESMBS dataset, anger shows a higher rate than that of StackingC, but sadness shows a lower classification rate. Vote is significantly less accurate for surprise than StackingC, with 70.78% and 78.84% respectively.

Table 7.6: Confusion matrices for the StackingC classifier on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**66.52**	33.48
Neutral	15.28	**84.72**

Correctly classified (%): **77.45**

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**94.81**	0.00	0.00	0.70	0.00	4.50
Disgust	0.08	**67.05**	16.98	10.08	2.25	3.57
Fear	0.08	19.15	**52.95**	10.70	7.75	9.38
Happiness	4.50	15.66	15.19	**44.73**	1.32	18.60
Sadness	0.00	1.94	2.64	0.70	**94.73**	0.00
Surprise	5.43	3.88	3.33	8.53	0.00	**78.84**

Correctly classified (%): **72.18**

7.2.3 Performance after feature selection

As mentioned above, because the SVM with RBF kernel is the most accurate, it is used for feature selection where a subset evaluator is required. A subset evaluator is required for the forward selection and the genetic search, since there must be a way of measuring the performance of the newly generated dataset at each stage in the process. For principal components analysis, no subset evaluator is needed.

Table 7.7: Confusion matrices for the unweighted vote classifier on the NATURAL and ESMBS datasets.

(a) NATURAL dataset

	Anger	Neutral
Anger	**69.03**	30.97
Neutral	15.97	**84.03**
Correctly classified (%): **78.04**		

(b) ESMBS dataset

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**95.89**	0.00	0.08	0.00	0.00	4.03
Disgust	0.39	**65.35**	17.75	10.93	0.62	4.96
Fear	0.31	20.93	**53.02**	12.64	4.96	8.14
Happiness	4.34	16.51	17.13	**45.12**	1.55	15.35
Sadness	0.00	1.78	3.95	1.16	**93.10**	0.00
Surprise	9.46	4.65	4.26	10.85	0.00	**70.78**
Correctly classified (%): **70.54**						

Feature selection is performed on both the NATURAL and ESMBS databases independently, meaning the process yields different feature subsets for each database. Table 7.8 shows the resulting feature subsets for each database.

Principal components analysis yields exactly the same feature sets for each database. To aid in the labelling of PCA selected features, we transform the principal components back into the original feature space and kept only the top 25 principal components[1]. Every attribute from the pitch and first formant frequency (F1) contour is retained. The majority of attributes from the F2 contour are retained, with the exception of the value at the last voiced segment and the range. No attributes for F3 are retained, hinting that this entire feature group may not add any variance to the dataset. All energy attributes are kept, save the range and mean.

The feature sets resulting from forward selection do not seem to show much correlation between the NATURAL and ESMBS datasets. The F0 attributes are shared except that the value at the last voiced segment and the range are retained for the ESMBS dataset. The F1 features mean, minimum, maximum, and standard deviation are favoured for NATURAL, while maximum is discarded for ESMBS. F2 attributes compare similarly for each dataset when compared with F1 attributes, except the value at the last voiced segment for F1/NATURAL is retained and the F2 standard deviation is discarded for ESMBS. F3 attributes are very different between datasets. They are sparsely retained for NATURAL but densely retained for ESMBS. This may be due to the subtlety of emotion in the NATURAL dataset and the clear, concise nature of emotion in ESMBS due to the different data collection methods described in Chapter 3. Mean energy is valued for both datasets, while the values at the first and last voiced segments seem important only for ESMBS. Forward selection on both datasets retain all of the rhythm-based statistics. Of interest is the fact that forward selection for the NATURAL dataset resulted in 23 features retained, where for the ESMBS dataset, 30 features are retained.

For the search using the genetic algorithm, features for F0, F1, and F2 are almost identical

[1]For classification, the principal components are not transformed back into the original feature space.

with a high number of attributes retained. For F1, the maximum is discarded for ESMBS and for F2, the value at the first voiced segment is discarded for NATURAL but retained for ESMBS. It is the opposite case for the value at the last voiced segment: retained in NATURAL but discarded in ESMBS. The F3 attributes seem quite useful for both datasets in comparison to the other selection techniques, but more so for NATURAL. All energy features are kept for NATURAL, whereas all except the minimum and value at the first voiced segment for ESMBS. All features relating to rhythm are retained. Interestingly, the genetic search retains the highest number of features for both datasets, compared to forward selection.

Table 7.9 shows a summary of results with feature selection on both the NATURAL and ESMBS data sets. All results presented are the average number of correctly classified instances over 10×10 cross-validation.

Forward selection proves to be the most accurate feature selection method for the NATURAL dataset, proving more accurate for every classifier except K^*, where the genetic search yields a better dataset. The genetic search proves more accurate than forward selection and PCA on the ESMBS dataset. PCA is always worse than even the original feature set, which is surprising, as PCA often has success for accurate feature reduction in other studies (Lee and Narayanan, 2005).

Observing the results more closely, we can see that forward selection on ESMBS actually worsens classification. This is likely due to the fact that the subset evaluation for the forward selection search is done using the SVM. Had each classifier evaluated its own intermediate subsets during selection, the resulting feature sets would likely have been better adapted to those classifiers.

With respect to the performance of the classifiers, there is a clear improvement with using the ensemble methods. For the NATURAL dataset, we can see that the voting scheme and StackingC perform slightly better than the base classifiers using the original feature set, and marginally better than the SVM using the forward selection set. The improvement is more significant when we look at the results for the genetic search.

For the ESMBS dataset, the ensemble methods perform significantly better on the original datasets compared to the base classifiers. Here, however, it is StackingC which shows the best performance by almost 2% over the voting scheme.

7.2.4 Summary of results

In summary, forward selection and the voting scheme performed best on the NATURAL dataset, while the genetic search and StackingC performed best on the ESMBS dataset. In general, the accuracies of between the different emotion classes remained constant over all classification methods.

In the ESMBS dataset, anger and sadness were most accurately classified, followed by sur-

Table 7.8: Resulting feature subsets after feature selection. PCA = principal components analysis; FW = forward selection; GA = genetic algorithm. PCA datasets have been transformed back into the original feature space for labelling purposes and have the top 25 principal components retained.

