1-1-2007

# Speech-based human emotion recognition

Talieh Seyed Tabtabae
*Ryerson University*

# SPEECH-BASED HUMAN EMOTION RECOGNITION

by

Talieh Seyed Tabtabaei

B.Eng, Azad Islamic University of Najafabad,
Iran, 2003

A thesis
presented to Ryerson University
in partial fulfillment of the
requirement for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering.

Toronto, Ontario, Canada, 2007

## Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.
Signature

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.
Signature

## Instructions on Borrowers

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

# Abstract

Talieh Seyed Tabatabaei
Speech-Based Human Emotion Recognition
MAS.c, Electrical and Computer Engineering Department,
Ryerson University, Toronto, 2007

Automatic Emotion Recognition (AER) is an emerging research area in the Human-Computer Interaction (HCI) field.

As Computers are becoming more and more popular every day, the study of interaction between humans (users) and computers is catching more attention. In order to have a more natural and friendly interface between humans and computers, it would be beneficial to give computers the ability to recognize situations the same way a human does. Equipped with an emotion recognition system, computers will be able to recognize their users' emotional states and show the appropriate reaction to that. In today's HCI systems, machines can recognize the speaker and also content of the speech, using speech recognition and speaker identification techniques. If machines are equipped with emotion recognition techniques, they can also know "how it is said" to react more appropriately, and make the interaction more natural.

One of the most important human communication channels is the auditory channel which carries speech and vocal intonation. In fact people can perceive each other's emotional state by the way they talk. Therefore in this work the speech signals are analyzed in order to set up an automatic system which recognizes the human emotional state. Six discrete emotional states have been considered and categorized in this research: anger, happiness, fear, surprise, sadness, and disgust.

A set of novel spectral features are proposed in this contribution. Two approaches are applied and the results are compared. In the first approach, all the acoustic features are extracted from consequent frames along the speech signals. The statistical values of features are considered to constitute the features vectors. Support Vector Machine (SVM), which is a relatively new approach in the field of machine learning, is used to classify the emotional states.

In the second approach, spectral features are extracted from non-overlapping logarithmically-spaced frequency sub-bands. In order to make use of all the extracted information, sequence discriminant SVMs are adopted.

The empirical results show that the employed techniques are very promising.

# Acknowledgments

I would like to sincerely thank my supervisors Dr. Aziz Guergachi and Dr. Sridhar Krishnan for their continual guidance and encouragement. I admire them for their generosity in sharing their knowledge and time with me. Without their extensive support and constructive feedbacks this work would not have been possible.

I would also like to thank all members of the Signal Analysis Research (SAR) Group for being so helpful and providing a friendly and peaceful environment for research.

I would like to acknowledge financial support of Electrical and Computer Engineering Department, NSERC, CFI, and OIT during the coarse of my studies.

Last but certainly not least, I would like to thank my husband. I could not have achieved this goal without his love, encouragement, and support.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

EMOTION is a fundamental component of being a human. Joy, sadness, anger, and fear, among the plethora of other emotions, motivate action and add meaning and richness to virtually all human experience. Traditionally, human-computer interaction (HCI) has been viewed as the ultimate exception; users must discard their emotional selves to work efficiently and rationally with computers, the quintessentially unemotional artifact. Emotion seemed at best marginally relevant to HCI, and at worst, an oxymoron. Recent research in psychology and technology suggests a very different view of the relationship between humans, computers, and emotion. After a long period of dormancy and confusion, there has been an explosion of research on the psychology of emotion (Gross, 1999). Emotion is no longer seen as limited to the occasional outburst of fury when a computer crashes inexplicably, excitement when a video game character leaps past an obstacle, or frustration at an incomprehensible error message. It is now understood that a wide range of emotions plays a critical role in every computer-related, goal-directed activity, from developing a three-dimensional computer-aided design (CAD) model and running calculations on a spreadsheet, to searching the Web and sending an e-mail, and to making an online purchase. Indeed, many psychologists now argue that it is impossible for a person to have a thought or perform an action without engaging, at least unconsciously, his or her emotional systems (Picard, 1997b). The literature on emotions and computers has also grown dramatically in the past few years, driven primarily by advances in technology. Inexpensive and effective technologies that en-

able computers to assess the physiological correlates of emotion, combined with dramatic improvements in the speed and quality of signal processing, now allow even personal computers to make judgments about the user's emotional state in real time (Picard, 1997a). Multimodal interfaces that include voices, faces, and bodies can now manifest a much wider and more nuanced range of emotions than was possible in purely textual interfaces (Cassell, Sullivan, Prevost, and Churchill, 2000). Indeed, any interface that ignores a user's emotional state or fails to manifest the appropriate emotion can dramatically impede performance and risks being perceived as cold, socially inept, untrustworthy, and incompetent.

This chapter reviews the psychology of emotion and the related technologies, with an eye toward identifying those concepts that are most relevant to the design and assessment of interactive systems. The seat of emotion is the brain; hence, a brief description of the psychophysiological phenomena that lies at the core of how emotions emerge from interaction with the environment is presented. Then we talk about Human Computer Interaction (HCI) and the position of emotion in HCI.

## 1.1 Human Emotion

The mainstream definition of emotion refers to a state of feeling involving thoughts, physiological changes, and an outward expression or behavior [32]. The contributions to the subject of emotions come from so many different disciplines. In recent years, especially the last decade, knowledge in the field of emotion has been steadily increasing. This knowledge comes from many different areas: psychology, neurology, ethology, physiological psychology, personality and social psychology, clinical psychology and psychiatry, medicine, nursing, and social work are all directly concerned with emotion. Professions such as law and architecture have an obvious concern with emotions as they affect human motives and needs. The various branches of art, specially the performing arts, certainly deal with the emotions and their expressions. A flurry of recent work in modeling emotional circuitry and recognition has come out of computer science, and engineering mostly in the applications of intelligent human-machine interaction. There is a wide range of scientific opinions regarding the nature

and importance of emotions. Some scientists (Duffy, 1962) have maintained that emotion concepts are unnecessary for the science of behavior. She, as well as others (i.e. Lindsley, 1957), suggested that the concept of activation or arousal has more explanatory power and is less confusing than emotion concepts. Others (Tomkins, 1962, 1963; Izard, 1971, 1972) have maintained that the emotions constitute the primary motivational system of human beings. Some say that emotions are only transient phenomena while others maintain that people are always experiencing some emotion to some extent (e.g. Schachtel. 1959). Some scientists have maintained that for the most part emotions disrupt and disorganize behavior, and are primarily a source of human problems (Arnold, 1960). Others have argued that emotions play an important role in organizing, motivating, and sustaining behavior (Rapaport, 1942; Leeper, 1948). Some scientists have taken the position that emotions are primarily a matter of visceral functions, activities of organs innervated by autonomic nervous system (Wenger, 1950). Other scientists have emphasized the importance of the externally observable behavior of the face, the voice tone and intonation, and the role of the nervous system (Gellhorn, 1964, 1970).

### 1.1.1 What Are Emotions?

Most theories either explicitly or implicitly acknowledge that an emotion is not a simple phenomenon. It cannot be described completely by having a person describe his emotional experience. It cannot be described completely by electrophysiological measures of occurrence in the brain, the nervous system, or in the circulatory, respiratory, and glandular systems. It cannot be described completely by the expressive or motor behavior that occurs in emotion. A complete definition of emotion must take into account all of these three aspects or components: (a) the experience or conscious feeling of emotion, (b) the processes in the brain and nervous system, and (c) the observable expressive patterns of emotion, particularly those on the face and in the vocal system [32].

### 1.1.2 How Do Emotions Occur?

Most people know what kind of conditions or situations interest them or disgust them or make them feel angry or guilty. Thus in general they know what brings about a given emotion. However, scientists do not agree on precisely how an emotion comes about. Some maintain that emotion is a joint function of a physiologically arousing situation and the person's evaluation or appraisal of the situation [32]. This explanation of the causal process comes from a cognitive theory of emotion (Schachter, 1971). Considering the problem at the neurological level, Tomkins (1962) maintains that emotions are activated by changes in the density of neural stimulation (the number of neurons firing per unit of time). This rather persuasive theory does not say much about the causes or conditions at the conscious level that trigger these changes in neural stimulation.

### 1.1.3 Components of Emotion

The component that seems to be the core of common sense approaches to emotion, the one that most people have in mind when talking about human emotions, is the feeling component, i.e., the passion or sensation of emotion. For example, people generally agree that the state of mind during anger is different from that when one is happy. However, this component is not observable and measurable by other people in order to distinguish the emotions and also is not considered a helpful component for a HCI system.

Another obvious descriptive component of emotion is the set of behaviors that may be performed and observed in conjunction with an emotion. These behaviors are produced by the muscular system and are of two general types: gross behaviors of the body effected by the skeletal muscles and the so-called *emotion expressions* [35]. These categories shade into each other because any behavior can be interpreted as an expression of emotion. The gross body behaviors may have no apparent adaptive value, e.g., wringing and rubbing the hands or tapping a foot, or they may be directed towards a goal, e.g., striking something or running away. The facial and bodily behaviors called emotion expressions are indicators of emotion, as opposed to effecting some action or achieving some goal. These expressions

can differentiate one emotion from another. The most widely discussed and investigated emotion expressions both in emotion communication between people and in HCI systems are the emotion faces and vocal intonation.

A less obvious component of emotion is the set of internal bodily changes caused by the smooth muscles and glands. Chemicals secreted by the body's various glands are activated during emotion and spread to other parts of the body, usually by the blood, to act in diverse ways on the nervous system and other organs. Smooth muscles of the digestive system, circulatory system, and other bodily components can shift from their typical level or type of operation during emotion under the effects of chemical and neural action. This component includes some behaviors that can be observed, such as the constriction or dilation of the iris of the eye, possibly piloerection, and sweating, blanching, and flushing of the skin, and other responses that are relatively hidden, such as heart rate, stomach activity, and saliva production. Autonomic activity has received considerable attention in studies of emotion, in part due to the relative ease in measuring certain components of the autonomic nervous system, including heart rate, blood pressure, blood pulse volume, respiration, temperature, pupil dilation, skin conductivity, and more recently, muscle tension (as measured by electromyography). However, the extent to which emotions can be distinguished on the basis of autonomic activity alone remains a hotly debated issue (Ekman and Davidson, 1994; Levenson, 1988). Although the debate is far from resolved, certain measures have proven fairly reliable at distinguishing among "basic emotions". Heart rate, for example, increases most during fear, followed by anger, sadness, happiness, surprise, and finally disgust, which shows almost no change in heart rate (Cacioppo, Bernston, Klein, and Poehlmann, 1997; Ekman, Levenson, and Friesen, 1983; Levenson, Ekman, and Friesen, 1990). Decreases in heart rate typically accompany relaxation, attentive visual and audio observation, and the processing of pleasant stimuli (Frijda, 1986). However, even assuming that we could distinguish among all emotions through autonomic measures, it is not clear that we should. In real-world social interactions, humans have at least partial control over what others can observe of their emotions. If another person, or a computer, is given direct access to users' internal states, they

may feel overly vulnerable, leading to stress and distraction. Such personal access could also be seen as invasive, compromising trust. It may, therefore, be more appropriate to rely on measurement of the external signals of emotion [35].

Another less observable component in emotion consists of Neurological Responses. The brain is the most fundamental source of emotion. The most common way to measure neurological changes is the electroencephalogram (EEG). In a relaxed state, the human brain exhibits an alpha rhythm, which can be detected by EEG recordings taken through sensors attached to the scalp. Disruption of this signal (alpha blocking) occurs in response to novelty, complexity, and unexpectedness, as well as during emotional excitement and anxiety (Frijda, 1986). EEG studies have further shown that positive emotions lead to greater activation of the left anterior region of the brain, whereas negative emotions lead to greater activation of the right anterior region (Davidson, 1992; see also Heller, 1990). Indeed, when one flashes a picture to either the left or the right of where a person is looking, the viewer can identify a smiling face more quickly when it is flashed to the left hemisphere, and a frowning face more quickly when it is flashed to the right hemisphere (Reuter-Lorenz and Davidson, 1981). Current EEG devices, however, are fairly clumsy and obstructive, rendering them impractical for most HCI applications.

Finally the ideation, imagery, and thoughts that occur during emotion can be considered as another component of the emotion process. These aspects of emotion are also cognitive activities, and can both give rise to an emotional event and be affected by it, e.g., thinking about a lost pet may evoke feelings of sadness, which may in turn evoke memories of a romance now finished. Since thoughts and other cognitions, like feelings, cannot be directly observed and are hard to measure, there is less understanding of how they fit into the emotion picture than other components.

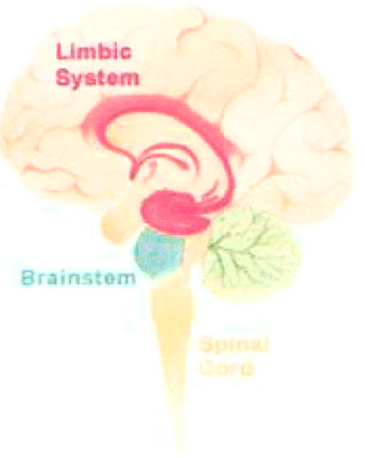


**Figure 1.1:** Limbic system of the brain [34]

### 1.1.4 Emotions and Brain

Emotions are thought to be related to activity in brain areas that direct our attention, motivate our behavior, and determine the significance of what is going on around us. Pioneering work by Broca (1878), Papez (1937), and MacLean (1952) suggested that emotion is related to a group of structures in the center of the brain called the limbic system, which includes the hypothalamus, cingulate cortex, hippocampi, and other structures (Fig. 1.1). The limbic system (Latin limbus: "border" or "edge") includes the structures in the human brain involved in emotion, motivation, and emotional association with memory. The limbic system influences the formation of memory by integrating emotional states with stored memories of physical sensations. More recent research has shown that some of these limbic structures are not as directly related to emotion as others are, while some non-limbic structures have been found to be of greater emotional relevance. The following brain structures are currently thought to be most involved in emotion [34]:

- Amygdala - The amygdalae are two small, round structures located anterior to the hippocampi near the temporal poles. The amygdalae are involved in detecting and learning what parts of our surroundings are important and have emotional significance. They are critical for the production of emotion, and may be particularly so for negative

emotions, especially fear.

- Prefrontal cortex - The term prefrontal cortex refers to the very front of the brain, behind the forehead and above the eyes. It appears to play a critical role in the regulation of emotion and behavior by anticipating the consequences of our actions. The prefrontal cortex may play an important role in delayed gratification by maintaining emotions over time and organizing behavior toward specific goals.

- Anterior Cingulate - The anterior cingulate cortex (ACC) is located in the middle of the brain, just behind the prefrontal cortex. The ACC is thought to play a central role in attention, and may be particularly important with regard to conscious, subjective emotional awareness. This region of the brain may also play an important role in the initiation of motivated behavior.

- Ventral striatum - The ventral striatum is a group of subcortical structures thought to play an important role in emotion and behavior. One part of the ventral striatum called the nucleus accumbens is thought to be involved in the experience of goal-directed positive emotion. Individuals with addictions experience increased activity in this area when they encounter the object of their addiction.

- Insula - The insular cortex is thought to play a critical role in the bodily experience of emotion, as it is connected to other brain structures that regulate the body's autonomic functions ( heart rate, breathing, digestion, etc.). This region also processes taste information and is thought to play an important role in experiencing the emotion of disgust.

Based on discoveries made through neural mapping of the limbic system, the neurobiological explanation of human emotion is that emotion is a pleasant or unpleasant mental state organized in the limbic system of the human brain.

In fact, emotions are human elaborations of general arousal patterns, in which neurochemicals (e.g., dopamine, noradrenaline, and serotonin) step-up or step-down the brain's

activity level, as visible in body movements, gestures, and postures [34]. In human beings, feelings are displayed as emotion cues.

### 1.1.5 Are Emotions Innate or Learned?

The early work of Darwin (1872, 1877) and the more recent work of Ekman et al. (1972) and Izard (1971) has shown that certain emotions, referred to as basic (or alternatively fundamental) emotions (see Section 1.1.6), have the same expressions and experiential qualities in widely different cultures from virtually every continent of the globe, including isolated preliterate cultures having had virtually no contact with civilization [32]. Therefore it can be concluded that the fundamental emotions are subserved by innate neural programs. However, the fact that there are genetically based mechanisms for the fundamental emotions does not mean that no aspect of an emotion can be modified through experience. Almost anyone can learn to modify the innate emotion expressions [32]. This cognitive part actually has contributed to a relatively new field, called Emotional Intelligence (EI) [36], which describes an ability, capacity, or skill to perceive, assess, and manage the emotions of one's self, of others, and of groups. People of different social backgrounds and different cultures may learn quite different facial movements for modifying innate expressions. In addition to learning modifications of emotion expressions, sociocultural influences and individual experiences play an important role in determining what will trigger an emotion and what a person will do as a result of emotion.

### 1.1.6 Classification of Emotions

One broad classification of emotion is to classify emotions simply as positive or negative. Scientists as well as laymen agree that there are both positive and negative emotions. While this very broad classification of emotions is generally correct and useful, the concepts of positiveness and negativeness as applied to the emotions require some qualification [32]. Emotions such as anger, fear, and shame cannot be considered categorically negative or bad. Anger is sometimes positively correlated with survival, and more often with the defense of

maintenance of personal integrity and the correction of social injustice [33]. So instead of saying that emotions are merely positive or negative, it is more accurate to say that there are some emotions which tend to lead to psychological entropy, and others which tend to facilitate constructive behavior or the converse of entropy.

One of the most influential classification approaches in the study of emotion is Robert Plutchik's eight primary emotions. The emotions that Plutchik lists as primary are: anger, fear, sadness, joy, disgust, surprise, curiosity, and acceptance. Similar to the way primary colors combine, primary emotions are believed to blend together to form the full spectrum of human emotional experience. Plutchik reasons that these eight are primary on evolutionary grounds, by relating each to behavior with survival value. For example: fear motivates flight from danger, anger motivates fighting for survival. They are considered to be part of our biological heritage and built into human nature [33].

Paul Ekman [32] devised a similar list of basic emotions from cross-cultural research. He found that even members of an isolated, stoneage culture could reliably identify the expressions of emotion in photographs of people from cultures which they were not yet familiar with, and concluded that the facial expression of some basic emotions is innate. The following is Ekman's list of basic emotions: *anger*, *fear*, *sadness*, *happiness*, *surprise*, and *disgust* [32]. Ekman believes that there are *discrete*, *basic*, *universal* emotions each of which has unique physiological arousal patterns, behavioral display patterns, motivational values, etc. Ekman's list of basic emotions is perhaps the most well-known classification of emotions, which is also used in this thesis.

### 1.1.7 Emotional Communication in Human Beings

In some theories, emotional expression is regarded as an integral aspect of the emotion process. Some theorists have proposed that emotional expression underlies the experience of emotion, which includes the felt quality of emotion.

A large body of literature shows that emotions are communicated both nonverbally and verbally. On the nonverbal side, emotions are typically accompanied by nonverbal expres-

sions such as facial expression (Buch, 1984), body gesture, and voice (Burgoob, Buller, and Woodall, 1996). Emotions are also expressed through verbal communication that implicitly or explicitly reveals the emotions that a person is experiencing [37]. In fact, body language and tone of voice are important parts of emotional communication.

Tone of voice reflects psychological arousal, *emotion*, and mood. Tone of voice may provide a hint of the feelings that a person is unable to put into words. Emotions have a global impact on speech since they modulate the respiratory system, larynx, vocal tract, muscular system, heart rate, and blood pressure [32]. Changes in the speaker's autonomic nervous system can account for some of the most significant changes, where the sympathetic and parasympathetic subsystems regulate arousal in opposition. For instance, when a subject is in a state of fear, anger, or joy, the sympathetic nervous system is aroused. This induces an increased heart rate, higher blood pressure, changes in depth of respiratory movements, greater sub-glottal pressure, dryness of the mouth, and occasional muscle tremor. The resulting speech is faster, louder, and more precisely enunciated with strong high frequency energy, a higher average pitch, and wider pitch range. In contrast, when a subject is tired, bored, or sad, the parasympathetic nervous system is more active. This causes a decreased heart rate, lower blood pressure, and increased salivation. The resulting speech is typically slower, lower-pitched, more slurred, and with little high frequency energy [38]. Body language gives an additional clue. Study of emotion on facial expressions constitutes a vast part of emotion expression in literature. Sometimes the way that the body is positioned or even the hands can express what a person is feeling. But the connection between gesture and emotional state is less distinct, in part due to the greater influence of personality and culture (Casseli and Thorisson, 1999; Coffier, 1985).

Typically, a facial or vocal expression of emotion is presented to another person(s), who then indicates which emotion it signals. The impressive empirical foundation for this theory is the repeated finding that, despite differences in culture, age, or background, receivers agree on the emotion signaled more often than could be achieved by chance.

## 1.2 Human Computer Interaction

Human-computer interaction (HCI), alternatively man-machine interaction (MMI) or computer-human interaction (CHI) is the study of interaction between people (users) and computers. It is an interdisciplinary subject, relating computer science with many other fields of study and research. Interaction between users and computers occurs at the user interface (or simply interface), which includes both software and hardware. A basic goal of HCI is to improve the interaction between users and computers by making computers more usable and receptive to the user's needs [39].

A long term goal of HCI is to design systems that minimize the barrier between the human's cognitive model of what they want to accomplish and the computer's understanding of the user's task. HCI is an interdisciplinary area. It is emerging as a specialty concern within several disciplines, each with different emphases: computer science and engineering (application design and engineering of human interfaces), psychology (the application of theories of cognitive processes and the empirical analysis of user behavior), sociology and anthropology (interactions between technology, work, and organization), and industrial design (interactive products). Because human-computer interaction studies a human and a machine in communication, it draws from supporting knowledge on both the machine and the human side. On the machine side, techniques in computer graphics, machine learning algorithms, signal processing techniques, programming languages, and development environments are relevant. On the human side, communication theory, graphic and industrial design disciplines, linguistics, social sciences, cognitive psychology, and human performance are relevant.

HCI has been an important research area in the fields of multimedia and telecommunication. Some applications in the HCI are speaker recognition, speaker verification, speech recognition, face recognition, gesture recognition, and more recently emotion recognition.

The topic of emotion in Human-Computer Interaction is of increasing interest to the HCI community. Since Picard's fundamental publications on affective computing [40], research in this field has gained significant momentum. Emotion research is largely grounded in psychology yet spans across numerous other disciplines. The challenge of such an inter-

disciplinary research area is developing a common vocabulary and research framework that a mature discipline requires. What is increasingly needed for advanced and serious work in this field is to place it on a rigorous footing, including developing theoretical fundamentals of HCI-related emotion research, understanding emotions' function in HCI, ethical and legal issues, and the practical implications and consequences for the HCI community. The first workshop on emotion in HCI held in Edinburgh in 2005 brought an interdisciplinary group of practitioners and researchers together for a lively exchange of ideas, discussion of common problems, and identification of domains to explore.

## 1.3 Emotion in HCI

Research related to emotion in HCI often tends to focus on how a computer can autonomously detect the emotional state of a user and then adapt itself accordingly. Another important strand of emotion-related research in HCI is the simulation of emotional expressions made by computer agents. Interface designers often include emotional expressions and statements in their interfaces through the use of textual content, speech (synthetic and recorded), and synthetic facial expressions.