		Dataset					
		NATURAL			ESMBS		
Number	Description	PCA	FW	GA	PCA	FW	GA
1	F0 mean	•	•	•	•	•	•
2	F0 minimum	•	•	•	•	•	•
3	F0 maximum	•	•	•	•	•	•
4	F0 standard deviation	•	•	•	•	•	•
5	first F0 value	•		•	•		•
6	last F0 value	•		•	•	•	•
7	F0 range	•		•	•	•	•
8	F1 mean	•	•	•	•	•	•
9	F1 minimum	•	•		•	•	
10	F1 maximum	•	•	•	•		
11	F1 standard deviation	•	•	•	•	•	•
12	first F1 value	•		•	•	•	•
13	last F1 value	•	•	•	•	•	•
14	F1 range	•		•	•	•	•
15	F2 mean	•	•	•	•	•	•
16	F2 minimum	•	•	•	•		•
17	F2 maximum	•	•	•	•	•	•
18	F2 standard deviation	•	•	•	•		•
19	first F2 value	•			•	•	•
20	last F2 value			•		•	
21	F2 range			•		•	•
22	F3 mean			•		•	
23	F3 minimum		•	•		•	•
24	F3 maximum		•	•		•	
25	F3 standard deviation			•		•	•
26	first F3 value					•	•
27	last F3 value		•	•		•	•
28	F3 range			•			•
29	energy mean		•	•		•	•
30	energy minimum	•	•	•	•		
31	energy maximum	•	•	•	•	•	•
32	energy standard deviation	•		•	•		•
33	first energy value	•		•	•	•	
34	last energy value	•		•	•	•	•
35	energy range		•	•			•
36	speaking rate		•	•		•	•
37	average length of unvoiced segments	•	•	•	•	•	•
38	average length of voiced segments		•			•	•
	Count	25	23	34	25	30	31

Table 7.9: Average percentages of correctly classified instances from the NATURAL and ESMBS datasets for all classification methods. For acronyms in the dataset column, ORIG = original feature set; PCA = principal components analysis; FW = forward selection; GA = genetic algorithm.

	Dataset							
	NATURAL				ESMBS			
	ORIG	PCA	FW	GA	ORIG	PCA	FW	GA
SVM (RBF)	76.93	75.98	79.20	75.95	71.85	63.94	70.72	72.05
MLP	74.25	72.06	75.15	73.99	65.37	61.12	66.71	66.86
KNN (K = 5)	75.85	73.51	76.75	73.43	61.83	52.36	57.29	61.64
K^*	70.67	66.55	71.19	71.68	60.34	44.32	58.13	61.43
RF	73.07	66.73	73.99	72.47	67.36	53.85	66.77	69.04
StackingC	77.45	75.49	79.28	77.73	72.18	63.59	72.44	73.29
Vote	78.04	75.57	79.43	77.83	70.54	59.97	69.38	72.30

prise, disgust, fear, and happiness. For the NATURAL dataset, neutral speech was always classified more accurately than angry speech. The likely reason for this is that the dataset was slightly unbalanced (60% neutral versus 40% angry). Overall, classification rates on the NATURAL dataset were lower than expected. The main problem in using spontaneous emotional speech is the lack of control the researcher has over the experiment. Lack of control leads to unbalanced data often having some background noise and indefinite levels of specific emotions.

However, because the ESMBS dataset showed success comparable to human listeners, we can be confident that the features outlined in Chapter 5 for describing the variation between emotional classes are very good.

The results also show us the inherent differences between the two data collection methods. Acted data may lead to inflated results, while data from real-world situations yields much lower classification rates, but paints a more realistic picture of applied automatic emotion recognition.

7.3 Prototype implementation

Following the development of the algorithms for feature extraction and the design of a suitable artificial neural network architecture, a simple prototype application was designed and implemented. In this section, we describe the program architecture and design issues and implications of the prototype implementation.

The application was developed in the C and C++ programming languages. The main program was written in C++ and the modules which facilitated feature extraction were linked as C objects. The classification module was written as a shared library which could be replaced when needed. The network was generated using the Stuttgart Neural Network Simulator (SNNS)[2]

[2]The Stuttgart Neural Network Simulator can be downloaded from `http://www-ra.informatik.`

which allows a trained network to be compiled as C code.

Figure 7.1 shows the data flow diagram for the implementation. The speech signal is read by way of microphone input or file. The endpoints of utterances in the signal are detected, and this utterance is cropped out and processed by the feature extraction module. Statistics are calculated based on the various frequency contours and are compiled into a feature vector which is passed through the classification module where a prediction of either angry or neutral is returned.

Figure 7.1: Data flow diagram for prototype system

7.3.1 Endpoint detection

During real-time processing, it is important for the system to be able to detect the endpoints of an utterance so that an assessment can be constructed. For an utterance the endpoints are defined as the beginning and the end. The endpoints can be detected by measuring the short-time energy on a frame-by-frame basis. When several frames have surpassed a threshold value for silence, signifying the beginning of an utterance, these frames can be buffered until the short-time energy falls below that threshold, signifying the end of the utterance. Figure 7.2 shows an example waveform with endpoints highlighted. The dark grey regions indicate the silence preceding and following the utterance. The beginning and ending crop points are denoted by the arrows.

Endpoint detection is performed automatically in the prototype. The prototype begins by sampling the ambient noise. This takes roughly 400 ms to gather enough frames, after which the mean energy is calculated and this value is used as a threshold. Once the energy surpasses this threshold, recording begins and does not finish until the energy drops below the threshold for a certain number of frames. After the energy drops below the threshold, the utterance is sent to the feature extraction module while the recording process is reset to the initial state.

7.3.2 Feature extraction

Feature extraction for the prototype implementation follows the procedure described in Chapter 5. With the endpoints of the speech sample detected, the utterance is saved in memory in PCM WAVE format. Next, the F0 contour is extracted by way of the Robust Algorithm for Pitch Tracking (RAPT) (see Chapter 5 for details) and the first three formant frequencies (F1,

uni-tuebingen.de/SNNS/.

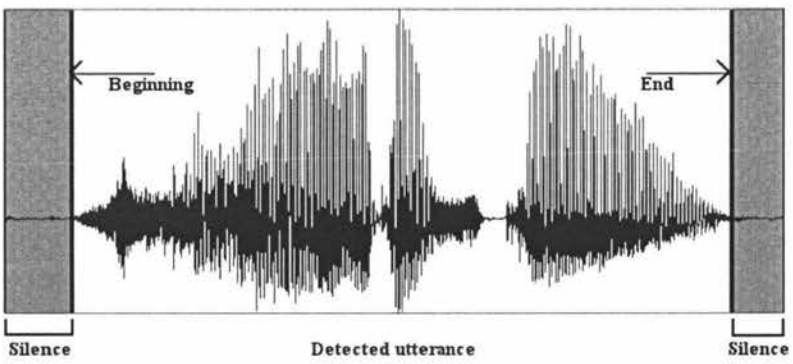


Figure 7.2: A sample utterance with endpoints highlighted. The dark grey regions indicate the silence preceding and following the utterance.