Today emotions are more accepted as an important ingredient of human life. Several studies show that emotions play a vital role in almost everything we do, for example in cognitive functions, including rational decision making and learning, and perception. In every day life we experience a rich variety of situations; from walking through a park full of fresh flowers to working out different functions of our new mobile phone or simply having a coffee with a friend. Emotions and affective responses are central parts of our experience and frequently shape and colour the kinds of experience we have [32]. Recently, the concepts of user experience and emotion have been receiving growing attention within the Human-Computer Interaction community as a way of adding value when designing products. It is no longer sufficient for a product to be simply usable or aesthetically pleasing, but it needs to evoke positive emotional responses [41]. In fact, emotion appears to be a necessary component of intelligent, friendly computers. The inability of today's computers to recognize, express, and

have emotions severely limits their ability to act intelligently and interact naturally with us.

## 1.4 Applications

As mentioned before, most of the potential applications of Automatic Emotion Recognition (AER) systems are in the field of HCI. In today's HCI systems, machines can recognize the speaker and also content of the speech, using speech recognition and speaker identification techniques. If machines are equipped with emotion recognition techniques, they can also know "how it is said" to react more appropriately, and make the interaction more natural. One example of computers with emotional intelligent is RoCo (Robotic Computer) [42]. Roco is a new type of desktop computer that has an articulated "neck" and "head" (a computer monitor that can be moved in a fluid manner relative to its base via motors and sensors). RoCo is capable of recognizing and physically responding to human socio-emotive cues such as postural shifts in principled ways. These cues are inspired by those found in human-human interaction, to foster a more natural, healthy, and productive human-computer interaction.

Other potential application of AER is intelligent toys such as Furby's EMOTO-TRONICS [43], Sony's AIBO [44], and Paro [45] as they are shown in Fig. 1.2. All these toys are equipped with AER systems, therefore they are able to recognize their owners' emotional state and be affectionate.

Another potential application of automatic emotion recognition is in e-learning applications, where affective computing can be used to adjust the presentation of a computerized tutor when a learner is bored, interested, frustrated, or pleased. Psychological health services such as counseling, can also benefit from AER applications, for example, when determining a client's emotional state. AER has also been suggested to apply in monitoring society. For example a car which can monitor the emotion of its occupants may engage additional safety measures, such as alerting other vehicles, if it detects the driver is angry. Another example is in telephone call center conversations in order to provide feedback to an operator or a supervisor for monitoring purposes. Other potential applications of AER consist of lie

**Figure 1.2:** Samples of intelligent toys which are able to show affection to people (from left to right: Sony's AIBO, Furby's EMOTO-TRONIC, and Paro).

detection, customer service, and educational software.

## 1.5 Organization of Thesis

The main objective of this thesis is to suggest a speech-based automatic emotion recognition system using a novel set of acoustic features and utilizing powerful and state-of-the-art machine learning methods. The remainder of the thesis is organized as shown in Fig. 1.3.

**Chapter 2** reviews some of the previous works on AER. Since audio and visual channels are the most important communication channels in humans, the concentration of this chapter is also on the researches which have utilized audio or/and visual information as input to their systems. The second part of the chapter addresses some of the existing databases for the application of emotion recognition and also explains the database which is used in this thesis.

**Chapter 3** presents a frame-based approach to emotion recognition where a set of novel acoustic features is proposed. Least square support vector machine is used to classify the emotional classes. The corresponding technical procedures are explained in detail.

**Chapter 4** proposes a novel sub-band approach AER system, where spectral features are extracted from non-overlapping logarithmically-spaced frequency sub-bands. The problem of variable-length sequences is addressed in this chapter, some of the alternative solutions are discussed and Sequence discriminant SVM (Fisher kernel) is proposed to overcome this issue.

**Chapter 5** summarizes the implemented methods and compares the achieved results. Some of the advantages and disadvantages of the adopted techniques are also discussed. The

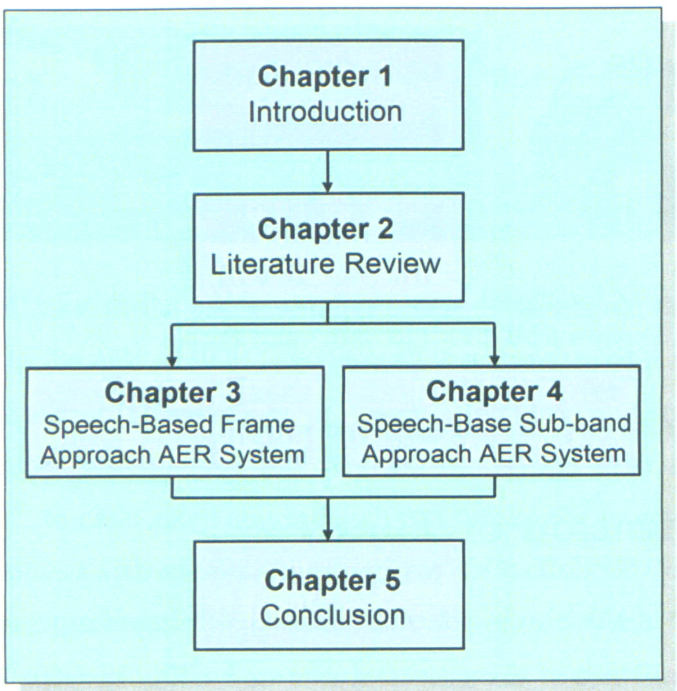


**Figure 1.3:** Organization of thesis.

last section of the chapter points out directions for future research.

# Chapter 2

# Automatic Emotion Recognition

THE popularity of computers has rapidly increased due to the progress of information technologies. Accordingly, research on human and computer interlace is gaining more interest in order to have a more natural and friendly interface between humans and machines. Related to this, various research projects on emotion recognition have been performed. Researchers have used diverse techniques and approaches aiming at getting a satisfactory result. In this chapter a brief review of previous works on emotion recognition and their employed techniques are presented and eventually our proposed method is addressed. The database utilized in this thesis is also described in the second part of the chapter.

## 2.1 Literature Review

The important and very basic steps of almost every Automatic Emotion Recognition (AER) system as Fig. 2.1 shows, are extracting some emotional data (i.e. features) from some kind of input to the system and then classifying the extracted information from the input to one of the predefined emotions. Different studies differ in the type of inputs they choose for their systems, the kind of features they extract, and the methods of classification they adopt.

A great number of studies have been performed on emotion analysis utilizing inputs such as voice (i.e. paralinguistic information), facial expression, body language, physiological signals (e.g. EEG, ECG, skin temperate variation, etc.), linguistic information of the speech, etc., or combination of two or more of these (multimodal approach). However, most of the

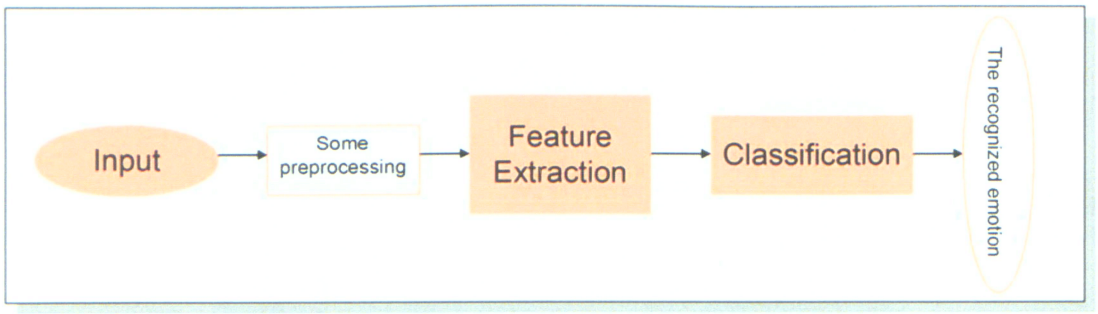


**Figure 2.1:** A very basic AER system.

researchers use voice or/and body language as input to their AER system. The main reason is that human-computer interaction follows human-human interaction in the basis, and as "nonverbal communication" (e.g. facial expressions, tone of voice, gesture, eye contact, etc.) plays a crucial role in humans' communication, it is also considered a very important factor in HCI. As mentioned in Chapter 1, the auditory channel (i.e. speech and tone of voice) and the visual channel (i.e. facial expression and body gesture) are the two primary emotion communication channels in humans and it is natural to use these channels in HCI systems as well. Also, using audio or visual information, it is much easier to collect data without causing any discomfort for users and as a result, it is more practical for the real-world applications.

Extracting the efficient and relevant features which can truly represent the state of emotion in the input data is a great challenge in every AER system. A vast variety of classifiers has also been adopted in different studies.

The next few chapters present different methods and schemes suggested by other researchers. Although facial expression is not used in this thesis as an input to the AER system, some of the existing works relating to this are reviewed due to the popularity and importance of the usage.

### 2.1.1 Emotion Recognition Using the Visual Channel

Reflection of emotions via the visual channel consists of emotion expression in face and in body gesture. As described in Section 1.1.7, it is hard to construct an emotion recognition

**Table 2.1:** Facial cues and emotions (Based on Ekman and Friesen, 1975)

| Emotion | Observed Facial Cues |
|---|---|
| Surprise | Brows raised (curved and high)<br>Skin below brow stretched<br>Horizontal wrinkles across forehead<br>Eyelids opened and more of the white of the eye is visible<br>Jaw drop open without tension or stretching of the mouth |
| Fear | Brows raised and drawn together<br>Forehead wrinkles drawn to the center<br>Upper eyelid is raised and lower eyelid is drawn he lower lid up<br>Mouth is open<br>Lips are slightly tense or stretched and drawn back |
| Disgust | Upper lip is raised<br>Lower lip is raised and pushed up to upper lip or it is lowered<br>Noes is wrinkled<br>Cheeks are raised<br>Lines below the lower lid, lid is pushed up but not tense<br>Brows are lowered |
| Anger | Brows lowered and drawn together<br>Vertical lines appear between brows<br>Lower lid is tensed and may or may not be raised<br>Upper lid is tense and may or may not be lowered due to brows' action<br>Eyes have a hard stare and may have a bulging appearance<br>Lips are either pressed firmly together with corners straight or down or open, tensed in a squarish shape<br>Nostrils may be dilated |
| Happiness | Corners of lips are dawn back and up<br>Mouth may or may not be parted with teeth exposed or not<br>A wrinkle runs down from the nose to the outer edge beyond lip corners<br>Cheeks are raised<br>Lower eyelid shows wrinkle below it and may be raise but not tense |
| Sadness | Inner corners of eyebrows are drawn up<br>Skin below the eyebrow is triangulated, with inner corner up<br>Upper lid inner corner is raised<br>Corners of the lips are drawn or lip is trembling |

system based on gesture since the way people express their emotion in body language highly depends on their cultural background and personality. Therefore, most of the studies use only facial expressions as visual emotional information. Table 2.1 describes characteristic facial features of six basic emotions (Ekman and Friesen, 1975).

Emotion recognition from facial expressions can be performed using either a single image or image sequence; either way the face region should be detected from the image first. In the case of still images, information is extracted from the detected face in the image. In the case of image sequence, the motion of the detected face and its features in the sequence is tracked.

Analyzing facial expression can be performed when the face is represented as a whole unit (holistic representation) or when prominent components of the face such as nose, mouth, eyes, and chin are considered as features. An example of these techniques used for the former approach is eigenfaces [1] which transforms face images into a small set of characteristic feature images, called "eigenfaces". Eigenfaces are the principal components of the initial training set of face images. Another example is using Gabor wavelet features to represent facial expression [6]. Facial Action Coding System (FACS) [18] is the most popular representation of facial expression for the latter approach. FACS is a system originally developed by Paul Ekman and Wallace Friesen in 1976, to taxonomize every conceivable human facial expression. It is the most popular standard currently used to systematically categorize the physical expression of emotions, and it has also proven useful both to psychologists and to animators. It defines expressions as one of 46 "Action Units" (AUs), which are a contraction or relaxation of one or more muscles.

A complete survey on the research regarding facial expression recognition can be found in [8] and [9].

Cowie *et al.* [75] have chosen to measure specific facial feature deformations (e.g. eyebrows, eyes, mouth) and create appropriate descriptive expression models to develop a rule-based system capable of analyzing image frames from a video stream of a speaker into MPEG-4 compliant Facial Definition Parameters (FDPs). FDPs are in turn used to calculate the

Facial Animation Parameters (FAPs). The FAPs can correlate strongly with emotionality and can be used to classify a face with respect to the emotional state it expresses. They use the Ekman dataset to categorize the six universal emotions. In their approach feature extraction results in a set of binary maps, indicating the position and extent of each facial feature (i.e. eyebrows, eyes, mouth and nose). The left, right, top and bottom-most coordinates of the eye and mouth, the left, right, and top coordinates of the eyebrow as well as the nose coordinates, are the facial feature points (FPs) which are used in [11] for defining the FAP values. By using unsupervised hierarchical clustering technique, they were able to achieve 84.7% accuracy.

Lien *et al.* [2] have developed a facial expression recognition system that automatically recognizes individual action units or action unit combinations in the upper face using Hidden Markov Models (HMMs). Their approach to facial expression recognition is based on the Facial Action Coding System (FACS), which separates expressions into upper and lower face action. They use three approaches to extract facial expression information: (1) facial feature point tracking, (2) dense flow tracking with principal component analysis (PCA), and (3) high gradient component detection (i.e., furrow detection). The recognition results of the upper face expressions using feature point tracking, dense flow tracking, and high gradient component detection are 85%, 93%, and 85%, respectively. Sixty subjects, both male and female, were used in their study. Their goal was to develop a system that recognizes subtle feature motion and complex facial expressions rather than six prototypic expressions.

Byun *et al.* [3] proposed a novel algorithm for hybrid feature extraction from still images, which uses not only emotional features that is perceived by human eye, but also various emotional information that is extracted by image processing. They apply the geometrical feature extraction method to extract the relative position, size, angle, and vector of numerical data from distinctive features such as eyes, eyebrows, nose, mouth, and chin, and also the RGB color distributed histogram method that is newly applied to the feature extraction stage. This paper applies face detection by RGB skin-color model. Skin colored region is selected and stored as the RGB type for the training data. Their overall accuracy is not reported.

Kotsia and Pitas [5] present two novel methods for facial expression recognition in facial image sequences. The user has to manually place some Candide grid nodes to face landmarks depicted at the first frame of the image sequence under examination. The grid-tracking and deformation system which is used based on deformable models, tracks the grid in consecutive video frames over time as the facial expression evolves, up to the frame that corresponds to the greatest facial expression intensity. The geometrical displacement of certain selected Candide nodes, defined as the difference of the node coordinates between the first and the greatest facial expression intensity frame, is used as an input to a multiclass Support Vector Machine (SVM) system of classifiers that are used to recognize either the six basic facial expressions or a set of chosen Facial Action Units (FAUs). Fig. 2.2 shows an example of the deformed frame facial expression models produced for each one of the six basic facial expressions in this research. In their proposed approach, the facial expression classification is performed based only on geometrical information, without taking directly into consideration any facial texture information. They also have developed a novel method of multiclass SVM by manipulating the original SVM's formulations. The Cohn-Kanade database [4] was used in this research to classify facial expressions into one of the six basic facial expression classes. This database is annotated with FAUs. They show a recognition accuracy of 99.7% for facial expression recognition using the proposed multiclass SVM and 95.1% for facial expression recognition based on FAU detection.

Gabor filters are used in [6] by Lyons *et al.* for facial expression recognition. Facial expression images are coded using a multi-orientation, multi-resolution set of Gabor filters which are topographically ordered and aligned approximately with the face. The similarity space derived from this facial image representation is compared with one derived from semantic ratings of the images by human observers. They have collected a database which consists of ten subjects each of which posed 3 or 4 examples for each of the six basic facial expressions (i.e. happiness, sadness, surprise, anger, disgust, fear) and also a neutral face for a total of 219 images of facial expressions. For simplicity of experimental design they have only employed Japanese female subjects. The classification is performed by comparing

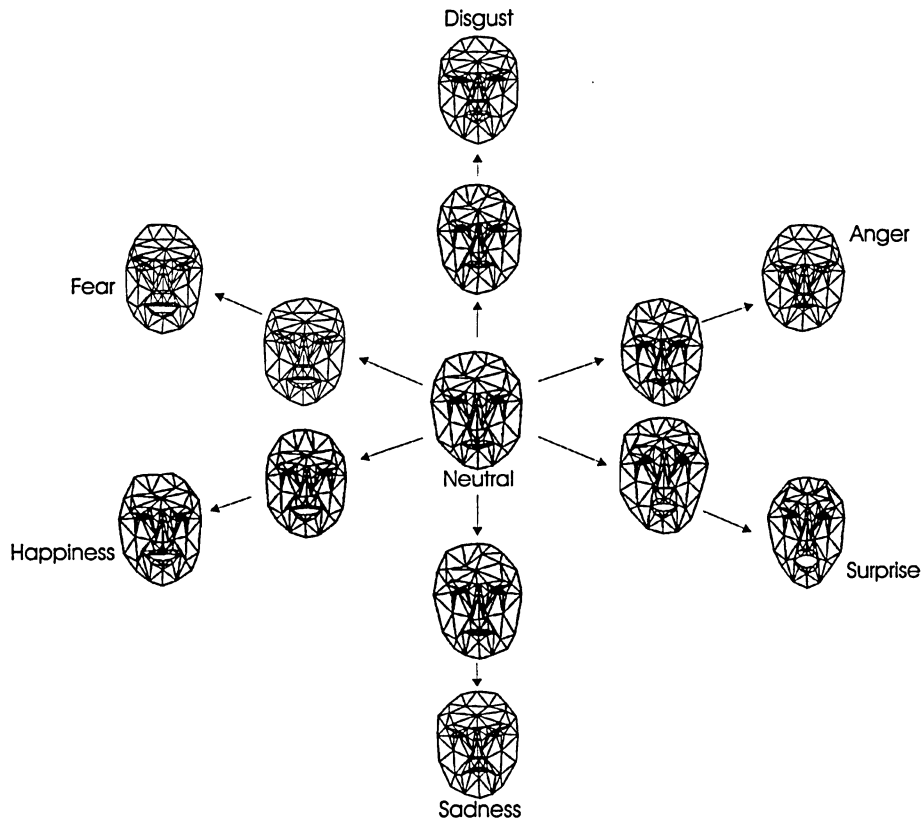


**Figure 2.2:** Example of the deformed Candide grids for each one of the six facial expressions created in [5]

the produced similarity spaces. The images are first transformed using a multi-scale, multi-orientation set of Gabor filters. The grid is then registered with the facial image region either automatically, using elastic graph matching or by manual clicking on fiducial face points. The amplitude of the complex valued Gabor transform coefficients are sampled on the grid and combined into a single vector, called a Labeled Graph Vector (LGV). The classification is performed using the distance of the LGV from each facial expression cluster center.

### 2.1.2 Emotion Recognition Using Auditory Channel

Speech is one of the indispensable communication channels in humans. Recognition of emotions via auditory channel consists of understanding the emotion expressed explicitly through

words and/or implicitly through tone of voice. However, research suggests that nonverbal communication is more important in understanding human behavior than words alone; the "nonverbal channels" seem to be more powerful than what people say. In its broadest definition, nonverbal communication is, according to Hecht, DeVito, and Guerrero, "all the messages other than words that people exchange in interactive contexts". Voice can provide indications of specific emotions through acoustic properties such as pitch range, rhythm, and amplitude or duration changes (Ball and Breese, 2000; Scherer, 1989). A bored or sad user, for example, will typically exhibit slower, lower-pitched speech, with little high-frequency energy, whereas a user experiencing fear, anger, or joy will speak faster and louder, with strong high-frequency energy and more explicit enunciation (Picard, 1997a). Murray and Arnott (1993) provide a detailed account of the vocal effects associated with several basic emotions (see Table 2.2).

Feature extraction is a very important and decisive part of every speech-based AER system. A substantial body of existing works on automatic emotion recognition based on speech use prosodic features. Prosody deals with the rhythmic patterns of spoken language, including stress and intonation. Acoustically, prosody describes changes in the syllable length, loudness, pitch, and certain details of the formant structure of speech sounds. Phonologically, prosody is described by tone, intonation (i.e. the contour of the pitch pattern; whether there is a rising or falling tone at the end of the pattern ), rhythm (i.e. how the words are grouped together ), and lexical stress (i.e. where the main accent occurs).

Schüller *et al.* in [10] use both acoustic features and language information of speech utterances in their database to construct their speech-based AER system. They are dealing with emotion recognition in an automotive environment. The emotion corpus used in [10] consists of German and English sentences of 13 speakers, one female and 12 male. They conduct their experiment in both person-dependent and person-independent situations to classify seven emotion categories: anger, disgust, fear, joy, neutral, sad, and surprise. The set of acoustic features used in this work is static features of prosodic analysis. For acoustic features classification they compare the performance of several classifiers throughout their

work:

1. k-means classifier, where the criterion function is the Euclidean distance between the class mean vectors.

2. k-nearest-neighbors (k-NN) classifier, where the unknown sample is assigned to the class with majority vote.

3. Gaussian Mixture Model (GMM), with 16 Gaussian models where the well-known Expectation Maximization (EM) algorithm is used to find the model parameters and a new sample is assigned to a model (class) according to maximum likelihood criterion.

4. Neural Networks (NNs), with one hidden layer. They use a multi-layer perceptron (MLP) neural network with back propagation algorithm and sigmoid transfer function.

5. Support Vector Machines (SVMs), with Radial Basis Kernel Function (RBF) to map the data points from input space to feature space.

To construct a multi-class SVM, they implement three different plots: One-Vs-All encoding scheme where the sample belongs to the class with the highest distance to others. In the second method the distances are fed into a MLP neural network. They also propose a Multi-Layer SVM (ML-SVM) depicted in Fig. 2.3 for extending the binary SVMs to a multi-category problem. They rank their extracted features according to Linear Discriminant Analysis (LDA) in order to choose the best subset of features. Their best classification result using just acoustic features is 81.29% achieved by ML-SVMs. They use standard Hidden-Markov-Model-based automatic speech recognition (ASR) engine with zero-grams as language model. They suggest using a MLP for fusion of the obtained acoustic and linguistic information, where the input feature vector consists of features derived by acoustic and linguistic analysis and 7 output neurons provide the final emotion probabilities by a softmax function.

Lin and Wei [11] in their research on AER use only acoustic information of speech signals. The emotional speech database used in this study is the Danish Emotional Speech

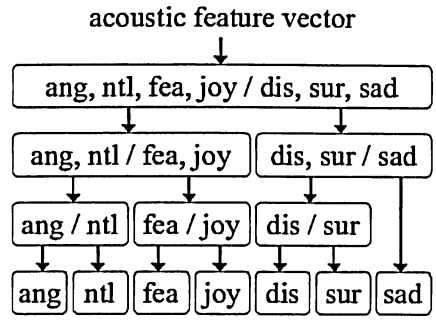


**Figure 2.3:** Optimal alignment of the emotions using ML-SVMs suggested in [10]

**Table 2.2:** Voice and emotion (Based on Murray and Amott, 1993)

| | Fear | Anger | Sadness | Happiness | Disgust |
|---|---|---|---|---|---|
| Speech rate | Much faster | Slightly faster | Slightly lower | Faster or slower | Very much slower |
| Pitch average | Very much higher | Very much higher | Slightly slower | Much higher | Very much slower |
| Pitch range | Much wider | Much wider | Slightly narrower | Much wider | Slightly wider |
| Intensity | Normal | Higher | Lower | Higher | Lower |
| Voice quality | Irregular voicing | Breathy chest tone | Resonant | Breathy blaring | Grumbled chest tone |
| Pitch changes | Normal | Abrupt on stressed syllables | Downward inflections | Smooth upward inflections | Wide downward terminal inflections |
| Articulation | Precise | Tense | Slurring | Normal | Normal |

(DES) database [14], which includes expressions by two male and two female actors familiar with radio theater. The whole database is divided into four parts for the purpose of cross-validation. The speech sentences in their database are expressed in five basic emotional states: anger, happiness, neutral, sadness and surprise. Their experiment is performed in both gender-dependent and gender-independent cases. In this work, the five groups of short-term features that were extracted relate to fundamental frequency (F0), energy, the first four formant frequencies (F1 to F4), two Mel Frequency Cepstrum Coefficients (MFCC1 and MFCC2), and five Mel frequency sub-band energies (MBE1 to MBE5). The performances of three different classifiers are investigated in this work: Hidden Markov Model (HMM), SVM with RBF kernel function, and k-NN. In order to select the best subset of features, Sequential Forward Selection (SFS) method is adopted. SFS algorithm is initialized with the single best feature as determined by maximal correct classification rate criterion. When combined with the selected ones, subsequent features that have the maximal correct classification rate are added in turn. The selection of features stops when adding new ones fails to increase the overall correct classification rate or when the number of the selected features reaches a pre-set number. The recognition rates reported in [11] are 98.9% by the HMM classifier for female subjects, 100% for male subjects, and 98.5% for gender independent cases. When the SVM classifier and the proposed feature vector are employed, correct classification rates of 89.4%, 93.6% and 88.9% are obtained for male, female and gender independent cases respectively.

Petrushin in [12] makes use of acoustic features of the speech signals. The study deals with a corpus of 700 short utterances expressing five emotions: happiness, anger, sadness, fear, and normal (unemotional) state, which were portrayed by thirty subjects. The database is randomly divided into two parts, 70% for training phase and 30% for evaluation. Some statistics for fundamental frequency F0, energy, speaking rate, first three formants (F1, F2, and F3), and their bandwidths (BW1, BW2, and BW3) are calculated as their acoustic features. Three different approaches for classifying the five different emotional states are implemented:

1. k-NN

2. NNs, with Back propagation algorithm and sigmoid transfer function.

3. Ensembles of NN classifiers, where an ensemble consists of an odd number of neural network classifiers that have been trained on different subsets of the training set using the bootstrap aggregation and the cross-validated techniques. The ensemble makes decision based on the majority voting principle.

RELIEF-F algorithm is used in this work to reduce the dimensionality of the feature vector. The overall accuracy achieved in this research is 70% corresponding to ensembles of NN classifiers. In this experimental study the performance of people for recognizing other's as well as their own emotional state is evaluated in order to investigate how well people without special training can recognize emotions in speech and how well people can recognize their own emotions that they recorded 6-8 weeks earlier. As a real-world application to this research, a developed emotion recognition software for call centers is also addressed at the end. According to the author, this emotion recognition agent was created in order to analyze telephone quality speech signals and distinguish between two emotional states: "agitation" and "calm". The agent can be used as a part of a decision support system for prioritizing voice messages and assigning a proper human agent to response the message at call center environment.

Chuang and Wu in [13] present an approach to emotion recognition from acoustic and textual content of speech signals in order to classify the aforementioned six universal emotional states. Their experiments were performed on a collected drama corpus with 1085 sentences in 221 dialogues from the leading man and 101 Sentences in 213 dialogues from the leading woman. In their proposed approach, some statistical values of four basic acoustic features: pitch, energy, formant 1 (F1), and zero crossing rate (ZCR) are calculated. The most significant features are selected utilizing principle component analysis (PCA) method to form the acoustic feature vector.

They adopt the SVM classifier that classifies the input data in a space and produce a con-

tinuous probability for emotion recognition. Given the test sample $x'$, the probability that $x'$ belongs to class $c$ is $P(Class_c|x')$. According to them this value relates to three factors: distance between the testing input and the hyperplane, distance from class centroid to the hyperplane, and the classification confidence of the class, which is the number of samples correctly recognized as class $c$ over total number of samples in class $c$. In the text analysis module, emotion content of an input sentence is essentially represented by its word appearance. Two primary word types "emotional keywords" and "emotion modification words" are manually defined and used to extract emotion from the input sentence. All of the extracted emotional keywords and emotion modification words have their corresponding "emotion intensity values" and "emotion modification values", which are manually defined. For each input sentence, the emotion intensity values are averaged and triggered by the emotion modification values to give the current emotion output. A front-end speech recognizer is firstly used to convert the input speech signal into the textual data. To extract the appropriate emotional state from textual input, they assume that every input sentence includes emotional keywords and emotion modification words. The emotional keywords provide basic emotion description of the input sentence, and the emotion modification words enhance or decrease the emotional state. The final emotion output is the combination of the previous two modules. They report the average recognition of 76.4%, 65.4%, and 81.4% for acoustic features, textual content, and the integrated system, respectively.

### 2.1.3 Emotion Recognition Using Bimodal Approach

In humans' face-to-face communication several different channels and modalities are functioning and thus the communication is very flexible and robust. In fact, failure of one channel is recovered by another channel and a message in one channel can be explained by another channel. As a result, some researchers have adopted bimodal approach (i.e. using both facial expressions and voice intonation) in their emotion recognition systems with the aim of extending the capability and performance of the system compared to when only single modal works alone. Some researchers [64] have found out that some emotions (sadness and fear) are

auditory dominant, some emotions (happiness, surprise and anger) are visually dominant, while some (disgust) are mixed dominant.

In addition to the applicable requirements and relevant steps in each of the single modules in a bimodal system (as partly mentioned in the two previous sections), the main challenge in such a system is to decide how and at which level the fusion of the information derived from each of the modules should happen to get the final result.

Hoch *et al.* [15] present a person-dependent emotion recognition system by adopting an acoustic and a visual monomodal recognizer and combining the individual results on the "decision level". In the visual analysis module, OpenCV (Open Source Computer Vision Library) [16] is used to detect the face areas in a sequence of images and a set of 18 Gabor wavelet filters results in 88200 different magnitude coefficients. In the acoustic module some statistical parameters of prosodic features (i.e. pitch, power, formants, duration of voiced segments) form the acoustic feature vector. In both modules an SVM with linear kernel function is employed. The output is transferred into a probability distribution $(\rho_1, \rho_2, \rho_3)$ by a soft-max function. Both monomodal emotion recognition systems provide an output vector containing the individual confidence measurements (posterior probability) of the monomodal classification process. Their proposed fusion approach combines these two monomodal results to a multimodal decision using a weighted linear equation:

$$\rho_{fus,n} = \eta.\rho_{ac,n} + (1 - \eta).\rho_{vis,n}$$

where $\rho_{ac,n}$ is acoustic confidence measurements, $\rho_{vis,n}$ is the corresponding visual results, $\rho_{fus,n}$ is the merged final result, and $\eta \in [0, 1]$ is called linear fusion coefficient (LFC). An evaluation of the recorded examples in [15] yields an average recognition rate of 90.7% for the fusion approach. According to them, this adds up to a performance gain of nearly 4% compared to the best monomodal recognizer.

Chen, Huang, and Cook in [17] show that combining in "feature level" outperforms combining in "decision level". They adopt both visual and acoustic features for categorizing the six universal emotions. In the facial expression analysis module, first they apply a facial feature tracking algorithm to track eyes, eyebrows, furrows, and lips. After collecting

all possible features, they employ FACS (Facial Action Coding System) [18] to generate the facial feature vectors. Eight acoustic features are calculated in their work, which can be categorized as three types of information: pitch contour, intensity contour, and energy spectrum. For bimodal feature analysis they directly combine the vocal and visual features and then they are fed into a SVM. They achieved the classification rate of 82% which according to them is an increased performance compared with each single mode.

### 2.1.4 The Objective of This Thesis

Between the two major modalities the audio channel is used as the input channel to the AER system developed in this thesis. In the real-world applications data acquisition is easier and faster if we deal only with speech signals; we don't have to be worried about problems caused by changes in the illumination and angle of the images. Processing one-dimensional speech signals using signal processing techniques is faster than processing two-dimensional images, especially when tracking in the image sequences is involved. Also there are some applications (e.g. in telephone conversations) where there is no access to the visual information.

While some researchers have utilized both acoustic characteristics and textual content of an emotional spoken utterance [10][13], this work is conducted using only acoustic features of the speech signal. Although adding the information derived from textual content of an utterance may provide some clue to recognize the emotion of the speaker and improves the overall performance of the AER system in some cases, in general human's speech emotional state is too complicated to be perceived from language information. People can recognize each other's emotional state mostly from intonation and speaking rate, rather than the said words. Two sentences could have the same lexical meaning but different emotional information. In fact, dependency on language information decreases the generalization of system and even can be misleading in some cases.

As reviewed in the previous three sections, some researchers have developed speaker-dependent speech emotion recognition systems [10][15]. We think that speaker independency is one of the intrinsic characteristics of an AER system. When a system is person-dependent

the accuracy increases, but for each new subject the system has to be trained all over again and that is a major drawback. So in this thesis it is tried to reach a very satisfying accuracy with a person-independent system by choosing adequate acoustic features and an appropriate classifier. Gender dependency also can decrease the generalization of the system.

Considering all these facts, the objective of this thesis is to develop a speaker-independent, gender-independent Automatic Emotion Recognition (AER) system based on acoustic information of speech signals. Another important issue in an emotion recognition study is to provide a representative database. The next section addresses some of the existing common databases for emotion recognition and also the emotion corpus which is used in this thesis is explained in detail.

## 2.2 Emotion Corpus

During the past decade, research on AER has attracted the interest of an ever-growing community of researchers. Numerous systems achieving emotion recognition from visual or acoustic features have been developed. However, since the achieved results strongly depend on the used databases, it remains very difficult to compare the relative performances of the existing prototypes due to the lack of common databases and protocols.

In the past few years, the Cohn-Kanade facial database [19] imposed itself as the main benchmark database for facial expression recognition algorithms. It includes over 2000 image sequences from over 200 different subjects, expressing up to six different emotions. Some other facial expression databases are also employed. To name just a few, the Japanese Female Facial Expression (JAFFE) database [20] contains 213 images of 7 facial expressions, posed by 10 Japanese female models. The AR Face Database [21] is a collection of over 4000 high-resolution color images of faces with different facial expressions, illumination conditions, and occlusions. For emotion recognition systems based on speech a relatively large number of databases are employed most of which are in languages other than English. Among these databases the Danish Emotional Speech (DES) database by Engberg *et al.* [14] and the Hebrew emotion speech database by Amir *et al.* [23], are widely used. DES database

contains 4 speakers (2 male and 2 female) expressing 5 emotions (neutral, surprise, happiness, sadness and anger) and Hebrew emotion speech data base contains 30 subjects recalling an emotional event in which they participated in order to collect samples of basic emotions. For multimodal emotion recognition, there are very few number of databases available. A detailed analysis of the state of the art, coupled with an attempt to fill the need for a multimodal emotion database has recently been made by Douglas-Cowie *et al.* [24], whose approach focuses on the generation of genuine emotions. Although their result is interesting, a database containing the 6 archetypal emotions defined by Ekman *et al.* [18] is still needed, as most of the existing systems aim at recognizing this set of archetypal emotions.

The database used in this research is the one created in [25]. This audio-visual emotion database is a professional reference database for testing and evaluating video, audio or joint audio-visual emotion recognition algorithms aiming for recognizing 6 archetypal emotions defined by Ekman *et al.* (see Section 1.1.6)

As described in [25], the protocol of constructing the database is as follows: First, the subject is asked to listen carefully to a short story which provokes a particular emotion and to immerse himself/herself into the situation. Once he/she is ready, the subject may read, memorize and pronounce (one at the time) the five proposed utterances, which constitute five different reactions to the given situation. These reactions are presented in Fig. 2.4. The subjects are asked to put as much expressiveness as possible, producing a message that contains only the emotion to be elicited. In the post processing stage two human experts judged whether the reaction expressed the emotion in an unambiguous way. If this was the case, the sample was added to the database. If not, it was discarded. The final version of the database contains 42 subjects among which 81% are men, while the remaining 19% are women.

All the experiments were driven in English. Allowing the subject to react in its own language has a main drawback: the acoustic features largely depend on the language itself. To illustrate by an example, the speaking rate is typically higher for an Italian than for a French-speaking Swiss subject. Thus, to have acoustic features that depend only on the

| ANGER |
| --- |
| R1: What??? No, no, no, listen! I need this money! |
| R2: I don't care about your coffee! Please serve me! |
| R3: I can have you fired you know! |
| R4: Is your coffee more important than my money? |
| R5: You're getting paid to work, not drink coffee! |

| SADNESS |
| --- |
| R1: Life won't be the same now |
| R2: Oh no, tell me this is not true, please! |
| R3: Everything was so perfect! I just don't understand! |
| R4: I still loved him (her) |
| R5: He (she) was my life |

| FEAR |
| --- |
| R1: Oh my god, there is someone in the house! |
| R2: Someone is climbing up the stairs. |
| R3: Please don't kill me... |
| R4: I'm not alone! Go away! |
| R5: I have nothing to give you! Please don't hurt me! |

| HAPPINESS |
| --- |
| R1: That's great, I'm rich now!!! |
| R2: I won: this is great! I'm so happy!! |
| R3: Wahoo... This is so great. |
| R4: I'm so lucky! |
| R5: I'm so excited! |

| DISGUST |
| --- |
| R1: That's horrible! I'll never eat noodles again. |
| R2: Something is moving inside my plate |
| R3: Aaaaah a cockroach!!! |
| R4: Eeeek, this is disgusting!!! |
| R5: That's gross! |

| SURPRISE |
| --- |
| R1: You have never told me that! |
| R2: I didn't expect that! |
| R3: Wahoo, I would never have believed this! |
| R4: I never saw that coming! |
| R5: Oh my God, that's so weird! |

**Figure 2.4:** Reactions to elicit the six emotions

**Table 2.3:** Geographic distribution of the subjects who participated in the database

| Country | Number of Subjects | Country | Number of Subjects |
|---|---|---|---|
| Belgium | 9 | Cuba | 1 |
| Turkey | 7 | Slovakia | 1 |
| France | 7 | Brazil | 1 |
| Spain | 6 | U.S.A. | 1 |
| Greece | 4 | Croatia | 1 |
| Italy | 1 | Canada | 1 |
| Austria | 1 | Russia | 1 |

expressed emotion, we have to deal only with one language. However, the subjects come from 14 different nationalities listed in Table 2.3, so they talk in different accents.

Regarding technical aspects, the video sequences were processed using a 720 × 576 Microsoft AVI format. The frame rate is equal to 25 frames per second, while pixel aspect ration is D1/DV PAL (1.067). The video was compressed using a DivX 5.0.5 Codec, to ensure easy portability. The audio sample rate is 48000 Hz, in an uncompressed stereo 16-bit format.

Eventually, the database consists of a total of 1287 video sequences, which is a large number of samples to train and to test a system compared with other databases. Out of 1287 video clips, 296 sequences are recordings from women (23%) and 991 sequences recordings from men (77%).

# Chapter 3

# Emotion Recognition Using LS-SVMs

HUMANS are capable of detecting emotions by listening to each other's voice. Although different languages and accents are used worldwide and the way people express their emotions in speech varies according to their cultural background, personality, age, gender, etc., in most of the cases we can perceive other peoples's feelings. For more natural HCI applications, we need to give computers the same capabilities as humans'. As one of the major indicators of humans' affective state, speech plays an important role in machine recognition of human emotion.

To build a more generic emotion recognition system, the extraction of features that can truly represent the universal characteristics of the intended emotion is a real challenge. A good reference model is the human hearing system. Previous works have explored several different types of features. Since prosody is believed to be the main indicator of a speaker's emotional state [46], most researchers adopt prosodic features. However, Mel Frequency Cepstral Coeffcient (MFCC) and formant frequency are also widely used in speech recognition and some of the other speech processing applications. In this work a set of novel cepstral features are proposed most of which are being used for the first time in this application.
As explained before, two indispensable characteristics of an AER system are user independency and gender independency. So the AER system developed in this contribution possesses both features. The proposed system is also independent of textual information of the speech signals.

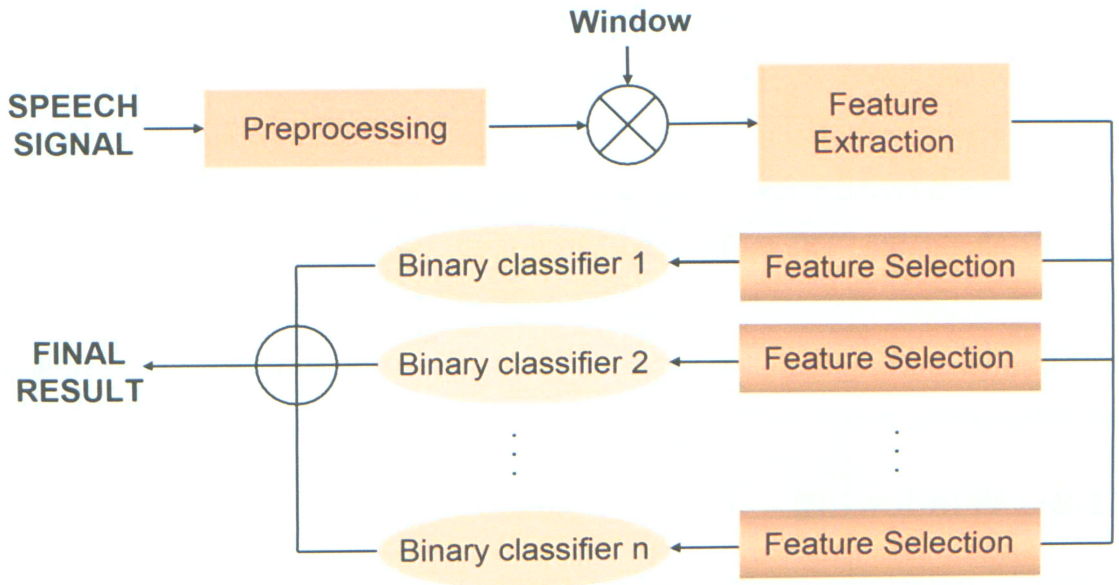


**Figure 3.1:** The structure of the speech emotion recognition system for frame approach.

In this chapter the proposed frame-approach AER system is explicated and the corresponding steps are expounded. The achieved results are reported and discussed.

## 3.1 Frame-approach AER System

The block diagram of the proposed speech emotion recognition system, when all features are extracted from each frame, is depicted in Fig. 3.1. It consists of five components. The preprocessing step performs noise reduction and silence elimination. Then, the preprocessed signal is passed through a windowing process to segment the original signals into short-time speech frames. Acoustic features are then extracted separately based on short time spectral analysis. Different steps of this procedure are elaborated in the following sections.

### 3.1.1 Preprocessing

In the preprocessing stage first each signal is de-noised by soft-thresholding the detail coefficients obtained by three levels of wavelet decomposition using db3 wavelet function. Also since the silent parts of the signals do not carry any useful information, those parts includ-

ing the leading and trailing edges are eliminated by thresholding the energy of the small intervals of the signal. In detail, a basic rectangular window with length of 23ms and zero percent frequency overlap is used to divide signals into adjacent frames; the energy content of the frames is calculated and thresholded. So the silent intervals with almost zero energy value are eliminated. The threshold values are set empirically. Also two-channel signals are converted to mono-channel by getting the average of the two channels.

Figs. 3.2 and 3.3 show the result of the preprocessing stage after de-noising and silent part elimination, respectively.

### 3.1.2 Windowing

In order to extract features from the emotional speech signal, we perform spectral analysis. The spectral analysis method is only reliable when the signal is stationary, i.e. the statistical characteristics of a signal are invariant with respect to time. Speech signals like any other audio signal are highly non-stationary, however; vocal tract can be considered stable over a very short period of time, typically around 10-30ms. A signal $x(n)$ is divided into a succession of windowed sequences $x_t(n)$, called frames. These speech frames can then be processed individually.

$$x_t(n) = w(n)x'_t(n) \quad n = 0, \ldots, N-1, \quad t = 0, \ldots, T-1 \tag{3.1}$$

where $w(n)$ is the impulse response of the window, $N$ is the size of the window, $T$ is the number of frames, and $x'_t(n)$ is the frame before applying the window function.
In this thesis, a Hamming window with length of 23ms and 50% frequency overlap is used. The impulse response of a Hamming window (Fig. 3.4) $w(n)$ is a raised cosine impulse [47]:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) \quad n = 0, \ldots, N-1. \tag{3.2}$$

Compared with the rectangular window shape, Hamming window has the advantage of decreasing the leakage effect and to smooth the transition and eliminate the possible gaps between blocks, overlapping windows are usually employed.

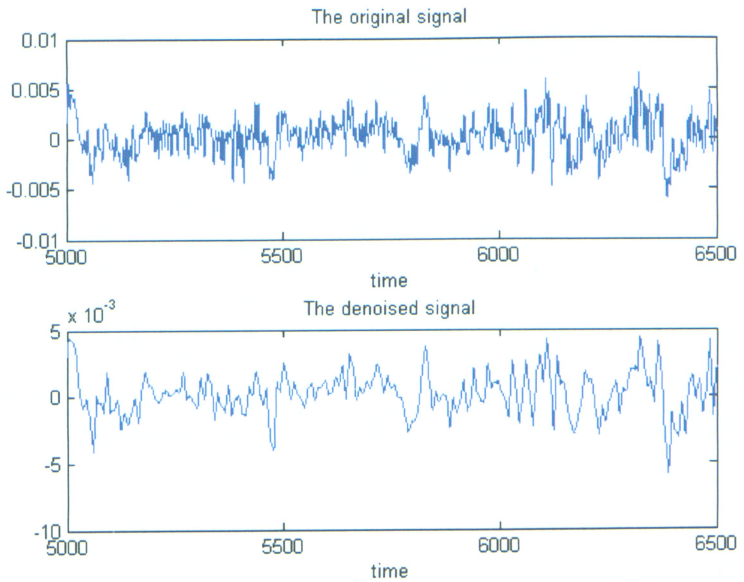


**Figure 3.2:** Signal after de-noising

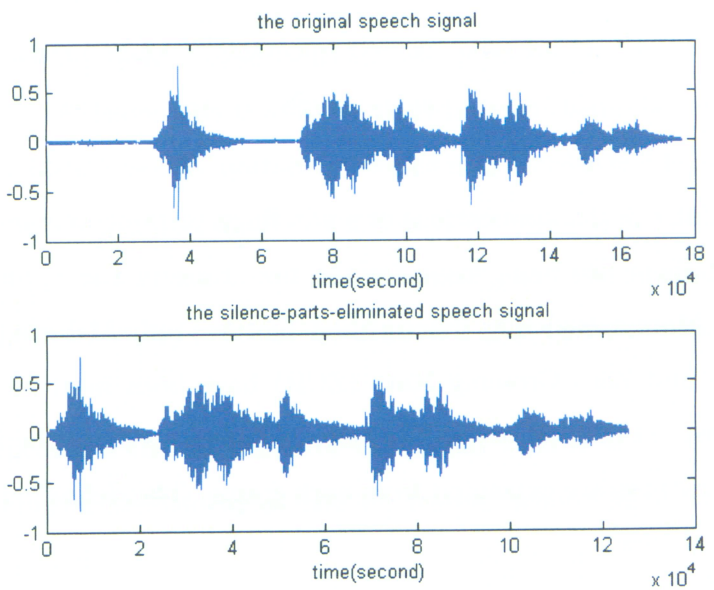


**Figure 3.3:** Signal after eliminating the silence parts

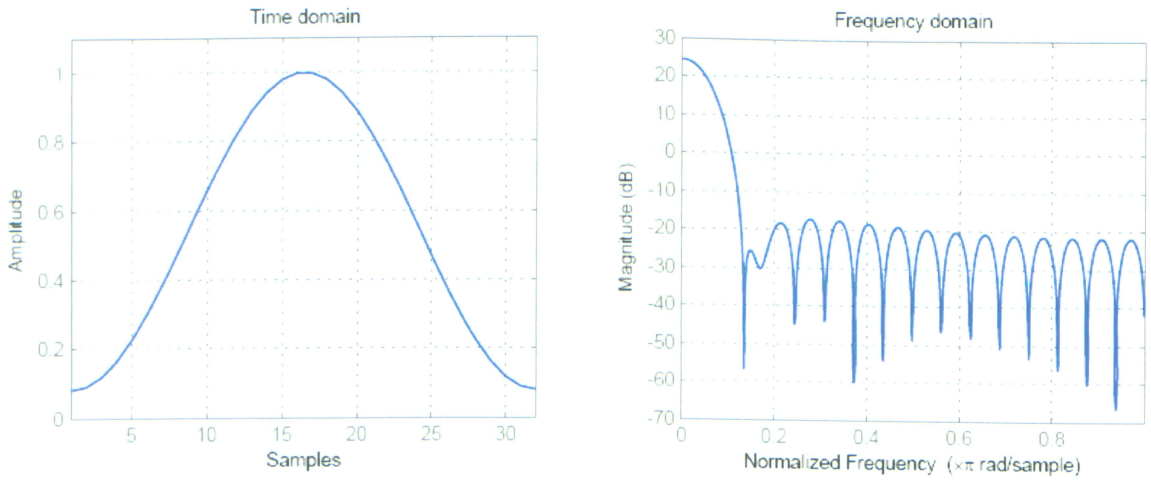


**Figure 3.4:** Hamming window (32 points). Left: time domain, Right: frequency domain

### 3.1.3 Feature Extraction

A set of novel acoustic features is proposed in this thesis. Most of the features used in previous works are prosodic features and their statistical characteristics [11][12][25][26][27]. Fig. 3.5 shows the list of features used in this contribution. These features have been previously utilized successfully in the applications of audio fingerprinting and speaker verification [28][29], but most of them are being used for the first time in the application of speech emotion recognition. More specifically, among these features only Mel Frequency Cepstrum Coefficients (MFCC) and Zero Crossing Rate (ZCR) have been used for speech emotion recognition in the past [11][13][30], while the rest are being used for the first time in this application. All the features are extracted from each frame. The definition of features used in this work are given below.