F2, F3) are extracted using LPC analysis and dynamic programming. The energy contour is then extracted, and the following statistics are calculated across these five contours: minimum, maximum, mean, standard deviation, value at first voiced segment, value at last voiced segment, and the range.

From the F0 contour, three additional features are calculated. The first is the speaking rate. This is calculated by estimating the number of transitions between voiced and unvoiced and vice versa. As mentioned in Chapter 5, a segment (one or more frames) is considered voiced if there is a valid value for the fundamental frequency and unvoiced if there is no valid value. These transitions are summed and divided by the length of the utterance (in seconds).

Next, the average duration of unvoiced segments (representing speaker pause) and voiced segments are calculated by summing the lengths of these segments and dividing by the number of transitions.

With all features extracted, the next step is to formulate a feature vector. A 38-element array is constructed and each group of statistics is copied into this new memory.

7.3.3 Real-time processing

As far as real-time processing is concerned, feature extraction is the most computationally expensive step. Computation times for detecting endpoints, cropping the utterance, and classifying the feature vector are all negligible. The generation of the fundamental and formant frequency contours require the most CPU time and this is the bottle-neck of performance in the prototype.

If feature extraction introduces too large a lag in the response time for the system, it may be necessary to sacrifice some classification accuracy (by reducing the number of features calculated) in exchange for increased speed. The feature extraction times for the algorithms described

in Chapter 5 are presented in Table 7.10. These were measured on a Pentium 4 1.4 GHz CPU. The extraction times constitute only a fraction of the utterance duration. From the example, we can see that an assessment can be made available in 283.30 ms for an utterance slightly over 3 seconds in length. After feature extraction is complete, the feature vector needs only to be passed through the classification model. The time needed to classify the utterance is negligible and sufficiently fast, given no modification to the model is required.

Breazeal and Aryananda (2002) take real-time processing into consideration. In their system, the robot was designed to respond within comfortable communcation constraints. They note that a human can tolerate small delays (one second or so) but that "long delays will break the flow of the interaction ... [and will] also interfere with the caregiver's ability to use the vocalization as a reinforcement signal." They strive for minimal latencies (< 500 ms), but later note that sometimes up to one or two seconds may pass before a response is given by the robot. In total it is stated that there is a delay of 500 ms from end of speech to classification output and a 1-2 second delay "associated with interpreting the classifier in affective terms and feeding it through emotional response."

Table 7.10: Average times for feature extraction compared with the average length of an utterance in the database. The statistics calculations include the maximum, minimum, mean, standard deviation, range (for pitch, energy, formants) and speaking rate.

Feature group	Average time (ms)
Fundamental frequency (F0) contour	83.81
Formant frequency analysis	92.73
Short-time energy contour	30.49
Statistics calculations	76.27
Total	283.30
Utterance	3254.20

7.3.4 Classification

The classification module in the prototype implementation consists of an artificial neural network compiled as C source code. The network is designed and trained using the Stuttgart Neural Network Simulator. A utility exists in this software package that allows the compilation of the generated network into C source code.

Figure 7.3 shows the ANN architecture used in the prototype implementation. Of note is the number of nodes in the hidden layer. Initially, the prototype was developed with 77 hidden nodes based on initial experiments using SNNS. However, it was later discovered that this resulted in overfitting, so the number was reduced to 60 nodes in the hidden layer as per the results from the hidden layer experiment in Chapter 6.

The network is built as a shared library to allow simpler retraining and replacement of this

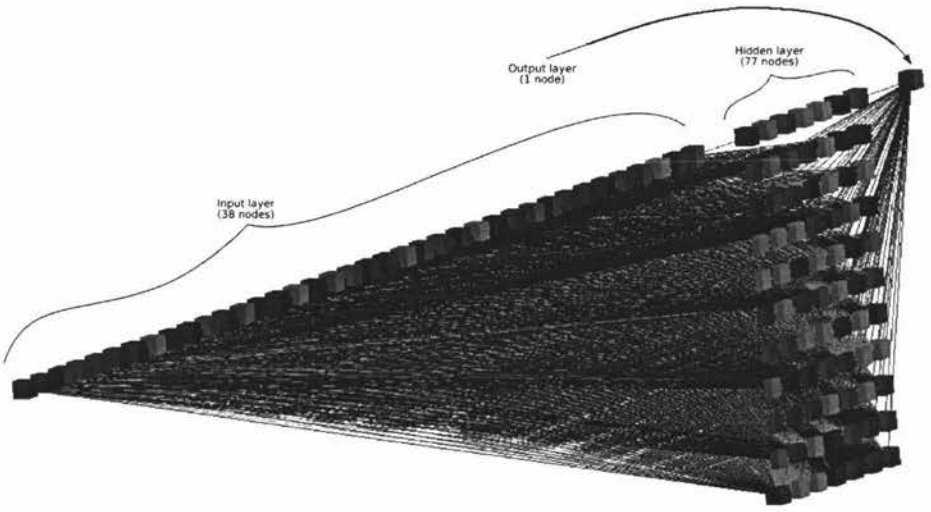


Figure 7.3: Graphical representation of the ANN architecture used in the prototype implementation.

component, separate from the rest of the prototype. An entry point is defined that allows a classification to be performed by simply loading the shared library, acquiring the address of the classification function, and calling this function with the feature vector as a parameter (see Figure 7.4). The result is returned in a new array and can be quickly examined for the predicted class.

Figure 7.5 shows a screen capture of the prototype implementation during a test. The prototype is built as a console application, meaning it has no graphical user interface. This choice was made to simplify the implementation by concentrating on the algorithms involved. However, the application can be easily extended to have a graphical interface to be more user-friendly.