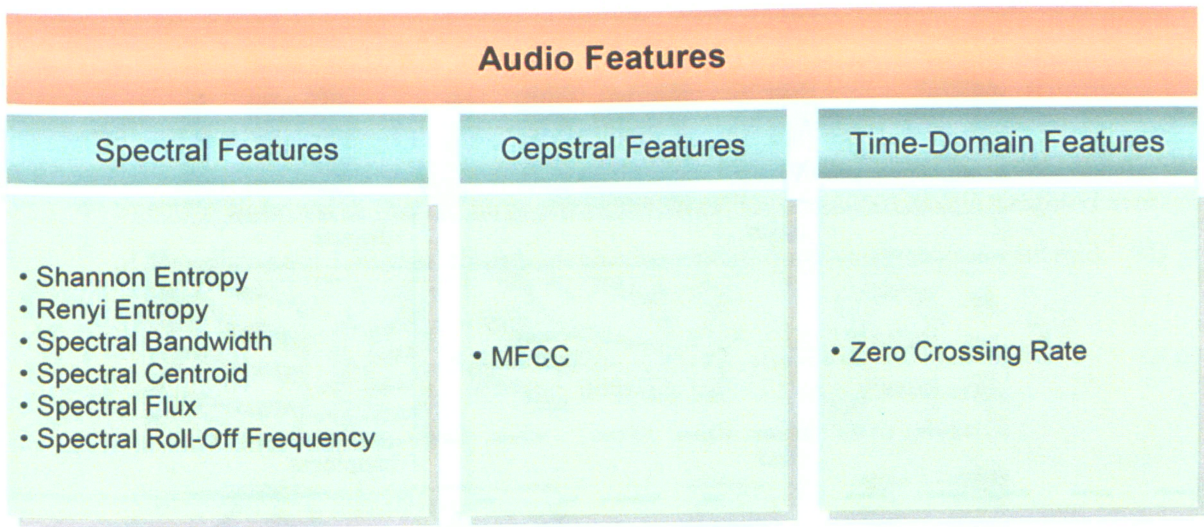


**Figure 3.5:** List of acoustic features used for speech emotion recognition.

### Spectral Features

In order to perform spectral analysis, the speech signals need to be transformed to the frequency domain. This is done by discrete Fourier transform. Figure 3.6 shows spectrograms and associated waveforms of the six emotions, as produced by one of the experimental subjects. On the spectrogram, time is represented along the horizontal axis, whereas frequency is plotted along the vertical axis. For a given spectrogram $S$, the strength of a given frequency component $f$ at a given time $t$ in the speech signal is represented by the darkness of the corresponding point $S(t, f)$. It can be observed that each emotion class exhibits different patterns.

Let $s_i(n)$ represents the $i^{th}$ frame of the signal with $n = 1, \ldots, N$. Let $F_i = f_i(u), u \in (0, M)$, be the Fourier transform of the $i^{th}$ frame, where $M$ is the index of the highest frequency band.

1. Shannon Entropy (SE) : The Shannon entropy of a signal is a measure of its spectral distribution. Shannon entropy is defined as

$$SE_i = \sum_{u=0}^{M} \mid f_i(u) \mid \log_2 \mid f_i(u) \mid \tag{3.3}$$

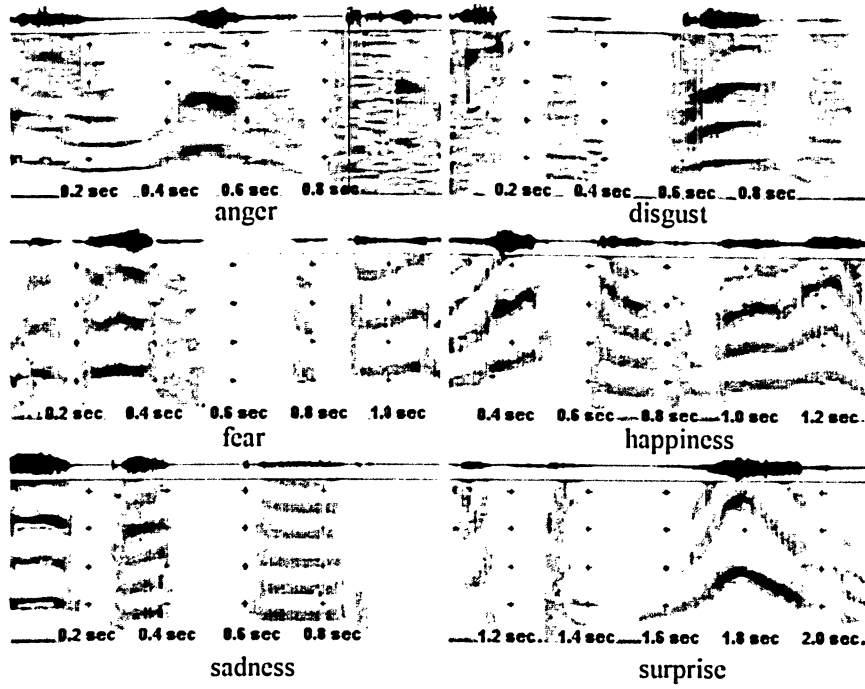


**Figure 3.6:** The spectrograms associated with different emotions (adopted from [76])

2. Renyi Entropy (RE): The Renyi entropy of a signal is also a measure of its spectral distribution. Renyi entropy is defined as

$$RE_i = \frac{1}{1-r}\log_2\left(\sum_{u=0}^{M} \mid f_i(u) \mid^r\right) \tag{3.4}$$

3. Spectral Centroid (SC): The spectral centroid is the center of gravity of the magnitude spectrum of the STFT and is a measure of spectral shape and "brightness" of the spectrum. SC is defined as

$$SC_i = \frac{\sum_{u=0}^{M} u. \mid f_i(u) \mid^2}{\sum_{u=0}^{M} \mid f_i(u) \mid} \tag{3.5}$$

4. Spectral Bandwidth (SB): The spectral bandwidth is measured as the weighted average of the distances between the spectral components and the spectral centroid. SB is defined as

$$SB_i = \frac{\sum_{u=0}^{M}(u - SC_i)^2 . \mid f_i(u) \mid^2}{\sum_{u=0}^{M} \mid f_i(u) \mid^2} \tag{3.6}$$

5. Spectral Flux (SF): The spectral flux is defined as

$$SF_i = \sum_{u=0}^{M} ||| f_{i+1}(u) \mid - \mid f_i(u) ||| \tag{3.7}$$

6. Spectral roll-off Frequency (SRF): The spectral roll-off frequency is defined as

$$SRF_i = max\left(h \middle| \sum_{u=0}^{h} f_i(u) < TH. \sum_{u=0}^{M} f_i(u)\right) \tag{3.8}$$

where $TH$ is a threshold between 0 and 1. A threshold value of 0.7 is used in this work.

### Cepstral Features

Cepstral based features are widely used in speaker recognition applications. Cesptral coeffcients enable us to obtain information about vocal tract configuration.

Essentially, the speech system can be modeled with an input, a filter, and an output. The input to the speech system is the periodic oscillations for the vocal cords or air from the lungs, the output is the speech signal, and the vocal tract, mouth, and lips, acts as a time varying filter that modifies the input signal to produce speech or other sounds in general. Of course, the shape and thickness of the vocal tract is controlled by a group of muscles, and

the shape of the mouth cavity and lips are controlled by the speaker; factors which depend on the anatomical structure of the speaker as well as the way the speaker learns to speak.

Modeling the entire speech system as a time varying excitation and a time varying filter, the speech signal ($s(t)$) is given by

$$s_{voiced}(t) = x(t) * h(t) \tag{3.9}$$

$$s_{unvoiced}(t) = n(t) * h(t) \tag{3.10}$$

where $x(t)$ is a periodic excitation, $n(t)$ is white noise, and $h(t)$ is a time varying filter which constantly changes to produce different sounds. However, $h(t)$ can be considered stable over a period of few milliseconds (ms); typically a period of about 10-30ms is commonly used in literature [48][49]. This convenient short-time stationary behavior can be exploited to characterize the vocal tract configuration given by $h(t)$. This information can be easily extracted from the speech spectrum using well established deconvolution techniques.

The cepstrum operator is often found in literature under homomorphic deconvolution and therefore, it can separate the components of speech found in Equ. 3.9 and Equ. 3.10. This powerful tool then permits for separate analysis of the vocal tract configuration (given by the filter component $h(t)$). The cepstrum ($C(t)$) of the signal $s_{voiced}(t) = x(t) * h(t)$ is given by

$$Cesptrum\{s(t)\} = FFT^{-1}\{|\log FFT[s_{voiced}(t)]|\} \tag{3.11}$$

Although Mel-frequency cepstral coefficient (MFCC) is perhaps the most popular solution in the field of speech recognition, identification, etc., since the purpose of MFCC is to mimic the behavior of human ears by applying cepstral analysis and as our goal is to identify possible acoustic features that can contribute to the recognition of human emotion, we also investigate this type of feature. Calculating the MFCCs for a speech signal consists of preprocessing, windowing, followed by Fourier transform, Mel-scaling and inverse cosine transform for each time frame. Prior to inverse transform, the magnitude of the spectrum is made logarithmic. This logarithmic scale is a characteristic of the human hearing system.

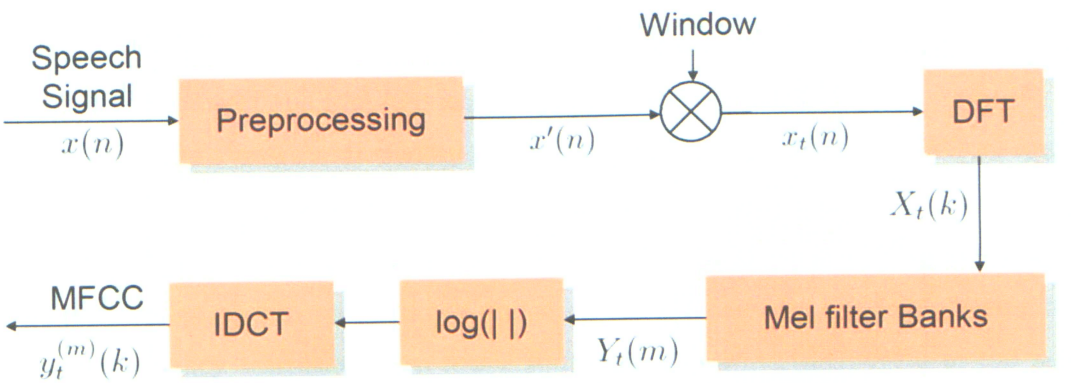


**Figure 3.7:** MFCC processing flow

Figure 3.7 shows the flow of the MFCC feature extraction procedure.

*Mel-scaled filter bank:* In order to simplify the spectrum without significant loss of data, the Fourier transformed signal is usually passed through a set of band-pass filters, which properly integrate a spectrum at defined frequency ranges. In a speech signal, most of the important and useful information is located at the lower frequency band. Mel-scale is the most widely used perceptual scale, which is designed to capture and emphasize the information in low frequency band. The filter bank is usually constructed of triangular-shaped filters with frequency overlap, so that the center frequency of a filter corresponds to the upper frequency of previous filter and lower frequency of next filter. The central frequency of each Mel filter bank (Fig. 3.8) is uniformly spaced before 1 kHz and it follows a logarithmic scale after 1 kHz. Furthermore, to emphasize the low frequency components, the filter magnitude is usually set to 1 at the low frequency band, while decreasing as the frequency increases.Usually, the range of the frequency covered by the filter bank lies between 20 Hz till half of the sampling frequency of the signal. Fig. 3.8 shows the diagram of an ideal Mel-scaled filter bank.

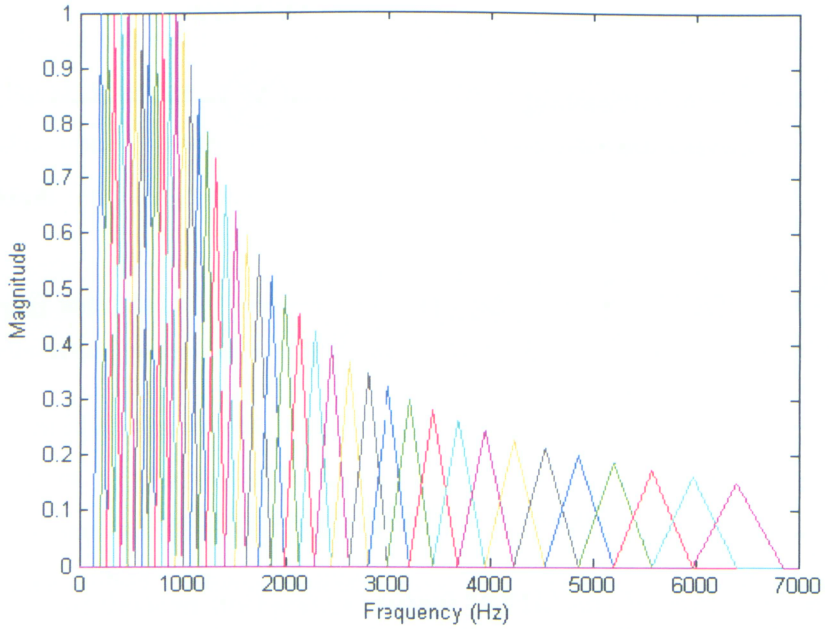


**Figure 3.8:** Mel-scaled filter bank design.

*Cepstral coefficients:* By using a Mel-scaled filter bank, the spectrum is smoothed in the same way it is in the human ear. The next step is to compute the logarithm of the square magnitude of the coefficients $Y_t(m)$. This reduces to simply computing the logarithm of the magnitude of the coefficients, because of the logarithm algebraic property which brings back the logarithm of a power to a multiplication by a scaling factor. By taking the log of the filter coefficients, the characteristics of the human auditory system can be simulated, because magnitude and logarithm processing are performed by the human ear as well. Furthermore, the magnitude operation discards the useless phase information, while a logarithm performs a dynamic compression, making feature extraction less sensitive to variations in dynamics [63].

MFCCs are the inverse discrete cosine transform of the logarithm of the magnitude of the filter bank output:

$$y_t^{(m)}(k) = \sum_{m=1}^{M} \log\{|Y_t(m)|\}.\cos(k(m - \frac{1}{2})\frac{\pi}{M}) \tag{3.12}$$

where $m$ is the index of the filter, and $M$ is the total number of the filters. $y_t^{(m)}(k)$, $m = 1, 2, \ldots, M$ represent the Mel-frequency cepstral coefficients.

By using the above mentioned techniques, we can calculate the MFCCs from each speech utterance. For each speech utterance, we calculate coefficients matrix of size $M \times N$, where $M$ is the number of coefficients, while $N$ denotes the total number of speech frames in an utterance where all processing is performed on 23ms frames with 50% of overlap to ensure smooth frequency transition from frame to frame. However, the lengths of the utterances are different, and thus the sizes of the coefficient matrix are different. In order to facilitate the classification, the features of each utterance that are mapped to the feature space should have the same length. It has been shown that the first several cepstral coeffcients from the cepstral domain represent the anatomical structure of a speaker's vocal tract. In speech recognition, the common number of used coefficients is between nine and thirteen [66][67]. In this work, we take the first thirteen coefficients. We then calculate the mean, and standard deviation of each order of $y_t^{(m)}(k)$, $m = 1, 2, \ldots, M$ as the extracted features.

### Time Domain Features

In time domain, only the zero crossing rate (ZCR) is considered. ZCR is a correlate of the spectral centroid. It is defined as the number of time-domain zero-crossings within the processing frame [69].

$$ZCR_i = \frac{f_s}{N}\left(\sum_{i=1}^{N-1} |sign(s_i(n)) - sign(s_i(n-1))|\right) \tag{3.13}$$

Let $X_i$ be the set of features extracted for the $i^{th}$ frame. We then have a sequence of feature vectors ($X$) for each signal. In order to represent only one feature vector for each utterance, mean and standard deviation of each variable is computed (see Section 4.1 for more detail).

$$X_i = [SE_i, RE_i, SC_i, SB_i, SF_i, SRF_i, MFCC_i^1, \ldots, MFCC_i^{13}, ZCR_i]^T$$

$$X = [X_1, X_2, \ldots, X_n]$$

where $n$ is the number of frames. Finally the feature matrix $X$ is mean subtracted and component wise variance normalized to get a normalized feature matrix. Having six spectral features, thirteen cepstral features, and one time-domain feature, and by computing mean and standard deviation of each of these variables, dimensionality of the final feature vectors comes to 40.

### 3.1.4 Feature Selection

The task of selecting relevant features in a classification task can be viewed as one of the most fundamental problems in the field of machine learning. The performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. By selecting the most relevant subset from the original feature set, we can increase the performance of the classifier and on the other hand decrease the computational complexity. Thus, selecting relevant features for the construction of classifiers has received a great deal of attention.

Feature Selection (FS) methods in Data Mining (DM) and Data Analysis problems aim at selecting a subset of the variables, or features, that describe the data in order to obtain a more essential and compact representation of the available information. The selected subset has to be small in size and must retain the information that is most useful for the specific application. The role of Feature Selection is particularly important when computationally expensive DM tools are used, or when the data collection process is difficult or costly.

In essence, the reduction of the original feature set to a smaller one preserving the relevant information while discarding the redundant one is referred to as feature selection. In many cases FS can be looked at as an independent task in the DM process, that pre-processes the data before they are treated by a DM method, which often may fail or have significant computational problems in treating data set with a large number of features directly [33].

The main benefits in using FS in DM may thus be outlined as follows:

- reduction in the amount of information needed to train a DM algorithm;
- better quality of the rules learned from data;

- easier acquisition and storage of the information related to a smaller number of "useful" features;

Many methods have been proposed in the literature for feature selection. The most common methods are explained in the following sections.

### General Feature Selection Method

In supervised learning, FS is often viewed as a search problem in a space of feature subsets and is based on four main steps, as follows [33]:

1. generation procedure;
2. evaluation function;
3. stopping criterion;
4. validation procedure.

The *generation procedure* is in charge of generating the subsets of features to be evaluated. From the computational standpoint, the number of possible subsets from a set of $N$ features is $2^N$. The generation starts with an empty set, and then adds a new feature at each iteration (*forward strategy*). Alternatively, it may start from the complete set of features removing one at each step (*backward strategy*). Finally, some methods propose to start from a randomly generated subset to which forward or backward strategy is applied.

The *evaluation function* is used to measure the quality of a subset. Such value is then confronted with the best available value obtained, and the latter is updated if appropriate. More specifically, the evaluation function measures the classification power of a single feature or of a subset of the features. Different types of evaluation functions can be used.

The *stopping criterion* is needed to avoid time consuming exhaustive search of the solution space without a significant improvement in the evaluation function. The search may be stopped if a given number of attributes has been reached, or when the improvement obtained by the new subset is not relevant.

Finally, the *validation procedure* measures the quality of the selected subset. This is typically accomplished by running the DM algorithm by using only the selected features on additional data.

According to the type of evaluation function adopted, FS methods are divided into two main groups: *filter methods* and *wrapper methods*[33]. In the former, the evaluation function is independent from the DM algorithm that is to be applied. In the latter, the DM algorithm is, to a certain extent, the essence of the evaluation function: each candidate subset is tested by using the DM algorithm and then evaluated on the basis of its performance. Wrapper methods are widely recognized as a superior alternative in supervised learning problems since can provide better results in terms of final accuracy [33].

Implementing a wrapper is a straightforward task in supervised learning, since there is always some external validation measure available. It is assumed that the goal of clustering is to optimize some objective function which helps to obtain good clusters and use this function to estimate the quality of different feature subsets.

Filter approach presents several weak points, amongst which are:

- They usually do not deal appropriately with noisy data;
- They often leave the choice amongst a number of "good" subsets to the user;
- In most methods the user is asked to specify the dimension of the final set of features, or to define a threshold value of some sort that drives the stopping condition of the algorithm;
- Some methods pose some constraints on the format of the data (e.g. they may require all data to be in binary format), introducing potential noise and furthermore increasing the number of features to start from.

However, it is faster than wrapper approach.
A drawback of wrapper method is that they are expensive from the computational standpoint.

### Feature Selection by Combining Features

*LDA-based feature selection:* Another approach to cope with the problem of excessive dimensionality is to reduce the dimensionality by combining features. Linear combinations are practically attractive because they are simple to compute and analytically tractable. In effect, linear methods project high-dimensional data onto a lower dimensional space. The classical approach to supervised linear dimensionality reduction is based on Linear Discriminant Analysis (LDA)[52]. This approach defines the optimal transformation matrix to be the one that maximizes the between-class covariance matrix and minimizes the within-class covariance matrix (Fisher criterion). There are two drawbacks with LDA-based methods: first the number of linear bases is limited by the number of classes and second the bases are not orthogonal in general [53].

*PCA-based feature selection:* Another method to find effective linear transformation is Principal Component Analysis (PCA), which seeks a projection that best represents the data in a least-squares sense. PCA is a multivariate procedure which rotates the data such that maximum variabilities are projected onto the axes. Essentially, a set of correlated variables are transformed into a set of uncorrelated variables which are ordered by reducing variability. PCA approach to feature selection has two drawbacks. The first is that it is based on variance of the features only and does not take the class labels into account, and the second is that there is no guarantee that selected feature are the most useful variables. In other words, PCA gives high weights to features with higher variabilities irrespective of whether they are useful for classification or not. This may give rise to the situation where the chosen principal component corresponds to the attribute with the highest variability but having no discriminating power [54].