7.3.5 Summary of prototype development

This section presented an overview of the development and implementation of the prototype for automatic and real-time spoken emotion classification. The prototype reads signal input from microphone or PCM WAVE file. Endpoints of speech are determined and feature extraction as detailed in Chapter 5 is performed. The feature vector is passed through a modular artificial neural network component which returns the predicted class.

```
1  int classify(float *in, float *out)
2  {
3          int retval = 0;
4
5  #ifdef WIN32
6          HMODULE mod = LoadLibrary("nn.dll");
7          nnProc_type nn;
8  #else
9          void* mod = dlopen("./libnn.so", RTLD_LAZY);
10         int (*nn)(float *,float *, int); // function declaration
11 #endif
12         if(mod)
13         {
14                 // Locate the procedure
15 #ifdef WIN32
16                 nn = (nnProc_type)GetProcAddress(mod, "nn");
17 #else
18                 *(void**)(&nn) = dlsym(mod, "nn");
19 #endif
20                 if(nn)
21                 {
22                         // Push features through neural network
23 #ifdef WIN32
24                         (nn)(in, out, 0);
25 #else
26                         (*nn)(in, out, 0);
27 #endif
28                 }
29                 else
30                 {
31                         retval = -1;
32                 }
33
34 #ifdef WIN32
35                 FreeLibrary(mod);
36 #else
37                 dlclose(mod);
38 #endif
39         }
40         else
41         {
42                 fprintf(stderr, "could not load neural network dll\n");
43                 retval = -1;
44         }
45
46         return retval;
47 }
```

Figure 7.4: C++ source code function for the dynamic loading of the ANN module for classification. The module is loaded (lines 6 and 9), the address of the classification procedure is then located (lines 16 and 18), the procedure is invoked (lines 24 and 26), and finally the module is unloaded (lines 35 and 37). The feature vector corresponding to the input layer is contained in the variable `in` and the prediction corresponding to the output layer is contained in the variable `out`.

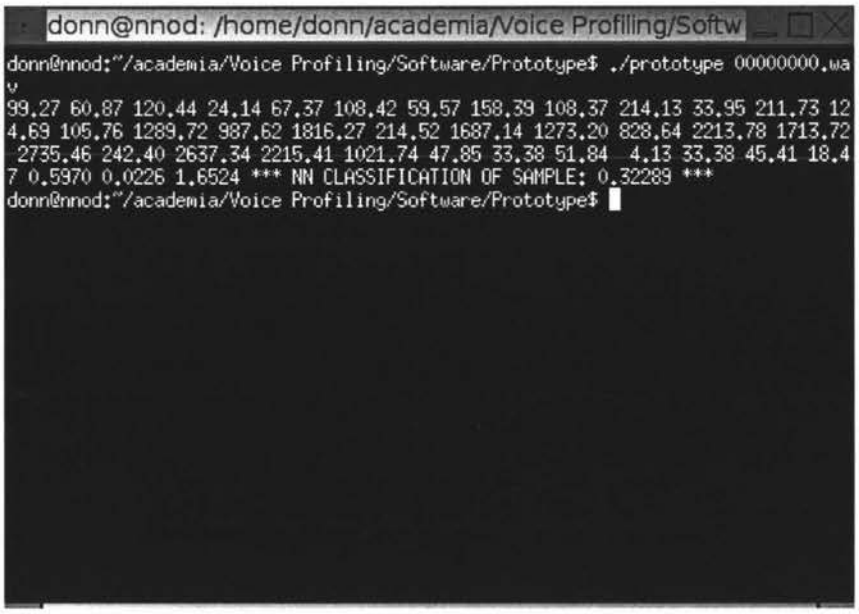


Figure 7.5: Screen capture of the prototype implementation.

7.4 Summary

This chapter showed the results of the ensemble methods compared to the results of the base classification algorithms. In all cases, the ensemble methods proved more effective at classification. Stacked generalisation performed the best for the ESMBS dataset while the voting scheme showed the highest accuracy for the NATURAL dataset. The results of feature selection were also discussed. It was seen that the genetic search yielded the most discriminative features on the ESMBS dataset while forward selection proved more useful on the NATURAL dataset.

Significant differences in classification accuracies were noted between the NATURAL and ESMBS datasets. The ESMBS dataset yielded a classification rate of 73.29% for 6 emotion classes while the NATURAL dataset yielded a rate of 79.43% for only 2 classes. Clearly the ESMBS database has clearer definition between the emotion classes than the NATURAL database, an issue which is attributed to the different methods of data acquisition.

We also detailed the development and implementation of the prototype system. The prototype was designed to automatically classify spoken emotion in real time. We presented the methods of endpoint detection, real-time feature extraction, and classification.

Chapter 8

Conclusion and Future Work

8.1 Summary of main findings

This thesis designed a system for speaker-independent, automatic and real-time emotion recognition from speech. Theoretical implications of emotion research were explored, and it was determined that a discrete emotion representation of the six basic emotions (anger, fear, disgust, sadness, happiness, and surprise, with neutral as a baseline state) is a preferred theoretical model. There are two reasons for this: first, the majority of previous research in this area covers a discrete representation and second, the alternative, a dimensional approach, would only require mappings to be drawn into the discrete emotion space in order for those emotions to be labelled.

Next, we reviewed the three main data acquisition methods: natural, simulated, and induced emotion. From these, two methods, natural and simulated acquisition, were used to gather emotional speech data into two datasets. The natural dataset (labelled NATURAL) comprised two emotion classes, anger and neutral, and was collected from a call-centre. The simulated dataset (labelled ESMBS), was collected from native Burmese and Mandarin speakers and comprised the six basic emotion classes: anger, happiness, sadness, fear, disgust, surprise. The NATURAL dataset required objective validation by listener-judges. For this purpose, a web-based human classification system was developed which can be found in Appendix A. The development and use of this system was crucial to obtaining objective labels for the NATURAL database.

We looked at some standard features correlating emotional expression and speech. We also extensively made use of the formant frequency contours as emotional markers. In contrast to past research which seems to suggest little or no emotional content is contained in the formants, we have found that by the use of feature selection methods, statistics based on the formant frequency contours are highly relevant. Feature extraction algorithms were introduced in Chapter 5. By the use of these methods we were able to extract the fundamental, formant, and energy contours from speech signals and use these for the generation of some simple statistics which

were subsequently compiled into feature vectors.

For classification, we compared several machine learning techniques and found that a support vector machine with the radial basis function kernel performed best on both datasets. The ensemble classifiers, stacked generalisation and unweighted vote, show an even further performance increase, especially when combined with forward selection and genetic search on the NATURAL and ESMBS datasets respectively. The voting scheme showed best performance on the NATURAL dataset while stacked generalisation performed best on the ESMBS dataset.

The classification accuracies were very similar to that from human listeners, as we saw in Chapter 3. The accuracy for human listeners for six emotions is typically 55% to 70%. For the six-class ESMBS dataset, StackingC showed an average accuracy of 73.29%. Due to the high classification rate on this dataset, we can conclude that the features chosen to represent emotional speech are sufficient for describing this variance.

Additionally, we uncovered some fundamental differences with respect to the nature of the two emotional speech databases. Natural emotional speech was seen to carry high in-class variance and faint markers. Conversely, simulated emotional speech is much clearer even over multiple classes. Overall, the results of the research indicate that emotion collected from natural sources contains subtle emotion which make it difficult for automatic classification. Acted or portrayed data show high correlations to the different emotion classes, and hence automatic classification is much easier, even for several classes.

The implications of the different data acquisition methods should always be explored. On one hand, the collection of acted data may artificially inflate the results, as the emotion expressed is often emphasised by the actors which leads to more obvious features. On the other hand, spontaneous data is collected from real-world situations, where humans are acting naturally, and hence emotion represented in the resulting data will be difficult to control, recognise, and categorise. This may lead to problems in the research: low classification accuracy, incorrect perceptions of actual emotion, as well as ethical issues.

Finally, Chapter 7 saw the design and implementation of a prototype application. This system will provide the basis for further study and will be introduced into a call-centre environment for testing. This prototype at present uses an artificial neural network for classification, but subsequent versions will see the introduction of more sophisticated machine learning algorithms.

8.2 Contributions

The contribution from this research to the field is two-fold. First, different data collection methods have long required a closer look. We have presented an examination of the differences between natural emotional expression and simulated or portrayed emotional expression in human speech.

Second, stacked generalisation and unweighted vote for classification have not been previously used in emotion recognition from speech despite a recent surge of research on these methods (Ting and Witten, 1999; Seewald, 2002b,a,c). In this research they have shown a significant improvement over traditional classification techniques. Despite the increased complexity due to training several algorithms, stacked generalisation and unweighted vote are a useful extension to traditional classification methods.

8.3 Future work

The avenues for future work in spoken emotion recognition are many, however, we have discovered a small number which would have a profound effect on the research presented in this thesis. First, to better adapt to the wide variety of speaking styles, emotion classes, and noise conditions in call-centres, an incremental learning system would be ideal. Incremental learning would allow the recognition models to be updated in a fast and efficient manner, typically without the need for retraining on large amounts of data. It would be useful to explore how incremental learning could be applied to stacked generalisation or the voting scheme. This would most likely require some degree of incremental learning on each of the base classifiers, and may be quite complex.

Second, it would be useful to collect a more even distribution of natural emotional speech data. This, as discussed in Chapter 3, is not an easy task and would require a significant investment of time and resources. However, to more accurately compare the two data acquisition methods this should certainly be a next step. Additionally, it would be useful to acquire data using induction techniques. Having data from all three methods of acquisition would allow a more complete picture to be made of the inherent differences.

Third, although ensemble methods were found to be optimal classifiers, the prototype uses an artificial neural network for classification. We would propose the integration of these ensemble methods with the prototype to determine the effectiveness in real-world situations.

Ubiquitous emotional intelligence in machines is still many years away. Although there is a need and most definitely economic incentive (as seen in Chapter 1), there are areas that still require additional research. The recognition of emotion from speech remains a hurdle because of the wide variety of speaking styles as well as cultural differences and display rules.

Appendix A

Emotional Speech Database Annotation System (ESDAS)

A.1 Introduction

An interactive web-based tool was developed to allow annotation of the speech database. The purpose of this tool is to create a ground truth for the emotional content contained in the dataset collected in the natural environment (NATURAL dataset). In Petrushin (2000) a similar system was developed to allow human participants to judge the emotional content of speech samples. Samples are presented to the user in a web browser in a random order. The user is asked to classify each sample in to one of the discrete emotion categories (anger, frustration, neutral, or unknown).

A.2 How it works

Users are presented with a logon screen. A user must sign up for an account to create a profile in the database. This profile stores a relation to each sample annotation. After creating an account, the user is presented with a welcome screen which gives a brief overview of the system and the requirements for use.

The user can select to begin the annotation process by clicking "Start". Upon starting, the first sample is presented. If the browser is configured correctly, the sample will automatically play. When the user has listened to the sample (it can be replayed if necessary), one of four buttons is clicked, corresponding to the perceived emotion in the sample. These choices are anger, frustration, neutral, unknown. After one of these buttons is pressed, the user is taken to the next sample, which is automatically played. This continues until either the maximum samples per session has been reached, or all samples have been classified.

Upon reaching the maximum samples per session, the user is reminded that the process can be resumed the following day. If all samples have been classified, the user is asked to reclassify any Unknown classifications into one of the other three classes.

A.3 Screenshot

A screenshot for ESDAS is shown in Figure A.1. Simplicity and efficiency were goals for the design of the interface.