Considering the advantages and disadvantages of the mentioned methods, in this thesis the Sequential Forward Selection (SFS) method is applied for each single binary classifier in the system in order to select the most efficient subset of features. The algorithm starts with zero feature selected and at each step the variable which increases the performance of the classifier the most is added to the feature subset. As a result, selection of the subset which

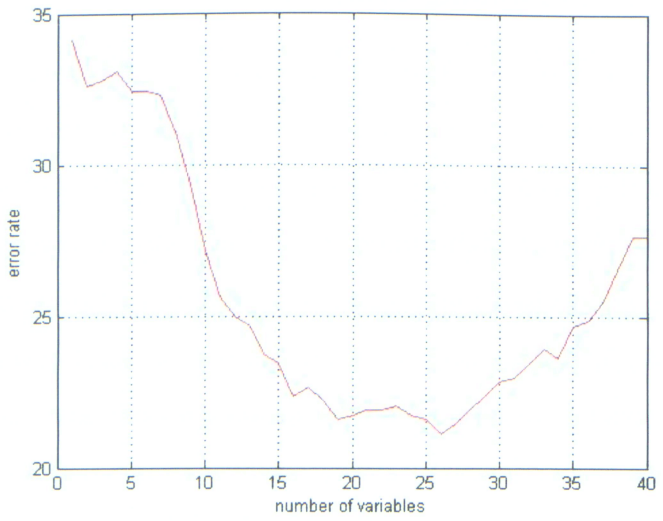


**Figure 3.9:** The performance of a binary LS-SVM by adding a new feature at each iteration of SFS algorithm

yields the best result is guaranteed. Fig. 3.9 illustrates the concept. In most researches one identical subset of features is used for all the classes to be separated. In this work, feature selection procedure is applied separately to each binary LS-SVM, so the variables which give the best result for different emotions can be captured.

### 3.1.5 Classification

The recognition of human emotion is essentially a pattern recognition problem where we want to categorize the emotional content of an utterance. Selection of a proper classifier has a significant impact on the overall result. There is no classifier referred in the literature as an optimal classifier. Therefore almost all the researchers try different classification methods to find a classifier which works well with their problem. In fact, choosing a proper classification method is highly related to the geometric distribution of the data points.

Machine Learning in general and the classification methods used in this thesis are elaborated in Section 4.12. In this research Least Squares Support Vector Machine (LS-SVM) explained in Section 3.3.2 is adopted in order to classify six categories of emotion. Since we are dealing with a multi-class classification problem, we need a method to extend our

binary SVMs to a multi-category problem. There are different strategies proposed to build a multi-class SVM. In this thesis the results achieved by one-against-all, fuzzy one-against-all, pairwise, and fuzzy pairwise [55] mentioned in Section 3.3.2 are compared, in order to achieve the best result.

For the purpose of comparative study, a linear classifier (described in Section 3.3.1) is also applied to examine if the data samples are linearly separable.

### 3.1.6 Implemented Results

Our database consists of 1287 instances of utterance. In the first part of our study for frame approach, 50% of data samples was used for the training phase exclusively and the remaining 50% for evaluating the trained classifiers (the division is done in random).

All the binary LS-SVM classifiers were trained using RBF kernel function defined in Equ. 3.34 with different optimal regularization and kernel parameters. Kernel parameters ($\sigma$) and regularization parameters were set empirically.The linear classifiers were trained using the gradient descent algorithm (see Appendix B). The initial weight vector and the step size in the algorithm were both set empirically.

Table 3.1 shows the final results. The abbreviation "FS" in the table means Feature Selection. As it is shown the best performance (81.3%) belongs to fuzzy pairwise LS-SVM using the features selected by SFS algorithm. In fact selecting pairwise method over One-Vs-All is a trade-off between accuracy and computational complexity where the improvement of accuracy is much more significant. The achieved result by the linear classifier is not an acceptable accuracy and we can come to the conclusion that the geometric distribution of the data samples in our experiment is not linearly separable.

Table 3.2 shows the confusion matrix for LS-SVM classifier using fuzzy pairwise method and SFS algorithm. The abbreviations in this table stand for the six different emotions: *anger, fear, disgust, happiness, sadness,* and *surprise.* We can deduce from Table 3.2 that the most difficult emotion to recognize in our experiment is surprise and the easiest ones are sadness and happiness.

**Table 3.1:** Final recognition results

| Classification method | Recognition Rate |
|---|---|
| One-Vs-All LS-SVM | 44.9% |
| fuzzy One-Vs-All LS-SVM | 53.6% |
| Pairwise LS-SVM | 74.5% |
| fuzzy Pairwise LS-SVM | 78.4% |
| fuzzy Pairwise LS-SVM, FS | **81.3%** |
| fuzzy pairwise LDA | 37.7% |

**Table 3.2:** Confusion matrix for fuzzy-pairwise LS-SVM with feature selection

| | Recognized Emotions (%) | | | | | |
|---|---|---|---|---|---|---|
| | Ang | Fea | Dis | Hap | Sad | Sur |
| Ang | **83.3** | 0 | 2.7 | 6.4 | 2.7 | 4.6 |
| Fea | 1.8 | **71.9** | 7.4 | 1.8 | 13 | 3.7 |
| Dis | 4.6 | 5.5 | **79.6** | 0 | 3.7 | 6.4 |
| Hap | 1.8 | 1.8 | 0 | **92.4** | 1.8 | 1.8 |
| Sad | 0 | 6.1 | 0.9 | 0 | **90.5** | 2.3 |
| Sur | 11.1 | 9.2 | 5.5 | 4.6 | 13.8 | **55.5** |

## 3.2 Machine Learning

Machine learning methodology is an artificial intelligence approach to establish and train a model to recognize the pattern or underlying mapping of a system based on a set of training examples consisting of input and output patterns. The construction of machines capable of learning from experience has, for a long time, been the object of philosophical and technical debate. The technical aspect of the debate has received an enormous impetus from the advent of electronic computers. They have demonstrated that machines can display a significant level of learning ability, though the boundaries of this ability are far from being clearly defined [51].

In other words, Machine Learning (ML) is the study of methods for programming computers to learn. Computers are applied to a wide range of tasks, and for some of these it is relatively easy for programmers to design and implement the necessary software. However, there are many tasks for which this can be difficult or impossible. These tasks can be divided

into four general categories.

First, there are problems for which there exist no human experts. For example, in modern automated manufacturing facilities, there is a need to predict machine failures before they occur by analyzing sensor readings. Because the machines are new, there are no human experts who can be interviewed by a programmer to provide the knowledge necessary to build a computer system. A ML system can study recorded data and subsequent machine failures and learn prediction rules.

Second, there are problems where human experts exist, but they are unable to explain their expertise. This is the case in many perceptual tasks, such as speech recognition, hand-writing recognition, and natural language understanding. Virtually all humans exhibit expert-level abilities on these tasks, but none of them can describe the detailed steps that they follow as they perform them. Fortunately, humans can provide machines with examples of the inputs and correct outputs for these tasks, so ML algorithms can learn to map the inputs to the outputs.

Third, there are problems where the underlying phenomena are changing rapidly. In finance, for example, people would like to predict the future behavior of the stock market, consumer purchases, or exchange rates. These behaviors change frequently, so that even if a programmer could construct a good predictive computer program, it would need to be rewritten frequently. A learning program can relieve the programmer of this burden by constantly modifying and tuning a set of learned prediction rules.

Fourth, there are applications that need to be customized for each computer user separately. Consider, for example, a program to filter unwanted electronic mail messages. Different users will need different filters. It is unreasonable to expect each user to program his or her own rules, and it is infeasible to provide every user with a software engineer to keep the rules up-to-date. A ML system can learn which mail messages the user rejects and maintain the filtering rules automatically.

ML addresses many of the same research questions as the fields of statistics, DM, and cognition, but with differences in emphasis. Statistics focuses on understanding the phenom-

ena that have generated the data, often with the goal of testing different hypotheses about those phenomena. DM seeks to find patterns in the data that are understandable by people. Cognitive studies of human learning aspire to understand the mechanisms underlying the various learning behaviors exhibited by people (concept learning, skill acquisition, strategy change, etc.). In contrast, ML is primarily concerned with the accuracy and effectiveness of the resulting computer system.

There are two phases in ML algorithms: "learning" or "training" the system with known data and "testing" where the system performance is tested with new data.

### 3.2.1 Learning

When computers are applied to solve a practical problem, it is usually the case that the method of deriving the required output from a set of inputs can be described explicitly. The task of system designer and eventually the programmer implementing the specifications will be to translate that method into a sequence of instructions which the computer will follow to achieve the desired effect. As computers are applied to solve more complex problems, however, situations can arise in which there is no known method for computing the desired output from a set of inputs, or where that computation may be very expensive. An example of this type of situations might be the handwriting recognition problem. These tasks cannot be solved by traditional programming approach since the system designer cannot precisely specify the method by which the correct output can be computed from the input data. An alternative strategy for solving this type of problems is for the computer to *learn* the input/output functionality from the examples. The approach of using examples to synthesize programs is known as the *learning methodology.* When the underlying function from inputs to outputs exists, it is referred to as the target function. The estimate of the target function which is learnt or output by the learning algorithm is known as the solution of the learning problem. In the case of classification this function is referred to as the *decision function.*

In the broadest sense, any method that incorporates information from training samples in the design of a classifier employs learning. Because nearly all practical or interesting pattern

recognition problems are so hard that we cannot guess the classification decision ahead of time, we shall spend the great majority of time considering learning. Learning tasks can be classified along different dimensions. One important dimension is the distinction between supervised (empirical) and unsupervised learning.

#### Supervised learning

In supervised learning, a teacher provides a category label or cost for each pattern in a training set, and we seek to reduce the sum of the costs for these patterns [50].

#### Unsupervised Learning

In unsupervised learning or clustering there is no explicit teacher, and the system forms clusters or "natural groupings" of the input patterns. "Natural" is always defined explicitly or implicitly in the clustering system itself, and given a particular set of patterns or cost function, different clustering algorithms lead to different clusters. Often the user will set the hypothesized number of different clusters ahead of time [50].

### 3.2.2 Testing

Once we have chosen a model, we have to estimate its performance. The training set error rate can be highly misleading and is usually an overoptimistic estimate of performance. Inaccuracies are due to the over-fitting of a learning machine to the data. In fact we want to know how well the model that has just learned from some training data is going to perform on future as-yet-unseen data (generalization). There are at least two reasons for wanting to know the generalization rate of the classifier on a given problem. One is to see if the classifier performs well enough to be useful; another is to compare its performance with that of a competing design.

Some of the most common validation methods are:

### Random Sub-sampling

Random sub-sampling performs $K$ data splits of the data set. Each split randomly selects a (fixed) number of examples without replacement. For each data split we retrain the classifier from scratch with the training examples and estimate the error rate, $E_i$, with the test examples. The true error estimate is obtained as the average of the separate estimates $E_i$.

### K-fold Cross Validation

This technique creates a K-fold partition of the data set. For each of $K$ experiments, it uses $K-1$ folds for training and the remaining one for testing. So the classifier is trained $K$ times, each time with a different set held out as a validation set. The estimated performance is the mean of these $K$ errors. The advantage of $K$-Fold cross validation is that all the examples in the data set are eventually used for training and testing.

### Leave-One-Out Cross Validation

Leave-one-out is the degenerate case of $K$-Fold cross validation, where $K$ is chosen as the total number of examples $N$. Each resulting classifier is tested on the single deleted point, and the estimate of accuracy is then simply the mean of these leave-one-out accuracies. Here the computational complexity may be very high, especially for large $N$.

## 3.3 Machine Learning Algorithms for Classification

The task of a classifier is to use the feature vector to assign the object to the proper category [50].

There are many different classifiers with different approaches, different cost functions and different algorithms. Considering the type of our data and the amount of information we have about it, we can decide on a more proper classifier; however, it is hard to find a perfect classifier unless we compare the result of several different algorithms. Classifications can be categorized as: supervised or unsupervised, and the classifiers as: linear or non-linear

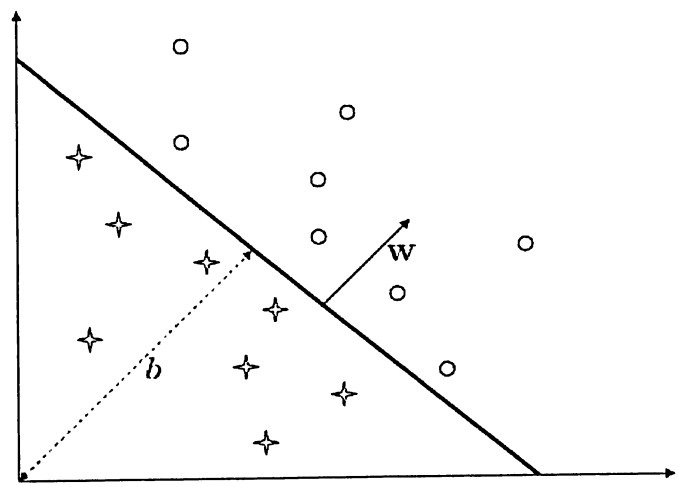


**Figure 3.10:** A linear separating hyperplane $(\mathbf{w}, b)$ for a two dimensional data set

and probabilistic or deterministic. Some of the existing classification methods are explained below.

### 3.3.1 Linear Discriminant Function

A discriminant function that is a linear combination of the components of $\mathbf{x}$ can be written as

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{3.14}$$

where $\mathbf{w}$ is the weight vector and $w_0$ the bias (Fig. 3.10)

A two-category linear classifier implements the following decision rule: Decide $w_1$ if $g(\mathbf{x}) > 0$ and $w_2$ if $g(\mathbf{x}) < 0$. Thus, $\mathbf{x}$ is assigned to $w_1$ if the inner product $\mathbf{w^T x}$ exceeds the threshold $w_0$ and $w_1$ otherwise. If $g(\mathbf{x}) = 0$, $\mathbf{x}$ can ordinarily be assigned to either class. Figure 3.11 shows the general structure of a linear classifier. The equation $g(x) = 0$ defines the decision surface that separates points assigned to $w_1$ from points assigned to $w_2$. When $g(x)$ is linear, this decision surface is a hyperplane. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are both on the decision

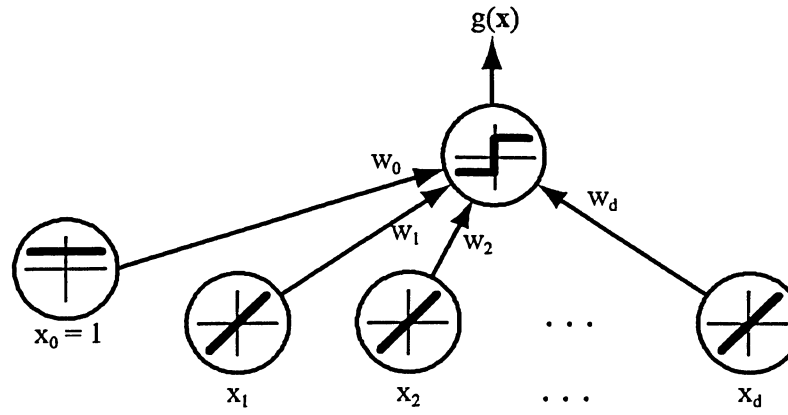


**Figure 3.11:** A simple linear classifier having $d$ input units

surface, then

$$\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0 \tag{3.15}$$

and this shows that $\mathbf{w}$ is normal to any vector lying in the hyperplane.

In general, linear discriminant function divides the feature space by a hyperplane decision surface. The orientation of the surface is determined by the normal vector, and the location of the surface is determined by the bias. The discriminant function $g(\mathbf{x})$ is proportional to the signed distance from $\mathbf{x}$ to the hyperplane, with $g(\mathbf{x}) > 0$ when $\mathbf{x}$ is on the positive side, and $g(\mathbf{x}) < 0$ when $\mathbf{x}$ is on the negative side.

In a two-category linearly separable problem, we have a set of $n$ samples $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$, some labeled $w_1$ and some labeled $w_2$. We want to use these samples to determine the weights $\mathbf{w}$ in a linear discriminant function $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ .

A sample $\mathbf{x}_i$ is classified correctly if $\mathbf{w}\mathbf{x}_i > 0$ and $\mathbf{x}_i$ is labeled $w_1$ or if $\mathbf{w}\mathbf{x}_i < 0$ and $\mathbf{x}_i$ is labeled $w_2$. This suggests a "normalization" that simplifies the treatment of the two-category case: the replacement of all samples labeled $w_2$ by their negatives. With this normalization we can forget the labels and look for a weight vector $\mathbf{w}$ such that $\mathbf{w}\mathbf{y}_i > 0$ for all of the samples (where $\mathbf{y}_i$ is the normalized sample). Such a weight vector is called a separating vector or more generally a solution vector [50]. The solution vector, if exists, is not unique.

The approach to finding a solution to the set of linear inequalities $\mathbf{w}\mathbf{y}_i > 0$ will be to

define a criterion function $J(\mathbf{w})$ that is minimized if $\mathbf{w}$ is a solution vector. This reduces our problem to one of minimizing a scalar function, a problem that can often be solved by a gradient descent procedure.

One good choice for our criterion function is *Perceptron Criterion Function*, which is defined as

$$J_p(\mathbf{w}) = \sum_{\mathbf{y}\in\gamma} -\mathbf{w}^T\mathbf{y} \tag{3.16}$$

where $\gamma(\mathbf{w})$ is the set of samples misclassified by $\mathbf{w}$ . Geometrically, $J_p(\mathbf{w})$ is proportional to the sum of distances from the misclassified samples to decision boundary.

### 3.3.2 Support Vector Machines

Support Vector machines (SVMs) are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVMs are kernel machines based on the principle of structural risk minimization, which are used in applications of regression and classification; however, they are mostly used as binary classifiers. Although the subject can be said to have started in the late seventies (Vapnik, 1979), it is receiving increasing attention recently by researchers. It is such a powerful method that in the few years since its introduction, it has outperformed most other systems in a wide variety of applications, especially in pattern recognition.

Linear learning machines are the fundamental formulations of SVMs. The objective of the linear learning machine is to find the linear function that minimizes the generalization error from a set of functions which can approximate the underlying mapping between the input and output data. Consider a learning machine that implements linear functions in the plane as decision rules

$$f(\mathbf{x};\mathbf{w},b) = sign(\mathbf{w}^T\mathbf{x} + b) \tag{3.17}$$

with $N$ given training data with input values $x_k \in \Re^n$ and output values $y_k \in \{1, -1\}$.

The empirical error is defined as

$$\Re_{emp}(\theta) = \frac{1}{N}\sum_{k=1}^{N}|y_k - f(\mathbf{x};\mathbf{w},b)| = \frac{1}{N}\sum_{k=1}^{N}|y_k - sign(\mathbf{w}^T\mathbf{x} - b)| \tag{3.18}$$

where $\theta = (\mathbf{w}, \mathbf{b})$.

The generalization error can be expressed as

$$\Re(\theta) = \int |y - f(\mathbf{x},\theta)|p(\mathbf{x},y)dxdy \tag{3.19}$$

which measures the error for all input/output patterns that are generated from the underlying generator of the data characterized by the probability distribution $p(x, y)$ which is considered to be unknown.

According to statistical learning theory, the generalization (test) error can be upper bounded in terms of training error and a confidence term as shown in Equ. 3.20:

$$\Re(\theta) \leq \Re_{emp}(\theta) + \sqrt{\frac{h(ln(2N/h)+1) - ln(\eta/4)}{N}} \tag{3.20}$$

The term on left side represents generalization error. The first term on right hand side is empirical error calculated from the training data and the second term is called confidence term which is associated with the VC dimension $h$ of the learning machine. VC dimension is used to describe the complexity of the learning system. The relationship between these three items is illustrated in Fig. 3.12.

Thus, even though we don't know the underlying distribution based on which the data are generated, it is possible to minimize the upper bound of the generalization error in place of minimizing the generalization error. That means one can minimize the expression in the right hand side of the equation 3.20.

Unlike the principle of Empirical Risk Minimization (ERM) applied in Neural Networks which aims to minimize the training error, SVMs implement Structural Risk Minimization (SRM) in their formulations. SRM principle takes both the training error and the complexity of the model into account and intends to find the minimum of the sum of these two terms

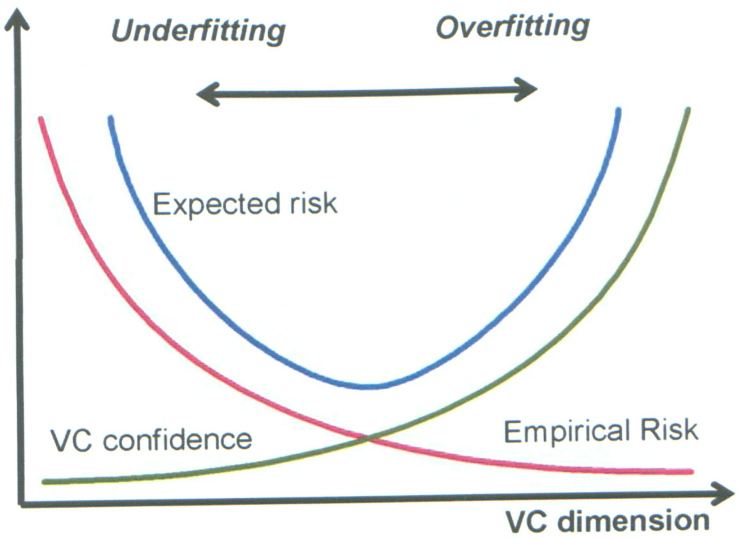


**Figure 3.12:** The relation between expected risk, empirical risk and VC confidence in SVMs.

as a trade-off solution (as shown in Fig. 3.12 ) by searching a nested set of functions of increasing complexity.

### Linear Support Vector Machines

Consider a binary classification problem with $\mathbf{x}_i \in R^d$ as its feature vector and $\mathbf{y}_i \in \{-1, +1\}$ the class labels (i.e. $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ are the training sets). The hyperplane which separates the two classes is

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0 \tag{3.21}$$

The function of SVM is based on choosing the hyperplane which minimizes the margin between two classes (Fig. 3.13) [51][56]. Thus, the hyperplane $(\mathbf{w}, b)$ that solves the optimization problem

$$minimize_{\mathbf{w},b} \quad \frac{1}{2} \parallel \mathbf{w} \parallel^2 \tag{3.22}$$

subject to

$$y_i(< \mathbf{w}.\mathbf{x}_i > +b) \geq 1 \qquad i = 1, \ldots, n$$

realizes the maximal margin hyperplane with geometric margin.

This is convex optimization problem (quadratic criterion with linear inequality constraints)

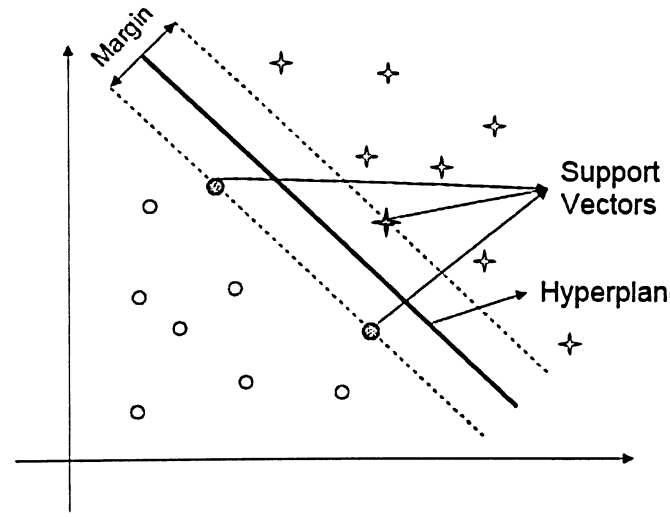


**Figure 3.13:** A linear SVM classifier. Support vectors are those elements of the training set which are on the boundary hyperplanes of two classes.

that has one unique solution.

The primal Lagrangian is

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} < \mathbf{w}.\mathbf{w} > - \sum_{i=1}^{n} \alpha_i [y_i (< \mathbf{w}.\mathbf{x}_i > + b) - 1] \tag{3.23}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

The corresponding dual is found by differentiating with respect to $\mathbf{w}$ and $b$:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j < \mathbf{x}_i.\mathbf{x}_j > \tag{3.24}$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \alpha_i \geq 0, i = 1, ..., n$$

The advantage of using the dual representation is derived from the fact that in this representation the number of free parameters relies on the number of data instead of the number of dimensions of the input space (equals the dimension of weight vector in the primal space). This property enables the classification in a high dimensional space.

But in many real-world problems the data is noisy; therefore there will in general be no

linear separation. In this case instead of hard margin, we use soft margin (the noise tolerant version), and slack variables denoted by $\xi$ , can be introduced to relax the constrains [51][56]. So our optimization problem would be

$$minimize_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3.25}$$

subject to

$$y_i(<\mathbf{w}.\mathbf{x}_i> +b) \geq 1-\xi_i \qquad \xi_i \geq 0, i=1,...,n$$

where $C$ is regularization parameter which is a trade off between the empirical risk (reflected by the second term in Equ. 3.25) and model complexity (reflected by the first term in Equ. 3.25). The dual form of this case is the same as Equ. 3.24 except that the constraint is different:

$$0 \leq \alpha_i \leq \frac{C}{N} \quad ,i=1,...,n \tag{3.26}$$

Considering the model complexity in the optimization problem, prevents overfitting.

Now if we suppose that $\alpha_i^*$ is the answer to 3.25, and with making use of so important Karush-Kuhn-Tucker (KKT) conditions, the optimal hyperplane can be expressed in the dual representation:

$$f(\mathbf{x}, \boldsymbol{\alpha}^*, b_0) = \sum_{i=1}^{N_{sv}} y_i\alpha_i^* <\mathbf{x}_i.\mathbf{x}> +b_0 \tag{3.27}$$

where $N_{sv}$ is the number of support vectors.

### Non-linear Support Vector Machines

In most of the real-world cases the data points are not linearly separable. In this case we use a non-linear operator $\varphi(.)$ to map the data to a higher dimensional space $\mathcal{F}$ (*Feature Space*), where it can be classified linearly. Figure 4.2 illustrates this mapping.

In other words, a linear learning machine can be employed in the feature space to solve the original non-linear problem. Kernel functions satisfying Mercer condition not only enable implicit mapping of data from input space to feature space but also ensure the convexity

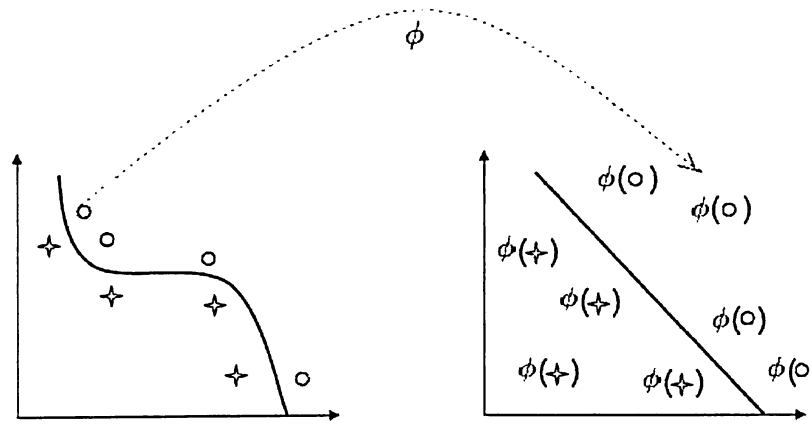


**Figure 3.14:** Mapping from input space to a higher dimensional feature space by means of a kernel function.

of the cost function which leads to the unique optimum. Mercer condition states that a continuous symmetric function $K(x, z)$ must be positive semi-definite to be a kernel function which can be written as inner product between the data pairs.

So the hypothesis in this case would be

$$f(\mathbf{x}) = \mathbf{w}^T\varphi(\mathbf{x}) + b \tag{3.28}$$

which is linear in terms of the mapped data ($\varphi(\mathbf{x})$ ). Now we can extend all the presented optimization problems for the linear case, for the transformed data in the feature space. If we define the Kernel function as

$$K(\mathbf{x}_i, \mathbf{x}_j) =< \varphi(\mathbf{x}_i).\varphi(\mathbf{x}_j)) > \tag{3.29}$$

where $\varphi$ is a mapping from input space to an (inner product) feature space $\mathcal{F}$ . Then the corresponding dual form is

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i.\mathbf{x}_j) \tag{3.30}$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \alpha_i \geq 0, i = 1, ..., n$$

The cost function $W(\alpha)$ in Equ. 3.30 is convex and quadratic in terms of the unknown parameters. This problem is solved through quadratic programming. The Karush-Kuhn-Tucker conditions for Equ. 3.30 lead to the following final decision rule

$$f(\mathbf{x}, \alpha^*, b_0) = \sum_{i=1}^{N_{sv}} y_i \alpha_i^* K(\mathbf{x}_i.\mathbf{x}) + b_0 \tag{3.31}$$

where $N_{sv}$ and $\alpha_i^*$ denote number of support vectors and the non-zero Lagrange multipliers corresponding to the support vectors respectively. Note that we don't have to know the underlying mapping function, however it is necessary to define the Kernel function.

Several typical choices of kernels are linear, polynomial, Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) kernel. Their expressions are as following:

$$K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}_i^T \mathbf{x} \quad \text{linear kernel} \tag{3.32}$$

$$K(\mathbf{x}, \mathbf{x}_i) = (\tau + \mathbf{x}_i^T \mathbf{x})^d \quad \text{polynomial kernel} \tag{3.33}$$

$$K(\mathbf{x}, \mathbf{x}_i) = exp(-\|\mathbf{x} - \mathbf{x}_i\|_2^2/\sigma_2) \quad \text{RBF kernel} \tag{3.34}$$

$$K(\mathbf{x}, \mathbf{x}_i) = tanh(k_1 \mathbf{x}_i^T \mathbf{x} + k_2) \quad \text{MLP kernel} \tag{3.35}$$

### Least Squares Support Vector Machines

Although SVMs have many appealing properties that avoid the problems (e.g., overfitting, inefficient of training and testing, a large set of parameters to be tuned) frequently associated with the classical supervised learning methods, they also have some drawbacks as any other techniques. The standard SVM requires the kernel matrix to be cashed to improve the computation speed that makes online learning infeasible. Also, the fact that SVM formulation is a convex quadratic programming (QP) guarantees the global optima, but QP is still difficult to solve, especially for the learning tasks where the speed is concerned. Based on the concept and formulation of SVMs, many researchers have been investigating the modification and improvements for different purposes. Least Squares Support Vector Machine (LS-SVM) is a reformulation of the standard SVM. LS-SVM models for classification and nonlinear regression are characterized by simplifying the quadratic optimization problem into a system

of linear equations. Such characterizations of LS-SVM allow fast training and less storage hence enable its use in on-line training.

LS-SVMs are reformulations to standard SVMs which lead to solving linear KKT systems [51]. In LS-SVMs the inequality constraints in SVM are replaced with equality constrains. As a result the solution follows from solving a set of linear equations instead of a quadratic programming problem which we have in original SVM formulation of Vapkin and obviously we can have a faster algorithm.

In LS-SVM's an equality constraint based formulation is made within the context of ridge regression [57] as follows

$$minimize_{\mathbf{w},b} J(\mathbf{w},b) = \frac{1}{2}\|\mathbf{w}\|^2 + \gamma\frac{1}{2}\sum_{i=1}^{n} e_i^2 \tag{3.36}$$

subject to

$$y_i[\mathbf{w}^T\varphi(\mathbf{x}_i) + b] = 1 - e_i \qquad i = 0, 1, ..., n$$

with Lagrangian

$$L(\mathbf{w}, b, e; \alpha) = J(\mathbf{w}, b) - \sum_{i=1}^{N} \alpha_i y_i[\mathbf{w}\varphi(\mathbf{x}_i) + b] - 1 + e_i \tag{3.37}$$

where $\alpha_i$'s are Lagrange multipliers (Support Values).

This LS-SVM formulation modifies Vapnik's SVM at two points. First, LS-SVM takes equality constraints instead of inequality constraints. Second, the error variable $e_i$ was introduced in the sense of least-square minimization. These error variables play a similar role as the slack variables in SVM formulation such that relatively small errors can be tolerated.

Taking the condition for optimality of the Lagrangian yields a set of linear equations shown in equation set 3.38:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \\ \frac{\partial L}{\partial b} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} \alpha_i = 0 \\ \frac{\partial L}{\partial e_i} = 0 \quad \rightarrow \quad \alpha_i = \gamma e_i, \quad i = 1, \ldots, N \\ \frac{\partial L}{\partial \alpha_i} = 0 \quad \rightarrow \quad w^T\varphi(x_i) + b + e_i - y_i = 0, i = 1, \ldots, N. \end{cases} \tag{3.38}$$

Solving this set of linear equations in $\alpha$ and $b$, the resulting LS-SVM model becomes the following equation:

$$y(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{3.39}$$

As it was shown in the previous section, SVMs solve the nonlinear regression problems by means of convex quadratic programs (QP). The use of least squares and equality constraints for the models leads to solving a set of linear equations, which is easier to use than QP solvers. It also has potential drawbacks such as the lack of sparseness which is indicated from the condition $\alpha_i = \gamma e_i$ in equation set 3.38 since the error would not be zero for most of data points. This is important in the context of an equivalence between sparse approximation and support vector machines. One can overcome the drawbacks using special pruning techniques for sparse approximation [58].

### Multi-class Support Vector Machines

Support vector machines were originally designed for binary classification; therefore we need a methodology to adopt the binary SVMs to a multi-class problem like our emotion recognition problem. How to effectively extend SVMs for multi-class classification is still an ongoing research issue. Currently the most popular approach for multi-category SVM is by constructing and combining several binary classifiers. Different coding and decoding strategies can be used for this purpose among which one-against-all and one-against-one (pairwise) are the most popular ones. Some of these methods are elaborated below.

*a) One-Against-All SVMs*: Assume that we have $n$ discrete classes. For a one-against-all SVM, we determine $n$ decision functions that separate one class from the remaining classes. Let the $i^{th}$ decision function, with the maximum margin, that separates class $i$ from the remaining classes be

$$D_i(\mathbf{x}) = \mathbf{w}_i^t g(\mathbf{x}) + b_i \tag{3.40}$$

The hyperplane $D_i(\mathbf{x}) = \mathbf{0}$ forms the optimal separating hyperplane and if the classification problem is separable, the training data $\mathbf{x}$ belonging to class $i$ satisfy

$$\begin{cases} D_i(\mathbf{x}) \geq 1 & ,\mathbf{x} \text{ belong to class } i \\ D_i(\mathbf{x}) \leq -1 & ,\mathbf{x} \text{ belong to remaining classes} \end{cases} \tag{3.41}$$

In other words, the decision function is the sign of $D_i(\mathbf{x})$ and therefore it is a discrete function. If 3.41 is satisfied for plural $i's$ , or there is no $\mathbf{x}$ that satisfies 3.41, $\mathbf{x}$ is unclassifiable.

*b) Fuzzy One-Against-All SVMs:* One way of avoiding unclassifiable regions is to introduce membership functions [55]. For class $i$ we define one-dimensional membership functions $m_{ij}(\mathbf{x})$ as

1) $for i = j$

$$m_{ij}(\mathbf{x}) = \begin{cases} 1 & \text{for } D_i(\mathbf{x}) \geq 1 \\ D_i(\mathbf{x}) & \text{otherwise} \end{cases} \tag{3.42}$$

2) $for i \neq j$

$$m_{ij}(\mathbf{x}) = \begin{cases} 1 & \text{for } D_i(\mathbf{x}) \leq -1 \\ -D_i(\mathbf{x}) & \text{otherwise} \end{cases} \tag{3.43}$$

For $i \neq j$ , class $i$ is on the negative side of $D_j(\mathbf{x}) = \mathbf{0}$ .

After computing the membership values $m_{ij}(\mathbf{x})$ for $j = 1, \ldots, n$ , we define the membership function of $\mathbf{x}$ for class $i$ as

$$m_i(\mathbf{x}) = \frac{1}{n} \sum_{j=1,\ldots,n} m_{ij}(\mathbf{x}) \tag{3.44}$$

And finally the data point $\mathbf{x}$ classified into the class with the maximum membership value:

$$arg \quad \max_{i=1,\ldots,n} \quad m_i(\mathbf{x}) \tag{3.45}$$

*c) One-Against-One SVMs:* Another encoding method for converting the binary classifier into a multi-class one is one-against-one. In this method we construct a binary classifier for each possible pair of classes and therefore for $n$ classes we will have $\frac{(n)(n-1)}{2}$ decision

functions. The decision function for the pair of classes $i$ and $j$ is given by

$$D_{ij} = \mathbf{w}_{ij}^t g(\mathbf{x}) + b_{ij} \tag{3.46}$$

where $D_{ij}(\mathbf{x}) = -D_{ij}(\mathbf{x})$.

The final decision is achieved by maximum voting scheme. That is for the datum $\mathbf{x}$ we calculate

$$D_i(\mathbf{x}) = \sum_{j \neq i, i=1} sign(D_{ij}(\mathbf{x})) \tag{3.47}$$

And the datum is classified into the class

$$arg \quad \max_{i=1,\ldots,n} \quad D_i(\mathbf{x}) \tag{3.48}$$

*d) Fuzzy One-Against-One SVMs:* If 3.48 is satisfied for plural $i's$, $\mathbf{x}$ is unclassifiable. To avoid this, similar to fuzzy one-against-all, we introduce the fuzzy membership function. First, we define the one-dimensional membership function $m_{ij}$ as

$$m_{ij}(\mathbf{x}) = \begin{cases} 1 & \text{for } D_{ij}(\mathbf{x}) \geq 1 \\ D_{ij}(\mathbf{x}) & \text{otherwise} \end{cases} \tag{3.49}$$

$m_i(\mathbf{x})$ of $\mathbf{x}$ for class $i$ is given by

$$m_i(\mathbf{x}) = \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} m_{ij}(\mathbf{x}) \tag{3.50}$$

And $\mathbf{x}$ is classified into class

$$arg \quad \max_{i=1,\ldots,n} \quad m_i(\mathbf{x}) \tag{3.51}$$

### 3.3.3 Advantages and Disadvantages of Machine Learning

It is not surprising that the promise of a learning methodology should be so tantalizing. First, the range of applications that can potentially be solved by such an approach is very large. Second, it appears that we can also avoid much of the laborious design and programming inherent in the traditional solution methodologies, at the expense of collecting some labeled

data and running an off-the -shelf algorithm for learning the input/output mapping. Finally, there is the attraction of discovering insights into the way that humans learn, an attraction that inspired early work in neural networks. There are, however, many difficulties inherent in the learning methodology, difficulties that deserve careful study and analysis. One example is the choice of class of functions from which the input/output mapping must be sought. The class must be chosen to be sufficiently rich so that the required mapping or an approximation to it can be found, but if the class is too large the complexity of learning from examples can become prohibitive, particularly when taking into account the number of examples required to make statistically reliable inferences in a large function class. In practice these problems manifest themselves in specific learning difficulties. The first is that the learning algorithm may prove inefficient as for example in the case of local minima. The second is that the size of the input hypothesis can frequently become very large and impractical. The third problem is when there is only a limited number of training examples to reach a hypothesis class, therefore it will lead to overfitting and hence poor generalization. The forth problem is that frequently the learning algorithm is controlled by a large number of parameters that are often chosen by tuning heuristics, making the system difficult and unreliable to use. Despite the drawbacks, there have been notable successes in the application of the learning methodology to problems of practical interest.

# Chapter 4

# Emotion Recognition Using Sequence Discriminant SVMs

SPEECH , vision, text and biosequence data can be difficult to deal with in the context of simple statistical classification problems, because the examples to be classified are often sequences or arrays of variable size that may have been distorted in particular ways. It is common to estimate a generative model for such data, and then use Bayes rule to obtain a classier from this model. However, many discriminant functions which predict the class labels directly, as in support vector machines, have proven to be superior to generative models for classification problems.

During the last decade Support Vector Machines (SVMs) have proven to be successful discriminative approaches to pattern classification problems. Excellent results have been reported in applying SVMs in multiple domains. However, the application of SVMs to data sets where each element has variable length remains problematic. On the other hand statistical models such as Gaussian Mixture Models (GMM) or Hidden Markov Models make strong assumptions about the data. They are simple to learn and estimate, and are well understood by the multimedia community. It is therefore attractive to explore methods that combine these models and discriminative classifiers. The Fisher kernel proposed by Jaakkola [59] effectively combines both generative and discriminative classifiers for variable length sequences.

In this chapter the issue of variable length in speech sequences is addressed, several

proposed alternative solutions are discussed and our adopted technique is explained. The experimental results are compared with the results achieved in the previous chapter.

## 4.1 The Problem of Variable-Length Sequences

In our approach to emotion recognition problem, the goal is classifying the whole speech utterances, rather than frame-level classification. Since the lengths of utterances are different, the sequences of feature vectors for each signal show different lengths.

A drawback of SVMs when dealing with audio data is their restriction to work with fixed-length vectors. Both in the kernel evaluation and in the simple input space dot product, the units under processing are vectors of constant size. However, when working with audio signals, although each signal frame is converted into a feature vector of a given size, the whole acoustic event is represented by a sequence of feature vectors, which shows variable length. In order to apply SVM to this kind of data, one needs either to somehow normalize the size of the sequence of input space feature vectors or to find a suitable kernel function that can deal with sequential data. Several methods have been suggested to cope with this problem [60] some of which are explained below:

### Extracting Statistical Parameters

The easiest and the most common approach is to extract some statistical parameters (e.g. mean and standard deviation) from the sequence of vectors and thus transform the problem into that of fixed-length vector spaces. This is the method we adopted in Chapter 3. Despite the good results we obtained using this approach, when frame-level features are transformed into statistical event-level features, there exists an unavoidable loss of information.

### Outerproduct of Trajectory Matrix

The time analysis of the data gives a sequence of $l$-dimensional parametric vectors. The sequence is considered as a trajectory in the $l$-dimensional space. If we define the $l$-by-$m$

trajectory matrix as $X = [x_1, x_2, \ldots, x_m]$, the outerproduct matrix $Z$ [60] is defined as

$$Z = XX^T \tag{4.1}$$

Thus the outerproduct matrix $Z$ is $l$-by-$l$ and no longer depends on the length of the sequence. The vectorized outerproduct thus can feed the SVM classifier directly. It is obvious that this method explicitly considers sequence duration information.

### Gaussian Dynamic Time Warping (GDTW)

This approach as well as the previous one does not assume a model for the generative class conditional densities. The GDTW [60] method addresses the problem of variable length sequences classification by introducing the DTW technique to SVM kernel. Recalling the standard RBF kernel for SVM

$$K(\mathbf{x}, \mathbf{y}) = \exp(\frac{-\|\mathbf{x} - \mathbf{y}\|_2^2}{\sigma_2}) \tag{4.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote two patterns to compare. As mentioned before, if the two patterns are sequences of different length, they cannot be compared in the kernel evaluation directly. An obvious modification of 4.2 is to substitute the squared Euclidian distance computation with the equivalent that can cope with temporally distorted, variable-length sequences. GDTW kernel is defined as

$$K(\mathbf{x}, \mathbf{y}) = \exp(\frac{-D(\mathbf{x}, \mathbf{y})}{\sigma_2}) \tag{4.3}$$

where $D(\mathbf{x}, \mathbf{y})$ is a DTW distance between sequences $\mathbf{x}$ and $\mathbf{y}$. The proposed method was applied to handwriting recognition in [70].

### Fisher Kernel

Fisher kernel is one of the most successful methods that enable SVM to classify whole sequences. Generative probability models such as hidden Markov models provide a principled way of treating missing information and dealing with variable length sequences. On the other hand, discriminative methods such as support vector machines enable us to construct flexible decision boundaries and often result in classification performance superior to that of

the model based approaches. An ideal classifier should combine these two complementary approaches.

A generative method focuses on explaining the training data. A discriminative model, on the other hand, focuses on finding the boundary between classes in some feature space. Because of this property, discriminative methods often outperform generative models at classification. A major difficulty with using a discriminative method for audio classification is that each audio file $X$ consists of a sequence $x_1, \ldots, x_n$ , where $n$ varies among audio files; discriminative methods require a fixed length feature vector.

Fisher kernels make use of the information obtained by underlying generative models. It was first developed and applied to biological sequence analysis by Jaakkola and Haussler [59].The basic theory of Fisher kernels is to map the variable-length sequences to a single point in a fixed-dimension (and comparatively high-dimension) space called score-space. The Fisher scores for a given input sequence $X$ and a generative model $M$ parameterized by $\boldsymbol{\theta}$ are computed as:

$$\mathbf{U}_X = \nabla_\theta \log P(X|M, \boldsymbol{\theta}) \tag{4.4}$$

In fact, Fisher scores are derivatives of the log-likelihood with respect to all single parameters of the model. In some sense, it is a measure of how well a sequence matches the model. When the model is considered Gaussian Mixture Model (GMM), the parameter set consists of mean vectors, covariance matrices, and weights. See Appendix A for explicit formulas for computing Fisher score when the generative model is the diagonal covariance matrix.

The Fisher kernel is defined as:

$$K(X_i, X_j) = U_{X_i}^T I^{-1} U_{X_j} \tag{4.5}$$

where $I$ is called Fisher information matrix and is computed by

$$I = E_X\{\ U_X U_X^T\} \tag{4.6}$$

where the expectation is over $P(X|\boldsymbol{\theta})$ .

Fisher kernel defined in 4.5 is a valid kernel function since $I$ is positive definite. Furthermore, in [59] it was shown that, under the condition that the class variable is a latent

variable in the probability model, the learning machines that use Fisher kernel, are asymptotically at least as good as making decision based on the generative model itself (maximum a posteriori).

Applying this approach results in a sparse data problem for which SVMs are well suited.

## 4.2 Sub-Band Approach AER System

In our second approach to the emotion recognition problem, spectral features are extracted from non-overlapping logarithmically scaled frequency sub-bands listed in Table 4.1 rather than frames. The sub-band approach will provide better discrimination since for different emotions, different energy distributions in different frequency sub-bands can be captured. Since the resolution of the human hearing approximately decreases according to a logarithmic relationship with increasing frequency, it is reasonable from a perceptual point of view to divide the frequency bands according to a logarithmic frequency scale. The Mel cepstrum scale, which is used in this thesis, is a widely-known scale aiming at resembling the critical bands of human hearing. The following equation shows the relationship between Hertz and Mel frequency (m):

$$m = 1127.01048 \log_e(1 + f/700) \tag{4.7}$$

As Fig. 4.1 shows, most of the important information in speech signals is located in the lower frequencies; therefore the $6^{th}$ sub-band is dismissed. As a result, we gain more precision and also less computation complexity.

The structure of the proposed AER system when features are extracted from frequency sub-bands is depicted in Fig. 4.2.

The preprocessing and windowing procedures are exactly the same as those of previous approach described in sections 3.1.1 and 3.1.2.