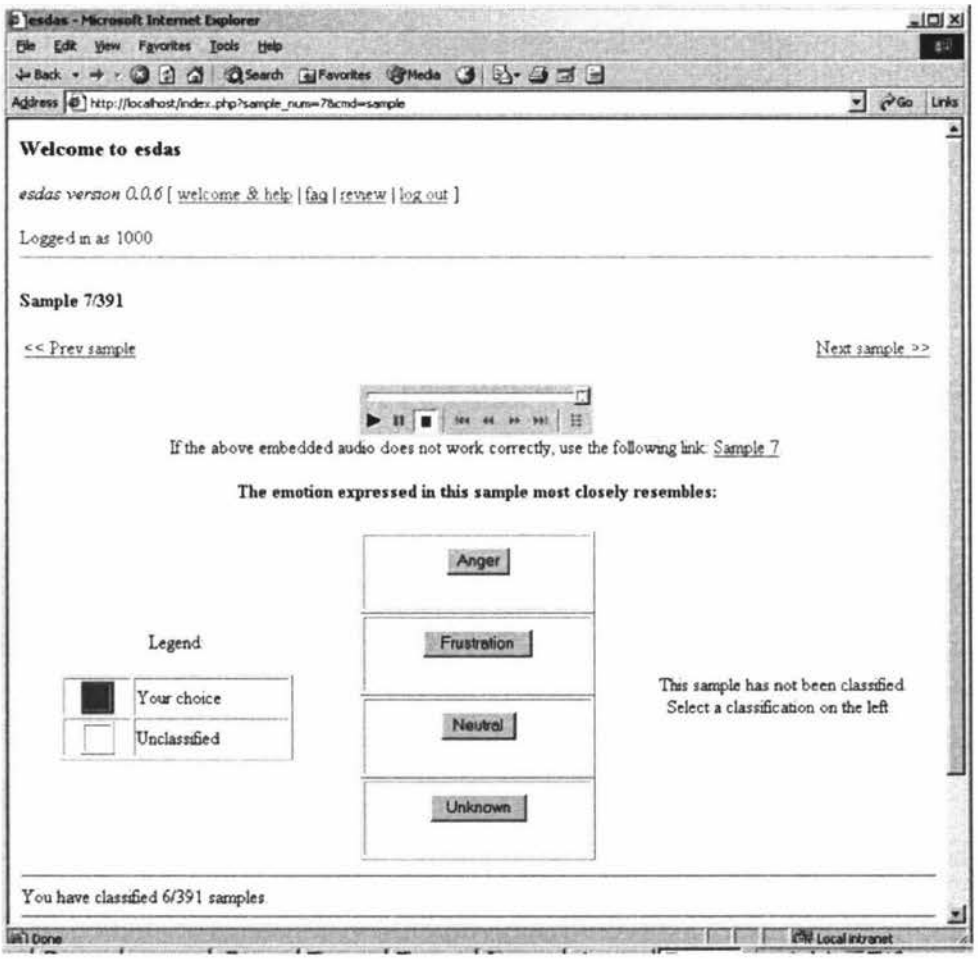


Figure A.1: ESDAS interface

A.4 Use studies

While 14 users began the annotation, only nine finished it entirely. There are several reasons for this. The first is that it is a fairly uninteresting process and there is little incentive for completion. The second is that because initial versions of the system limited the number of classifications per session to help eliminate speaker-recall, the users often forgot to return to resume annotation. Speaker-recall occurs because the user has recognised the speaker from previous samples, and is more inclined to maintain previous classifications because of this. Therefore, the process was broken up over several days to alleviate this effect. However, the users were less likely to remember to complete the process, and some ended up forgetting altogether. It was then decided to remove this restriction in order to allow users to complete the process in one sitting. This improved the completion rate dramatically.

A.5 Conclusions

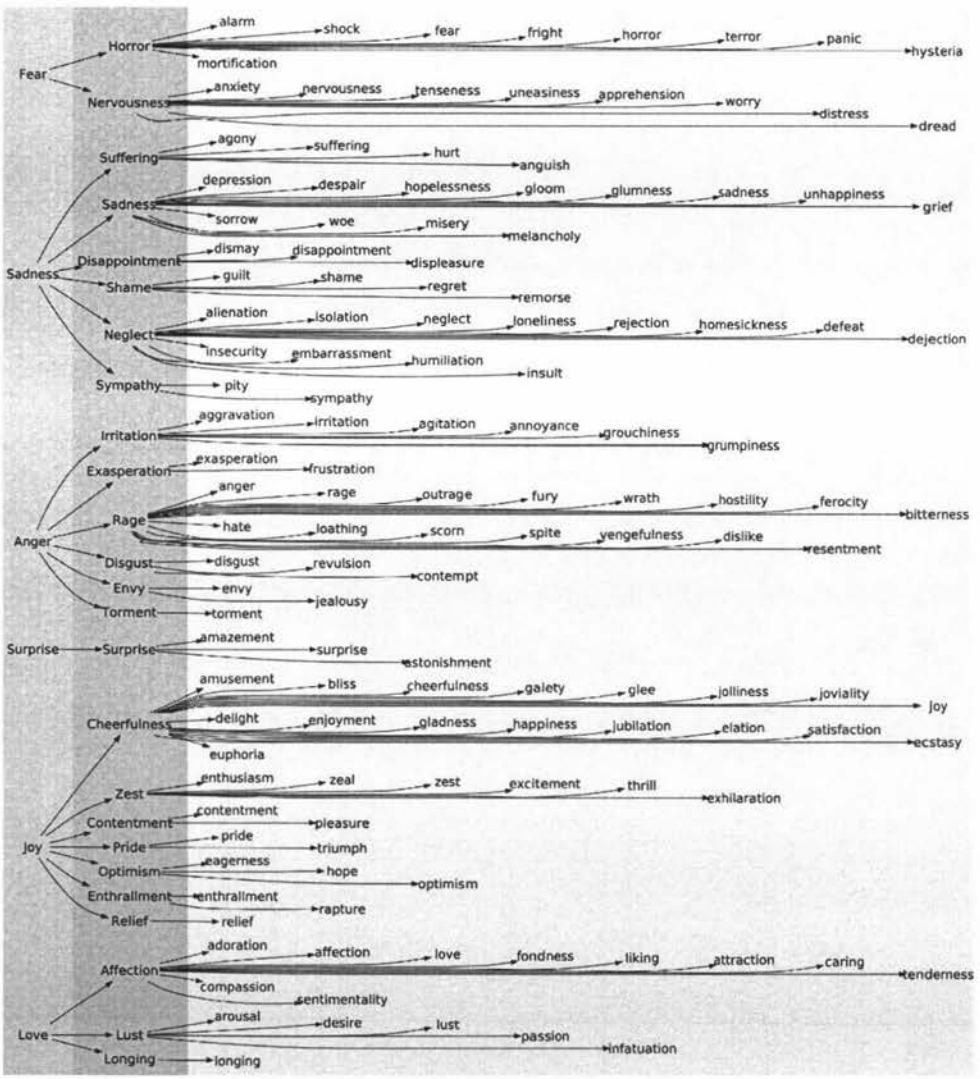
In conclusion, the system proved reliable in gathering a ground truth for the emotional content of the speech samples in the NATURAL dataset. The prcoess could be improved in several ways. Some recommendations were offered by the users to make the system both more objective and more closely model emotion classes.

- Allow for a scale of emotional intensity rather than the discrete categories anger, frustration, and neutral.
- Break the process up into shorter sessions. This would alleviate boredom as well as speaker-recall.
- Have listeners who cannot understand the language.
- Send a reminder by email until user completes or resigns from the process.

Appendix B

Other Figures

Figure B.1: The relationships between primary, secondary, and tertiary emotions (after (Parrot, 2001))



Author's Publications

[1] D. Morrison, R. Wang, L.C. De Silva, W.L. Xu. Real-time Spoken Affect Classification and its Application in Call-Centres. In *Proceedings of the International Conference on Information Technology and Applications (ICITA)*. Pages 583-586. Sydney, Australia, July 3-7, 2005.

[2] D. Morrison, R. Wang, L.C. De Silva. Spoken Affect Classification using Neural Networks. In *Proceedings of the IEEE Conference on Granular Computing*. Pages 483-486. Beijing, China, July 25-27, 2005.

[3] D. Morrison, R. Wang, W.L. Xu, L.C. De Silva. Incremental Learning for Spoken Affect Classification and its Application in Call-Centres. *Special Issue of the International Journal of Intelligent Systems Technologies and Applications (IJISTA)*. (accepted).

[4] D. Morrison, R. Wang, L.C. De Silva, W.L. Xu. Voting ensembles for Spoken Affect Classification. *Special Issue of the Journal of Networks and Computer Applications (JNCA)*. (accepted).

Bibliography

D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, 2002.

J. R. Averill. A constructivist view of emotion. In R. Plutchik and H. Kellerman, editors, *Emotion: Theory, research and experience*, volume 1, pages 305–339. Academic Press, New York, 1980.

J.-A. Bachorowski. Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8:53–57, 1999.

A. Batliner, R. Huber, H. Niemann, E. Noth, J. Spilker, , and K. Fischer. The recognition of emotion. In W. Wahlster, editor, *Verbmobil: Foundations of speech-to-speech translations*, pages 122–130. Springer, New York, Berlin, 2000.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. How to find trouble in communication. *Speech Communication*, 40:117–143, 2003.

C. Breazeal and L. Aryananda. Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12:83–104, 2002.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

J. G. Cleary and L. E. Trigg. K^*: An instance-based learner using and entropic distance measure. In *ICML*, pages 108–114, 1995.

R. R. Cornelius. *The Science of Emotion: Research and Tradition in the Psychology of Emotion*. Prentice Hall, Upper Saddle River, New Jersey, 1996.

T. T. Cover and P. E. Hart. Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

M. Cowan. Pitch and intensity characteristics of stage speech. *Arch. Speech, Supplement to December Issue*, 1936.

A. R. Damasio. *Descartes' Error: Emotion, reason, and the human brain*. G.P. Putnam's Sons, New York, 1994.

A. R. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. Ponto, J. Parvizi, and R. D. Hichwa. Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience*, 3:1049–1056, 2000.

F. Danes. Universality versus culture-specificity of emotion. In E. Weigand, editor, *Emotion in Dialogic Interaction: Advances in the Complex*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2004.

C. Darwin. *The expression of emotions in man and animals*. University of Chicago Press, Chicago, 1872/1965.

J. R. Davitz. Personality, perceptual, and cognitive correlates of emotional sensitivity. In J. R. Davitz, editor, *The Communication of Emotional Meaning*. McGraw-Hill, New York, 1964.

J. Deigh. Primitive emotions. In R. C. Solomon, editor, *Thinking about Feeling: Contemporary Philosophers on Emotions*. Oxford University Press, New York, New York, 2004.

F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, pages 1970–1973, Philadelphia, PA, 1996.

L. Devillers, I. Vasilescu, and L. Lamel. Annotation and detection of emotion in a task-oriented human-human dialog corpus. In *Proceedings of ISLE Workshop on dialogue tagging*, Edinburgh, 2002.

F. Dieterle. *Multianalyte Quantifications by Means of Integration of Artificial Neural Networks, Genetic Algorithms and Chemometrics for Time-Resolved Analytical Data*. PhD thesis, 2003.

T. G. Dietterich. Ensemble learning. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 405–408. The MIT Press, Cambridge, Massachusetts, 2002.

P. Ekman. *Darwin and Facial Expression*. Academic Press, New York, 1973.

P. Ekman. Are there basic emotions? *Psychological Review*, 99:550–553, 1992.

P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotics*, 1:49–98, 1969.

H. A. Elfenbein and N. Ambady. On the universality and cultural specificty of emotion recognition: a meta-analysis. *Psychological Bulletin*, 128:203–235, 2002.

C. Emmanouilidis, A. Hunter, J. MacIntyre, and C. Cox. Multiple-criteria genetic algorithms for feature selection in neurofuzzy modeling. In *Proceedings of the International Joint Conference on Neural Networks*, pages 4387–92, Washington, USA, 1999.

G. Fairbanks and L. W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monograph*, 8:85–91, 1941.

G. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monograph*, 6:87–104, 1939.

J. Flanagan, C. Coker, L. Rabiner, R. Schafer, and N. Umeda. Synthetic voices for computers. *IEEE Spectrum*, 7:22–45, October 1970.

J. L. Flanagan. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, New York, 1972.

I. Fonagy. A new method of investigating the pereption of prosodic features. *Language and Speech*, 21:34–49, 1978.

I. Fonagy. Emotions, voice and music. In J. Sundberg, editor, *Research Aspects on Singing*, pages 51–79. Royal Swedish Academy of Music No. 33, 1981.

I. Fonagy and K. Magdics. Emotional patterns in intonation and music. *Kommunikationsforsch*, 16:293–326, 1963.

N. Fragopanagos and J. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, Special Issue:1–17, 2005.

R. W. Frick. The prosodic expression of anger: Differentiating thread and frustration. *Aggressive Behaviour*, 12:121–128, 1986.

H. S. Friedman and T. Miller-Herringer. Nonverbal display of emotion in public and in private: Self-monitoring, personality, and expressive cues. *Journal of Personality and Social Psychology*, 61:766–775, 1991.

D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, Reading, Mass., 1989.

S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Upper Saddle River, New Jersey, 1999.

R. Huber, E. Noth, A. Batliner, J. Buckow, V. Warnke, and H. Niemann. You beep machine - emotion in automatic speech understanding systems. In *Proceedings of the Workshop on Text, Speech, and Dialog*, pages 223–228, Masark University, 1998.

R. Huber, A. Batliner, J. Buckow, E. Noth, V. Warnke, and H. Niemann. Recognition of emotion in a realistic dialogue scenario. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, volume 1, pages 665–668, Beijing, China, 2000.

C. Izard. *Human emotions*. Plenum, New York, 1977.

C. E. Izard. *Patterns of Emotions: A New Analysis of Anxiety and Depression*. Academic Press, New York, 1972.

W. F. Johnson, R. N. Emde, K. R. Scherer, and M. D. Klinnert. Recognition of emotion from vocal cues. *Arch. Gen. Psych.*, 43:280–283, 1986.

T. Johnstone. Emotional speech elicited using computer games. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1985–1988, Philadelphia, PA, 1996.

T. Johnstone, C. M. van Reekum, and K. R. Scherer. Vocal correlates of appraisal processes. In K. R. Scherer, A. Schorr, and T. Johnstone, editors, *Appraisal processes in emotion: theory, methods, research*. Oxford University Press, New York and Oxford, 2001.

I. Karlsson, T. Banziger, J. Dankovicova, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, and K. R. Scherer. Speaker verification with elicited speaking styles in the verivox project. *Speech Communiation*, 31:207–210, 2000.