**Table 4.1:** Sub-band allocation for calculating spectral features.

| Sub-band | Lower Edge (Hz) | Upper Edge (Hz) |
|---|---|---|
| 1 | 0 | 780 |
| 2 | 781 | 2000 |
| 3 | 2001 | 3900 |
| 4 | 3901 | 6800 |
| 5 | 6801 | 11500 |
| 6 | 11501 | 22050 |

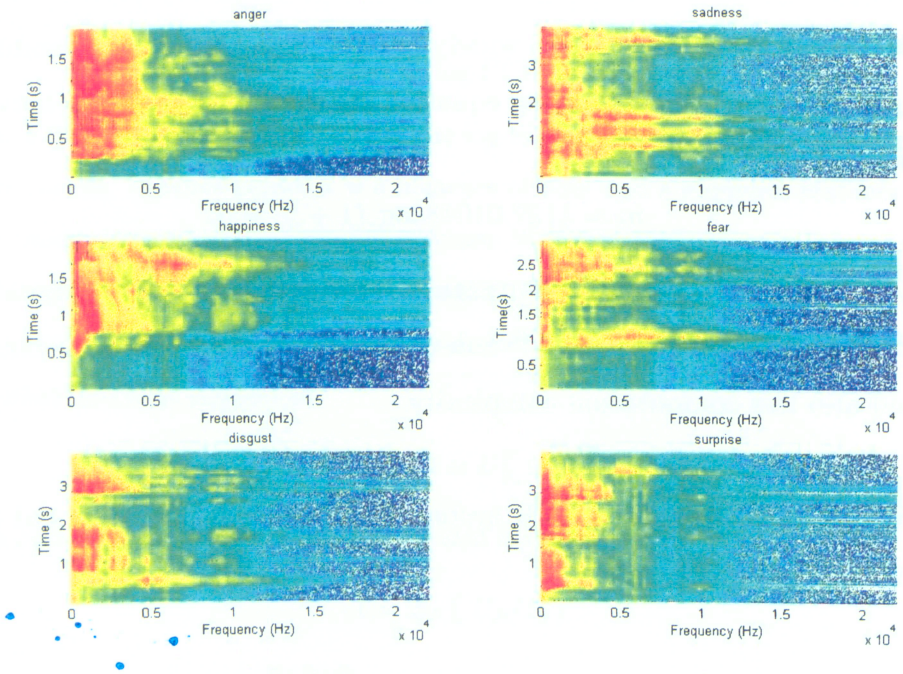


**Figure 4.1:** The corresponding spectrograms for six different emotions.

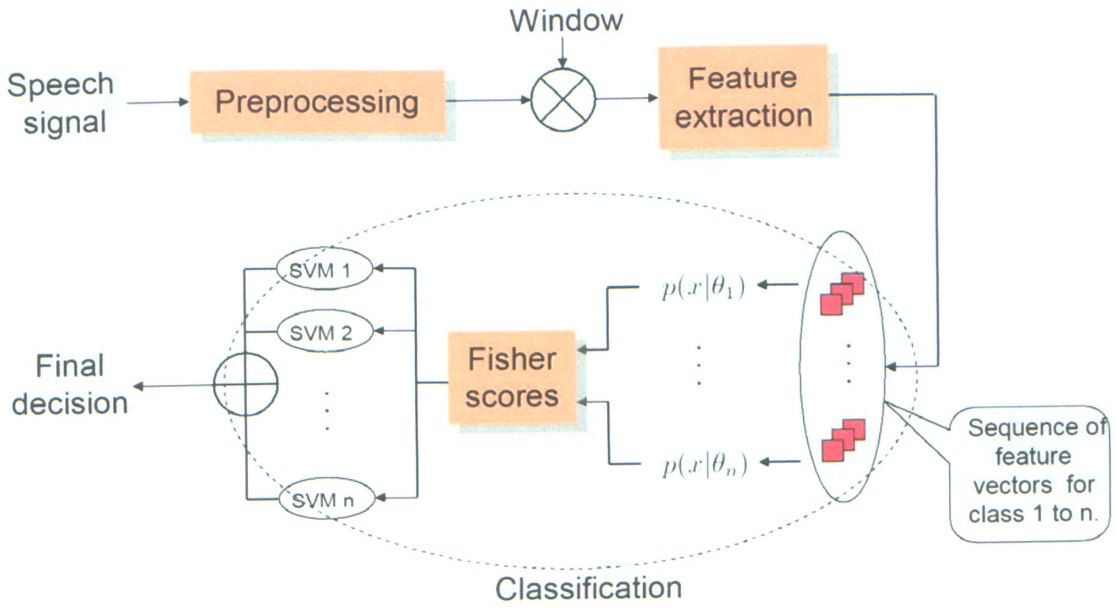


**Figure 4.2:** The structure of the speech emotion recognition system for sub-band approach.

### 4.2.1 Feature Extraction

The features used in this part are the same as those listed in Table 3.5 with three additional acoustic features. The complete list of features is listed in Table 4.3 and the definitions of the three new features are given below:

Let $s_i(n)$ represents the $i^{th}$ frame of the signal with $n = 1, \ldots, N$. Let $F_i = f_i(u)$,$u \in (0, M)$ be the Fourier transform of the $i^{th}$ frame, where $M$ is the index of the highest frequency band. As mentioned before, in order to capture more detail, spectral features are extracted from non-overlapping logarithmically spaced frequency sub-bands. Let $l_b$ and $u_b$ be the lower and upper edges of the frequency band $b$.

1. Spectral Flatness Measure (SFM): The spectral flatness measure quantifies the flatness of the spectrum and distinguishes between noise-like and tone-like signal. SFM is

defined as

$$SFM_{i,b} = \frac{[\prod_{u=l_b}^{u_b} |f_i(u)|^2]^{\frac{1}{u_b - l_b + 1}}}{\frac{1}{u_b - l_b + 1} \sum_{u=l_b}^{u_b} |f_i(u)|^2} \tag{4.8}$$

2. Spectral Crest Factor (SCF): The spectral crest factor is also a measure of tonality of the signal. SCF is defined as

$$SCF_{i,b} = \frac{max(|f_i(u)|^2)}{\frac{1}{u_b - l_b + 1} \sum_{u=l_b}^{u_b} |f_i(u)|^2} \tag{4.9}$$

3. Spectral Band Energy (SBE): The spectral band energy is the energy in the frequency bands normalized by the energy in the whole spectrum. SBE is defined as

$$SBE_{i,b} = \frac{\sum_{u=l_b}^{u_b} |f_i(u)|^2}{\sum_{u=0}^{M} |f_i(u)|^2} \tag{4.10}$$

Let $X_i$ be the feature vector extracted for the $i^{th}$ frame. So we have a sequence of feature vectors $(X)$ for each signal.

$$X_i = [SE_i^{s_1}, \ldots, SE_i^{s_5}, \ldots, SBE_i^{s_1}, \ldots, SBE_i^{s_5}, MFCC_i^1, \ldots, MFCC_i^{13}, ZCR_i]^T$$

where $s_i$ denotes the $i^{th}$ sub-band frequency.

$$X = [X_1, X_2, \ldots, X_n]$$

where $n$ is the number of frames. Finally the feature matrix $X$ is mean subtracted and component wise variance normalized to get a normalized feature matrix.

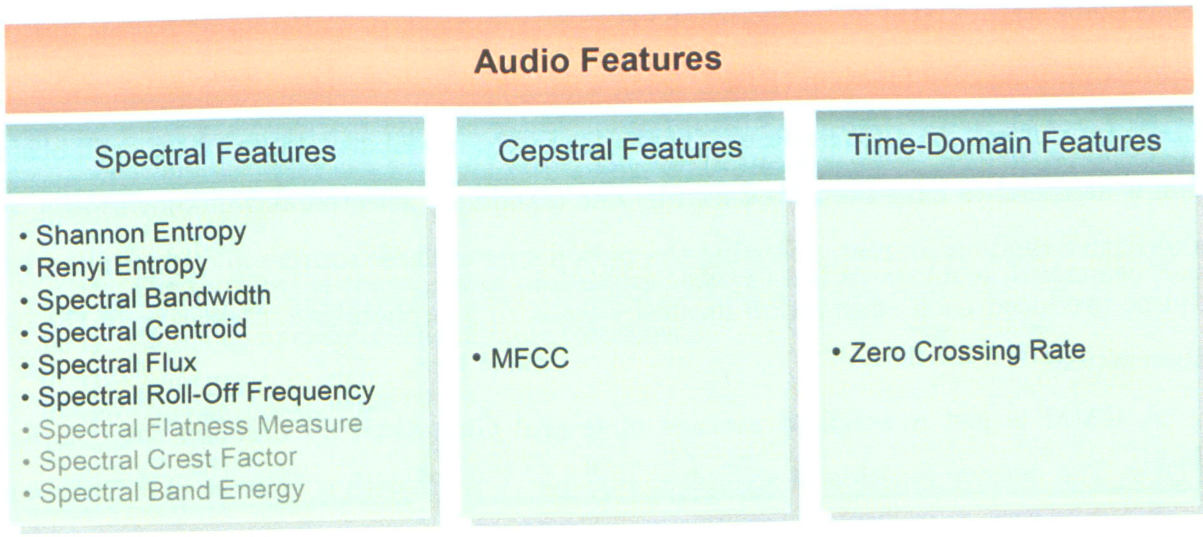


**Figure 4.3:** Complete list of acoustic features used for speech emotion recognition.

### 4.2.2 Classification

We could again consider some statistical values of the variables in order to present one feature vector per signal, but as explained in Sec. 4.1, this method causes some loss of information. As our goal in the sub-band approach is gathering more information, we have to adopt some other method in order to have an utterance-level classification. Fisher kernels are utilized in our second approach to the AER system. In order to compare the result obtained by Fisher kernel (sequence discriminant SVMs) and generative models, GMMs are also applied. These two methods are elaborated below.

#### Gaussian Mixture Models (GMMs)

Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering. Finite mixtures are a flexible and powerful probabilistic modeling tool for univariate and multivariate data. The usefulness of mixture models in areas that involve statistical modeling of data (such as pattern recognition, computer vision, signal and image analysis,and machine learning) is currently widely acknowledged. GMM is in fact the most popular and widely-used method in the fields of speak recognition and speaker verification

[29][71][72]. Here GMM is used to model the extracted emotion content of speech signals as a probability density function (PDF), using a weighted combination of Gaussian component PDFs (mixtures). In fact, finite mixtures adequately model situations where each observation is assumed to have been produced by one (randomly selected and unknown) of a set of alternative random sources. Inferring the parameters of these sources and identifying which source produced each observation naturally leads to a probabilistic clustering of the set of observations.

A GMM is just a weighted average of several Gaussian PDFs, called the component PDFs. The density function of a random variable $X \in R^d$ with a mixture of $k$ Gaussians is defined as:

$$f(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^{k} \alpha_j \frac{1}{\sqrt{(2\pi)^2|\Phi_j|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \Phi_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right\} \tag{4.11}$$

with parameter set $\boldsymbol{\theta} = \{\alpha_j, \mu_j, \Phi_j\}_{j=1}^{k}$ having the following parameters:

- weight $\alpha_j > 0$, $\sum_{j=1}^{k} \alpha_j = 1$
- mean $\boldsymbol{\mu}_j \in R^d$, and
- covariance matrix $\Phi_j$.

The standard method used to fit finite mixture models to observed data is the well-known Expectation-Maximization (EM) algorithm [61]. The EM algorithm is used to locate a maximum likelihood (ML) to estimate the mixture parameters.

Given a set of feature vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$, the maximum likelihood estimation of $\boldsymbol{\theta}$ is:

$$\boldsymbol{\theta}_{ML} = arg\max_{\theta} L(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \tag{4.12}$$

$$= arg\max_{\theta} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta})$$

Given the current estimation of the parameter set $\boldsymbol{\theta}$, each iteration of the EM algorithm re-estimates the parameter set according to the following two steps:

1. Expectation step:

$$w_{ij} = \frac{\alpha_j f(\mathbf{x}_i|\boldsymbol{\mu}_j, \Phi_j)}{\sum_{l=1}^{k} \alpha_l f(\mathbf{x}_i|\boldsymbol{\mu}_l, \Phi_l)} \qquad (4.13)$$
$$j = 1\ldots, k \qquad i = 1, \ldots, n.$$

The term $w_{ij}$ is the posterior probability that the feature vector $\mathbf{x_i}$ is sampled from the $j^{th}$ component of the mixture distribution.

2. Maximum step:

$$\alpha'_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij} \qquad (4.14)$$

$$\mu'_j = \frac{\sum_{i=1}^{n} w_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} w_{ij}} \qquad (4.15)$$

$$\sum_j' = \frac{\sum_{i=1}^{n} w_{ij}(\mathbf{x}_i - \boldsymbol{\mu'}_j)(\mathbf{x}_i - \boldsymbol{\mu'}_j)^T}{\sum_{i=1}^{n} w_{ij}} \qquad (4.16)$$

The parameter set $\boldsymbol{\theta}$ is updated repeatedly until the log-likelihood is increased by less than a predefined threshold from one iteration to the next to get the maximum likelihood parameters $\boldsymbol{\theta}_{ML}$. However, the EM algorithm for finite mixture fitting has several drawbacks: it is a local (greedy) method, thus sensitive to initialization because the likelihood function of a mixture model is usually multi-modal and for certain classes of mixtures it may converge to the boundary of the parameter space (where the likelihood is unbounded) thus leading to meaningless estimates [61].

A fundamental issue in mixture modeling is the selection of the number of components. The usual tradeoff in model order selection problems arises: with too many components, the mixture may over-fit the data, while a mixture with too few components may not be flexible enough to approximate the true underlying model [61]. The most common methods are setting the number of mixture components to an arbitrary constant, or a fraction of instances to the training set. However, there is no statistical justification for these methods,

and neither of these methods take the complexity of the data points distribution into account. To determine the number of components per GMM to best represent the true distribution of the data, model selection (MS) techniques will be used. This will provide a method to maximize the likelihood of the training data while attempting to avoid overfitting. Perhaps, the most widely-used model selection technique is the Akaike Information Criterion (AIC) [62], which penalizes the model based on its complexity. It is defined as:

$$AIC(\boldsymbol{\theta}) = -2\log p(X|\boldsymbol{\theta}) + 2k \tag{4.17}$$

where $\boldsymbol{\theta}$ is the model, $X$ is the input data, $\log p(X|\boldsymbol{\theta})$ is the log of the probability of $X$ given $\boldsymbol{\theta}$, and $k$ is the number of parameters in the model $\boldsymbol{\theta}$. The model selected will have the lowest AIC score.

After training the GMMs and obtaining the parameter sets for each class of data, the un-known feature vectors are used to evaluate the log-likelihood value (Equ. 4.12) of the all the models present in the database. We can choose the model that gives the highest log-likelihood.

### Exploiting GMMs in Discriminative Classifiers (SVMs)

The approach to emotion recognition problem outlined in the previous section operates at the frame-level with an overall sequence score obtained by averaging the likelihoods of each frame in the sequence rather than complete utterances. On the other hand, Support Vector Machines has shown a very good result in classification task in literature as well as our previously-reported results in Chapter 3 but they are restricted to fixed-length sequences. Therefore, by combining these two approaches we can benefit the strong performance of the SVMs and overcome the problem of variable-length (Fisher kernels, Sec 4.1).

The EM algorithm described in the previous section is again used to train the Gaussian Mixture Model. After finding the model parameters, the fisher score mapping is performed using Equ. 4.5. SVMs classify the mapped data points in the Fisher score space.

### 4.2.3 Implemented Results

As mentioned in previous sections, our database consists of 1287 instances of utterance. In this part of our experiment, the 5-fold cross validation technique (explained in section 3.2.2) is utilized to train and evaluate the classifiers.

Half of the procedure is exactly the same for both approaches (i.e. Fisher kernels and GMMs): a mixture of gaussian pdfs are fit to each class of data (emotion), and the corresponding parameters are found using EM algorithm. To initialize the parameters in the algorithm, *k-means algorithm*[73] is used. Also the covariance matrices are considered diagonal in order to reduce computational complexity. AIC criterion (Equ. 4.17) is used to determine the number of mixture components. Figure 4.4 shows the result of AIC criterion performance for one of the classes. The algorithm is stopped as soon as the AIC score is increased as the number of iterations is increased. In order to be consistent, the same number of components (i.e. the average of the result achieved according to AIC criterion) is assumed for GMM-based approach and for Fisher kernel-based approach, which is 20.

In the GMM-based approach emotion recognition, after finding the parameters of the models, the unknown samples are probabilistically classified, according to ML (Equ. 4.12) rule. In the Fisher kernel-based approach, after finding the parameters of the models, the data points are mapped into the Fisher score space (Equ. 4.4). The mapped data points are discriminatively classified using LS-SVMs. All the binary LS-SVMs are trained using linear kernel functions with different optimal regularization parameters. Fuzzy-pairwise method was adopted in order to extend the binary LS-SVMs to our multi-class problem, since this approach yielded the best result in our previous approach.

77.8% accuracy and 97.6% accuracy were achieved for GMM-based method and Fisher kernel-based method respectively, which shows the better performance of discriminative models at classification task.

Table 4.2 and Table 4.3 show the achieved confusion matrices. (The abbreviations in these tables stand for the six different emotions: *anger, fear, disgust, happiness, sadness,* and *surprise.*) We can see that in all cases the most difficult emotion to recognize in our

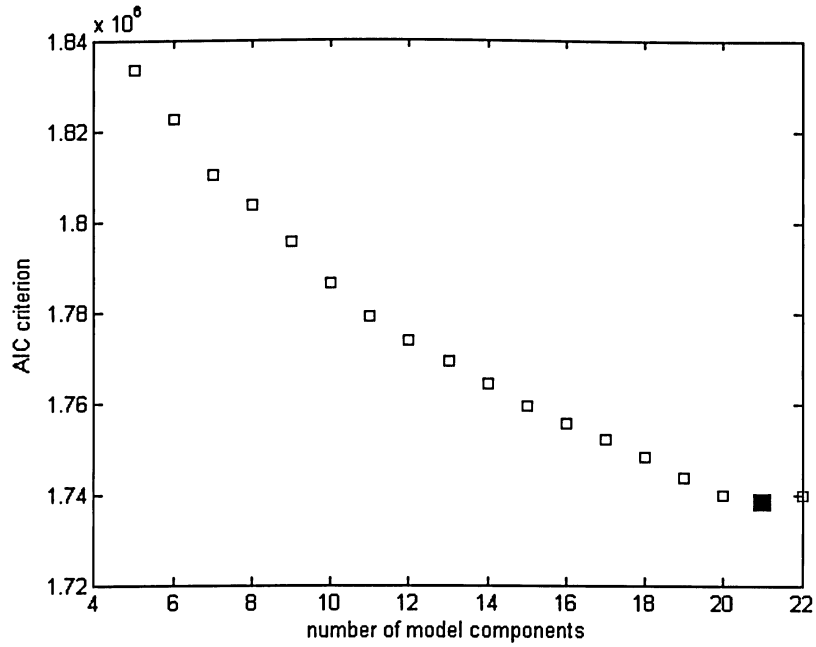


**Figure 4.4:** Deciding the number of Gaussian mixture components according to AIC criterion.

experiment is surprise.

The result obtained by sequence discriminant SVMs outperforms GMMs and all the methods adopted in Chapter 3. The flowchart in Fig. 4.5 enables us to have a comparative glance at all the performances obtained by the methods adopted in this thesis.

The performance of different implemented classifiers are also compared by means of Receiving Operating Characteristics (ROC) curves. ROC curves are a useful technique for organizing classifiers and visualizing their performance. ROC graphs have long been

**Table 4.2:** Confusion matrix for GMMs

| | Recognized Emotions (%) | | | | | |
|---|---|---|---|---|---|---|
| | Ang | Fea | Dis | Hap | Sad | Sur |
| Ang | **81.0** | 1.0 | 4.3 | 0 | 6.4 | 7.3 |
| Fea | 2.7 | **70.2** | 1.6 | 3.4 | 13.5 | 8.6 |
| Dis | 5.4 | 6.3 | **78.5** | 0 | 4.1 | 5.7 |
| Hap | 0.9 | 1.3 | 0 | **88.7** | 0 | 9.1 |
| Sad | 2.4 | 1.2 | 7.7 | 0 | **87.8** | 0.9 |
| Sur | 9.3 | 8.4 | 4.2 | 12.1 | 12.3 | **53.7** |

**Table 4.3:** Confusion matrix for fuzzy-pairwise LS-SVM with Fisher kernel

| | Recognized Emotions (%) | | | | | |
|---|---|---|---|---|---|---|
| | Ang | Fea | Dis | Hap | Sad | Sur |
| Ang | **99.8** | 0 | 0.15 | 0 | 0 | 0 |
| Fea | 0 | **100** | 0 | 0 | 0 | 0 |
| Dis | 0 | 0 | **99.8** | 0 | 0.15 | 0 |
| Hap | 0.15 | 0 | 0 | **99.8** | 0 | 0 |
| Sad | 0 | 0 | 0.31 | 0 | **99.6** | 0 |
| Sur | 0 | 0 | 1.24 | 0 | 0 | **98.7** |

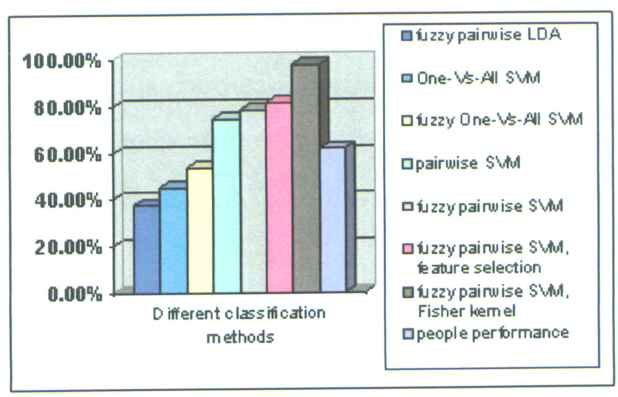


**Figure 4.5:** Flowchart of achieved accuracies for different methods.

used in signal detection theory to depict the tradeoff between hit rates and false alarm rates of classifiers (Egan, 1975; Swets et al., 2000). ROC analysis has been extended for use in visualizing and analyzing the behavior of diagnostic systems (Swets, 1988). One of the earliest adopters of ROC graphs in machine learning was Spackman (1989), who demonstrated the value of ROC curves in evaluating and comparing algorithms. Recent years have seen an increase in the use of ROC graphs in the machine learning community. ROC graphs are two-dimensional graphs in which True Positive (TP) rate is plotted on the $Y$ axis and False Positive (FP) rate is plotted on the $X$ axis. An ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives). Informally, one point in ROC space is better than another if it is to the northwest (TP rate is higher, FP rate is lower, or both) of the first. The common method of comparison between different ROC curves is to calculate the area under the ROC curve, abbreviated AUC. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between $(0, 0)$ and $(1, 1)$, which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. The corresponding ROC curves are presented in Fig. 4.6. As it shows the AUC for Fisher kernels is almost 1.0, which is an indication of their excellent performance.

In fact, sequence discriminant SVM (Fisher kernel) is the first time being used in the application of speech emotion recognition (it has been successfully used in other applications before [59][60]) and together with all other carefully-chosen methods in the rest of stages of the whole process, it gives a superior result (97.6%) in comparison with the previous works [25][26][65][27].