C. M. Lee and S. Narayanan. Towards detecting emotion in spoken dialogs. 13(2), 2005.

C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. Emotion recognition based on phoneme classes. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004)*, Jeju Island, Korea, 2004.

D. B. Lindsley. Emotion. In S. S. Stevens, editor, *Handbook of Experimental Psychology*. Wiley, New York, 1951.

G. F. Mahl. The lexical and linguistic levels in the expression of the emotions. In *Expression of the Emotions in Man*. International University Press, New York, 1963.

D. Matsumoto. Cultural influences on the perception of emotion. *Journal of cross-cultural psychology*, 20:92–104, 1989.

S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve. Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 200–205, Belfast, Northern Ireland, 2000.

I. Murray and J. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Amer.*, 93:1097–1108, 1993.

R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the International Conference on Multimedia Computing and Systems*, Florence, Italy, 1999.

T. Nwe, S. Foo, and L. De Silva. Stress classification using subband based features. *IEICE Transactions on Information and Systems, Special Issue on Speech Information Processing*, E86-D:565–573, 2003a.

T. L. Nwe. *Analysis and Detection of Human Emotion and Stress from Speech Signals*. PhD thesis, Department of Electrical and Computer Engineering, National University of Singapore, 2003.

T. L. Nwe, S. W. Foo, and L. C. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, 2003b.

K. Oatley. *Emotions: A brief history*. Blackwell Publishing, Malden, MA, 2004.

D. O'Shaughnessy. *Speech Communications: Human and machine. 2nd Edition*. IEEE Press, New York, 2000.

A. Oster and A. Risberg. The identification of the mood of a speaker by hearing impaired listeners. *Speech Transmission Lab.-Q. Prog. Stat. Rep.*, 4:79–90, 1986.

M. Pantic and J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. pages 1370–1390, 2003.

W. Parrot. *Emotions in Social Psychology*. Psychology Press, Philadelphia, 2001.

V. Petrushin. Emotion in speech: Recognition and application to call centers. In *Artificial Neural Networks in Engineering (ANNIE)*, pages 7–10, St. Louis, Missouri, 1999.

V. Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, 2000.

R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, Massachusetts, 1997.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, Massachusetts, 1998.

R. Plutchik. Emotions, evolution, and adaptive proceesses. In M. B. Arnold, editor, *Feelings and Emotions*. Academic Press, New York, 1970.

T. S. Polzin and A. Waibel. Emotion-sensitive human-computer interfaces. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast, Northern Ireland, 2000.

L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

P. Salovey, M. Kokkonen, P. Lopes, and J. Mayer. Emotional intelligence: What do we know? In A. S. R. Manstead, N. H. Frijda, and A. H. Fischer, editors, *Feelings and emotions: The Amsterdam Symposium*, pages 321–340. Cambridge University Press, Cambridge, UK, 2004.

H. Sato, Y. Mitsukura, M. Fukumi, and N. Akamatsu. Emotional speech classification with prosodic parameters by using neural networks. In *Proceedings of the Australian and New Zealand Intelligent Information Systems Conference*, pages 395–398, 2001.

K. Scherer. Emotion. In M. Hewstone and W. Stroebe, editors, *Introduction to Social Psychology: A European perspective*. Blackwell, Oxford, 2001.

K. R. Scherer. Vocal affect expression: A review and a model for future research. *Psychology Bulletin*, 99:143–165, 1986.

K. R. Scherer. Adding the affective dimension: A new look in speech analysis and synthesis. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, PA, 1996.

K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256, 2003.

M. Scheutz and P. Schermerhorn. The role of signaling action tendencies in conflict resolution. *Journal of Artificial Societies and Social Simulation*, 7, 2004.

B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *Proceedings of the International Conference on Artificial Neural Networks*, Berlin, 1996. Springer Verlag.

A. K. Seewald. Exploring the parameter state space of stacking. 2002a.

A. K. Seewald. How to make stacking better and faster while also taking care of an unknown weakness. In *Proceedings of the 19th International Conference on Machine Learning*, San Francisco, California, 2002b.

A. K. Seewald. Towards a theoretical framework for ensemble classification. 2002c.

C. A. Shipp and L. I. Kuncheva. Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3:135–148, 2002.

E. R. Skinner. A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. *Speech Monogograph*, 2:81–137, 1935.

H. Spencer. *The principles of psychology: Volume I.* Appleton, New York, 1890.

G. Stemmler, M. Heldmann, P. C. A., and T. Scherer. Constraints for emotion specifity in fear and anger: The context counts. *Psychophysiology Review*, 38:275–291, 2001.

D. Talkin. A robust algorithm for pitch tracking (rapt). In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier Science B.V., The Netherlands, 1995.

V. C. Tartter. Happy talk: Perceptual and acoustic effects of smiling on speech. *Percept. Psychophys.*, 27:24–27, 1980.

K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence*, 10:271–289, 1999.

H. Vafaie and K. De Jong. Genetic algorithms as a tool for feature selection in machine learning. In *Proceedings of the 4th International Conference on Tools with Artificial Intelligence*, Arlington, VA, November 1992.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, 1995.

E. Weigand. Emotions: The simple and the complex. In E. Weigand, editor, *Emotion in Dialogic Interaction: Advances in the Complex*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2004.

A. Wierzbicka. *Lexicography and conceptual analysis*. Karoma, Ann Arbor, MI, 1985.

C. E. Williams and K. N. Stevens. Emotions and speech: Some acoustical correlates. In *Nonverbal Communication: Readings with commentary. 2nd Edition.* Oxford University Press, New York, 1972.

I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers, San Fransisco, California, 2000.

D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–260, 1992.

R. S. Woodsworth. *Experimental Psychology*. Holt, New York, 1938.

W. Wundt. *Grundriss der psychologie*. C. H. Judd, translator, 1896.

S. Yacoub, S. Simske, X. Lin, and J. Burns. Recognition of emotion in interactive voice systems. In *Proceedings of Eurospeech 2003, 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

F. Yu, E. Chang, Y.-Q. Xu, and H.-Y. Shum. Emotion detection from speech to enrich multimedia content. In *Proceedings of the Second IEEE Pacific-Rim Conference on Multimedia*, Beijing, China, 2001.

N. Zagalo, A. Barker, and V. Branco. Story reaction structures to emotion detection. In *Proceedings of the 1st ACM workshop on Story representation, mechanism, and context*, pages 33 – 38, New York, NY, 2004.