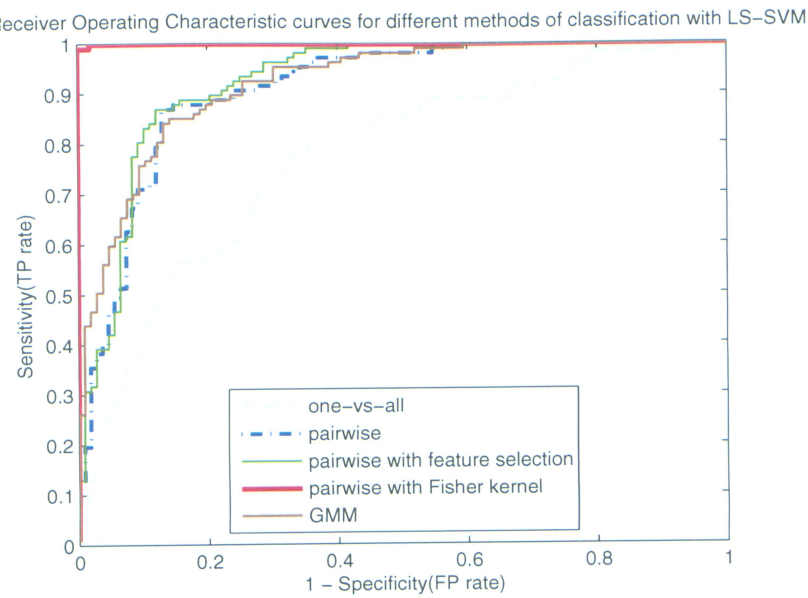


**Figure 4.6:** Receiver Operating Characteristic Curves for different methods of classification.

# Chapter 5

# Conclusions

## 5.1 Summary of the Thesis and Contributions

AFFECTIVE computing has grown an important field of research in today's man machine interaction. Applications span a wide range of fields going from entertainment and intelligent toys to emotion aware board computers in cars. Speech analysis is among the most promising information sources considering AER. While performance obtained by automatic systems based on this channel are among the most reliable ones, it is still not sufficient for usage in real-life scenarios. We therefore strive to bridge the gap between the commercially highly interesting multiplicity of potential applications and current accuracies.

In this thesis an AER system based on paralinguistic information of speech signals is presented aiming at enhancing human-computer interaction. The proposed system is independent of speakers and gender in order to offer more versatility. A genuine database with large number of data samples, which is exclusively created for the applications of automatic emotion recognition is used for training and testing the system. Six basic and universal emotions are being categorized throughout this research: anger, happiness, fear, disgust, surprise, and sadness. These basic emotions are said to be independent of cultural background.

As opposed to prosodic features used in most of the studies, a set of novel acoustic features has been proposed in this contribution most of which are being used for the first time in this application. The proposed set of acoustic features can be divided into three

main categories: spectral features, time-domain features, and cepstral features. In order to select the most effective set of features, the Sequential Forward Selective (SFS) method is employed with the goal of maximizing the performance accuracy criteria. By applying SFS algorithm, selection of the subset which yields the best performance is guaranteed. Rather than performing the feature selection algorithm on all the classes, it is performed on each pair of classes separately. It helps us to obtain more detailed insight into individual emotions and their corresponding significant features.

To categorize the extracted features to the six basic emotions, some sort of machine learning and classification method has to be adopted. While most people use the very common methods of classification such as Neural Networks or Hidden Markov Models, this research made use of the Least Squares Support Vector Machine (LS-SVM), which is a relatively novel and powerful method of classification. In essence two different approaches have been considered and experimented which are summarizes in the next two sections.

### Frame Approach

In the first approach after applying a Hamming window of length 23 ms with 50% frequency overlap in order to divide the speech signals into sequential frames which are considered stationary, all the features are extracted from each frame. Mean and standard deviation for each feature is computed in order to represent one feature vector for each speech utterance. To extend the binary LS-SVMs to our multi-category problem, four different coding and decoding strategies are implemented and compared: one-against-all, fuzzy one-against-all, one-against-one (pairwise), and fuzzy pairwise. For the purpose of comparative study we are also applying a Linear Classifier with gradient descent optimization algorithm.

Fifty percent of data samples was used for the training phase exclusively and the remaining 50% for evaluating the trained classifiers (the division is done in random). All the binary LS-SVM classifiers are trained using RBF kernel function with different optimal regularization and kernel parameters. The linear classifiers are trained using the gradient descent algorithm and perceptron criterion function.

### Sub-band Approach

In the second approach spectral features are extracted from six non-overlapping logarithmically scaled (according to Mel scale) frequency sub-bands rather than frames. Sub-band approach will provide better discrimination since for each emotion different energy distributions in different frequency sub-bands can be captured. Also because most of the important information in speech signals is located in the lower frequencies, the $6^{th}$ sub-band is dismissed, as a result more precision and also less computation complexity is gained.

In this approach Fisher kernels are adopted to have sequence discriminant SVMs in order to conquer the problem of variable-length feature sequences in speech signals and also to avoid an inevitable loss of data caused by computing the statistical parameters (as was performed in the first approach). Fisher kernel has been successfully used in the applications of biosequence analysis and classification, speech recognition, speaker verification, and acoustic event classification before, but it is the first time it is being employed in the application of emotion recognition. In order to compare the result of sequence discriminant SVMs, which is the combination of discriminative classifiers and generative models, with probabilistic classification, Gaussian Mixture Models are also applied to classify the different emotions.

Five-fold cross validation method is utilized to achieve the final result. To classify the mapped data points in the Fisher score space, all the binary LS-SVMs are trained using linear kernel functions with different optimal regularization parameters.

### Discussion of Results

Among different methods of multi-class SVM, fuzzy-pairwise method shows the best performance. In fact, since in pairwise method more binary classifiers construct the final result compared to one-vs-all method, there exists a trade off between accuracy and computation cost, where in this experiment the improvement of accuracy is much more potent. Also by using the best subset of features selected by SFS algorithm 2.9% improvement was achieved in the accuracy. The linear discriminant classifier shows a very poor performance, which

| Reference | Input to the system | Features | Classifier | Overall Accuracy |
|---|---|---|---|---|
| [10] | Acoustic features and language information | Statistics of prosodic features | ML-SVM and HMM | 81.2% |
| [11] | Acoustic information | F0, F1-F4, MBE1-MBE5, MFCC1-MFCC2 | SVM | 88.9% |
| [12] | Acoustic information | Prosodic features | Ensemble of NN classifier | 70% |
| [13] | Acoustic and textual content | Statistical values of F0, F1, energy, and ZCR | SVM | 81.4% |
| [25] | Acoustic content | Statistics of prosodic features | ANN | 73% |
| [26] | Acoustic information | Statistics of pitch and energy | HMM | 77.8% |
| This thesis | Acoustic information | Spectral features, MFCC, ZCR (Fig. 4.3) | Sequence discriminant SVM | 97.6% |

**Figure 5.1:** Comparison of some of the existing works with this work

could be due to the complexity of geometric distribution of the data points.

By upgrading our system to the sub-band approach and using Fisher kernel for sequence discriminant SVMs, 16.3% improvement in the best overall classification accuracy was achieved. The inferior classification rate achieved by GMMs shows the better performance of discriminative models at classification task compared to probabilistic models. In more detail, exploiting generative models in discriminative classifiers (Fisher kernel) outperforms both the generative model (GMM) and discriminative classifier (LS-SVM) per se.

The achieved classification rate (97.6%), which is a very promising and satisfying result, proves a very good choice of features along with very powerful employed methods for assigning those features to the corresponding emotions. A comparison of this work with some of the existing works are given in Fig. 5.1.

The sub-band approach has also become popular in recent years in speech recognition and speaker verification [74]. In addition to capture more detailed information, the main motivation has been to achieve robust recognition in the face of noise. It is often found that the sub-band approach delivers performance improvements (especially in the presence of noise) [74]. In other words, the system is robust in the case of speech corrupted by a noise affecting a limited number of sub-bands. A disadvantage of sub-band approach compared to frame-based approach could be its increased computational complexity.

Using spectral features makes the system faster and therefore more suitable for real-time applications, since they are easy to compute compared to prosodic features. Also because prosodic features represent speaker's habitual speaking style, including duration and pausing patterns, and intonation contours [75], they are relatively speaker-dependent. This is another disadvantage of using prosodic features in an AER system since we are avoiding any sort of subject dependency.

SVM is a good choice of classifier not only because of the excellent empirical results, but also because of the following advantages:

- Generalization capabilities in the high dimensional manifold are ensured by enforcing the largest margin classifier.
- Projection onto a high-dimensional manifold by means of kernel function is only implicit.
- There are few model parameters to select: the penalty term C and the parameter(s) of kernel function (e.g. $\sigma$ in the case of RBF kernel).
- The final results are stable and repeatable (i.e. no random initial values).
- SVM provide a method to control complexity independently of dimensionality.

## 5.2 Future Work

For future work, a combined time and frequency approach can be explored in order to take advantage of both frame-based and sub-band approach and take into account the temporal evolution of the speech signals.

One of the important issues in sub-band approach is choosing the number and locations of the frequency bands. It could be a future improvement to the system to achieve an optimal division of the frequency domain. Also as one of the benefits of sub-band approach is that different sub-bands can be processed separately, a good potential future work in this regard can be to investigate and localize the most emotional-dependant sub-bands. In other words, those sub-bands which carry the most emotional-relevant information can be detected, those information might then be emphasized\weighted to improve recognition.

As a potential future work, the intensity of emotions also could be considered in the system.

As the main drawback of LS-SVM is lack of sparseness (i.e. number of support vectors is equal to number of data points), one potential future research could be replacing LS-SVM with original SVM and investigating the resulting trade-off between accuracy and speed. Another alternative can be applying some pruning techniques to overcome this drawback.

Wrapper methods for feature selection, such as SFS algorithm which is adopted in this thesis, are very high in computation cost, especially when dealing with large data sets. This drawback might make a system impractical in on-line applications. Therefore, as a future work it would be beneficial to try other feature selection algorithms which are more time efficient in order to obtain a satisfying result between time and accuracy.

To extend this research and apply the reported methods to a real-world application, a potential future research is to apply the proposed techniques to recognize the emotional state in infants using their crying sounds as input information to the system.

# Appendix A

# Computing the Score Vectors for Fisher kernel

In this section, we derive the formulas for computing derivatives of the log likelihoods when the generative model is a diagonal covariance GMM.

Let

$$R(i,j) = \prod_{l=1}^{N_d} \frac{1}{\sigma_j^l \sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left( \frac{x_i^l - \mu_j^l}{\sigma_j^l} \right)^2 \right\} \tag{A.1}$$

so that the diagonal covariance GMM likelihood is

$$P(\mathbf{x}_i|M,\theta) = \sum_{j=1}^{N_g} a_j R(i,j) \tag{A.2}$$

where $\theta = \{a_j, \mu_j^l, \sigma_j^l\}$ is the set of parameters in the GMM, $M$. In particular, $a_j$ is the prior of the $j^{th}$ Gaussian component of the GMM, $\mu_j$ is the mean vector of the $j^{th}$ component, and $\sigma_j$ is the corresponding diagonal covariance vector. The superscript on the mean and covariance enumerate the components of the vectors. $N_g$ is the number of Gaussian components that make up the mixture model and $N_d$ is the dimensionality of the input vectors with components $f(\mathbf{x}_i) = [x_i^1, x_i^2, \ldots, x_i^{N_d}]$.

The global log likelihood of a sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_v}\}$ is

$$\log P(X|M,\theta) = \sum_{i=1}^{N_v} \log P(\mathbf{x}_i|M,\theta) \tag{A.3}$$

where $N_v$ is the number of frames in the sequence.

The score-vector is the vector of derivatives with respect to each parameter of A.3. The derivatives are with respect to the covariances, means, and priors of the Gaussian mixture model. The derivative with respect to the $j^{th}$ prior is

$$\frac{d}{da_{j^*}} \log P(X|M,\theta) = \sum_{i=1}^{N_v} \frac{R(i,j^*)}{\sum_{j=1}^{N_g} a_j R(i,j)} \tag{A.4}$$

The derivative with respect to the $l^{th}$ component of the $j^{th}$ mean is

$$\frac{d}{d\mu_{j^*}^{l^*}} \log P(X|M,\theta) = \sum_{i=1}^{N_v} \frac{R(i,j^*)}{\sum_{j=1}^{N_g} a_j R(i,j)} \cdot \frac{1}{\sigma_{j^*}^{l^*}} \left( \frac{x_i^{l^*} - \mu_{j^*}^{l^*}}{\sigma_{j^*}^{l^*}} \right) \tag{A.5}$$

The derivative with respect to the $l^{th}$ component of the $j^{th}$ covariance is

$$\frac{d}{d\sigma_{j^*}^{l^*}} \log P(X|M,\theta) = \sum_{i=1}^{N_v} \frac{R(i,j^*)}{\sum_{j=1}^{N_g} a_j R(i,j)} \cdot \left( \frac{(x_i^{l^*} - \mu_{j^*}^{l^*})^2}{(\sigma_{j^*}^{l^*})^3} - \frac{1}{\sigma_{j^*}^{l^*}} \right) \tag{A.6}$$

The likelihood score-vector can then be expressed as

$$\psi_{Fisher}(X) = \left[ \frac{d}{da_{j^*}}, \ldots, \frac{d}{d\mu_{j^*}^{l^*}}, \ldots, \frac{d}{d\sigma_{j^*}^{l^*}} \right]^T \tag{A.7}$$

for $j^* = 1, \ldots, N_g$ and $l^* = 1, \ldots, N_d$.

# Appendix B

# Steepest Gradient Descent Procedure for Optimization

Basic gradient descent is very simple. We start with some arbitrarily chosen weight vector $\mathbf{a}(1)$ and compute the gradient vector $J(\mathbf{a}(1))$. The next value $\mathbf{a}(2)$ is obtained by moving some distance from $\mathbf{a}(1)$ in the direction of steepest descent [22]. In general, $\mathbf{a}(k+1)$ is obtained from $\mathbf{a}(k)$ by equation

$$\mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k)\nabla J(\mathbf{a}(k)) \tag{B.1}$$

where $\eta$ is learning rate that sets the step size. We hope that such a sequence weight vector will converge to a solution minimizing $J(\mathbf{a})$. We should be careful with the choice of learning rate. If it is too small, convergence is needlessly slow, whereas if learning rate is too large, the correction process will overshoot and can even diverge.

# Bibliography

[1] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586-591, Maui, June 1991.

[2] J. J. Lien, T. Kanade, J. Cohn, and C. Li, "Automated facial expression recognition based on FACS action units," $3^{rd}$ *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 390-395, April, 1998.

[3] K. S. Byun, C. H. Park, and K. B. Sim, "Emotion recognition from facial expression using hybrid-feature extraction," *SICE Annual Conference in Sapporo*, August 44, 2004, Japan.

[4] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," *Proceedings of IEEE International Conference on Face and Gesture Recognition*, Mar. 2000, pp. 46-53.

[5] I. Kotsia and I. Pitas "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *In Proceeding of IEEE Transactions on Image Processing*, vol.16, no.1, January 2007.

[6] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," $3^{rd}$ *IEEE Int. Conference on Automatic Face and Gesture Recognition,* pp. 200-205, April 1998.

[7] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: Face

expression recognition," *IEEE Trans. Systems., Man, and Cybernnetics*, Part B, vol. 35, no. 3, pp. 477-488, June 2005.

[8] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[9] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no.1, pp. 259-275, 2003.

[10] B. Schüller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing,* vol.1, PP. I-577-80, 17-21 May, 2004.

[11] Y. L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," *Proceeding of International Conference on Machine Learning and Cybernetics*, Vol. 8, PP 4898-4901, 18-21 Aug. 2005.

[12] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," *Proceedings of the Six$^{th}$ International Conference on Spoken Language Processing*, October 2000.

[13] Z. J. Chuang and C. H. Wu, "Emotion recognition using acoustic features and textual content", *IEEE International Conference on Multimedia and Expo*, Vol. 1, PP. 53-56, 27-3- June 2004.

[14] I. S. Engberg and A. V. Hansen, "Documentation of the Danish Emotional Speech Database," *Technical report,* Center for PersonKommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark, September 1996.

[15] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," *Proceeding of IEEE International Conference on Acoustic, Speech, and Signal Processing,* Vol. 2, pp. 1085-1088, March 2005.

[16] Intel, "Intel open source computer vision library".

[17] C. Chen, Y. Huang, and P. Cook, "Visual/acoustic emotion recognition," *IEEE International Conference on Multimedia and Expo,* pp. 1468-1471, 6-8 July 2005.

[18] P. Ekman and W. V. Friesen, *Facial Action Coding System.* Palo Alto: Consulting Psychologist Press, 1978.

[19] T. Kanade, J. F. Cohn, and Y. L. Tian: "Comprehensive database for facial expression analysis", *Proceeding of* $4^{th}$ *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-53, March 2000.

[20] *http://www.irc.atr.jp/ mlyons/jaffe.html*

[21] A. M. Martinez and R. Benavente: "The AR Face Database," *technical report*, 1998.

[22] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 2933, 2000.

[23] E. Douglas-Cowie, R. Cowie, and M. Schrder: "A New Emotion Database: Considerations, Sources and Scope," *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 3944, 2000.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," *Proceedings of the* $22^{nd}$ *International Conference on Data Emgineering Workshop*, 3-7 April 2006.

[25] C. A. Martinez and A. B. Cruz, "Emotion recognition in non-structured utterance for human-robot interaction," *IEEE international workshop on robot and human interactive communication*, pp. 19-23, August 2005.

[26] B. Schüller, G. Rigoll, and M. Lang,"Hidden Markov model-based speech emotion recognition, " *Proceeding of the IEEE International Conference on Acoustic, Speech, and Signal Processing,* pp. I-401-04, April, 2003.

[27] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using Nueral Networks," *Proceedings of the 6th International Conference on Neural Information Processing*, Vol. 2, PP. 495-501, 1999.

[28] A. Ramalingam and S. Krishnan, "Gaussian mixture modelling using short-time Fourier transform features for audio fingerprinting," *Proceedings of International Conference on Multimedia and Expo*, pp. 1146-1149, July 2005.

[29] D. Hosseinzadeh and S. Krishnan, "Speaker recognition using spectral-based features," *EURASIP journal on Information Security*, 2006.

[30] T. Nguyen and I. Bass, "Investigation of combining SVM and decision tree for emotion classification," *seven$^{th}$ IEEE international symposium on multimedia,* pp. 540-544, 2005.

[31] V. Angelis, G. Felici, and G. Mancinelli, *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques.* Springer US, 2006.

[32] C. E. Izard, *Human Emotions.* New York: Plenum Press, 1977.

[33] T. J. Mayne and G. A. Bonanno, *Emotions: Current Issues ans Future Directions.* New York: Guilford Press, 2001.

[34] *http://en.wikipedia.org/wiki/Emotion*

[35] *http://face-and-emotion.com/dataface/emotion/theories.jsp*

[36] D. Goleman, *Emotional Intelligence: Why It Can Matter More Than IQ.* NewYork: Bantam Books, 1997.

[37] K. T. Strongman, *The Psychology of Emotion: From Everyday Life to Theory.* England: J. Wiley & Sons, 2003.

[38] J. A. Jacko and A. Sears, *Human-Computer Interaction Handbook: Fundamentals, Evolving Thechnologies, and Emerging Applications.* Lawrance Erlbaum and Associates, 2002.

[39] *http://en.wikipedia.org/wiki/Human-Computer-Interaction*

[40] R. W. Picard, *Affective Computing.* MIT Press, 1997.

[41] H. Stelmaszewska and B. Fields, "Emotion and technology: an empirical study," *Workshop on The Role of Emotion in Human-Computer Interaction*, London, September 2006.

[42] *http://robotic.media.mit.edu/projects/RoCo.html*

[43] *http://www.ultimatetoys.com.my/New/Furby/Furby.htm*

[44] *http://www.sonydigital-link.com/aibo/index.asp*

[45] *http://paro.jp/english/*

[46] S. Mozziconacci,"Prosody and emotions," Proceedings of Speech Prosody, France, April 2002.

[47] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* NJ: Prentice Hall, 1999.

[48] D. Morrison, R. Wang, L. C. De Silvia, and W. L. Xu, "Real-time spoken affect classification and its applications in call-centres," *Proceedings of the* $3^{rd}$ *Internationl Conference on Information Technology and Applications*, Vol. 02, pp. 483-487, 2005.

[49] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, Vol. 01, pp. I-593-6, May 2004.

[50] R. O. Duda, P. E. Hart and D. G. Strok, *Pattern classification.* USA: Wiley, 2001.

[51] N. Cristianini and J. SH. Taylor, *An introduction to Support Vector Machines and other kernel-based methods.* United Kingdom: Cambridge University Press, 2000.

[52] C. Unsalan and A. Ercil, "Shapes of features and a modified measure for linear discriminant analysis," *Proceedings of* $15^{th}$ *International Conference on Pattern Recognition*, Vol. 2, pp. 410 - 413, Sept 2000.

[53] W. Y. Zhao, "Discriminant component analysis for face recognition," *Proceedings of* $15^{th}$ *International Conference on Pattern Recognition*, vol. 2, pp. 818-821, Sep. 2000.

[54] M. Pechenizkiy, S. Puuronen, and A. Tsymbal, "Feature extraction for classification in knowledge discovery systems," *Technical Report*, Department of Computer Science and Information Systems, University of Jyväskylä, Finland.

[55] D. Tsujinishi, Y. Koshiba, and SH. Abe, "Why pairwise is better that one-against-all or all-at-once," *Proceedings of IEEE International Conference on Neural Networks*, vol. 1, pp. 693-698, July 2004.

[56] C. J. Burges, "A tutorial on support vector machine for pattern recognition," *Knowledge Discovery and Data Mining*, Vol. 2, pp. 121-167, June, 1998.

[57] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least square support vector machines.* Singapore: World Scientific Publishing Co. Pte. Ltd., 2002.

[58] J. A. K. Suykens, L. Lukas, and J. Vandewalle, "Sparse Least Squares Support Vector Machine Classiffiers," *Proceedings of European Symposium on Artificial Neural Networks,* pp. 37-42, Belgium, April 2000.

[59] T. S. Jaakkola and Haussler, "Exploiting generative models in discriminative classifiers," *Advances in Neural Information Processing Systems*, 1998.

[60] A. Temko, E. Monte, and C. Nadeu, "Comparison of sequence discriminant support vector machines for acoustic event classification," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 5, pp. v-721-724, May 2006.

[61] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp. 381-396, March 2002.

[62] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, Vol. AC-19, No. 6, December 1974.

[63] C. Becchetti and L. P. Ricotti, *Speech Recognition: Theory and C++ Implementation.* Toronto: John Wiley and Sons, 1999.

[64] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition us- ing multimodal information", *Proceedings of IEEE International Conference on Information, Communications and Signal Processing*, vol. 1, pp. 397-401, Singapore, September 1997.

[65] V. A. Petrushin, "Creating emotion recognition agents for speech signal, " *in Socially Intelligent Agents*, K. Dautenhahn, A. H. Bond, L. Canamaro, and B. Edmonds (eds.), Kluwer Academic Publishers, pp. 77-84, 2002.

[66] H. Seddik, A. Rahmouni, and M. Sayadi, "Text independent speaker recognition using the Mel frequency cepstral coefficients and a neural network classifier," *First International Symposium on Control, Communication, and Signal Processing*, pp. 631-634, 2004.

[67] R. B. Sória and E. F. cabral Jr., "Speaker recognition with artificial neural network and mel-frequency cepstral coefficients correlations," *In Proceedings of Eight European Signal Processing Conference*, pp. 1051-1054, Italia, Sep. 1996.

[68] F. Gouyan, F. Pachet, and O. Delerue, "On the use of zero-crossing rate for an application of classification of percussive sounds," *In Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00),* Verona, Italy, Dec. 2000.

[69] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line Handwriting Recognition using Support Vector Machines - A kernel approach," *Proceedings of the* $8^{th}$ *International Workshop on Frontiers in Handwriting Recognition*, pp. 49-54, Korea, August 2002.

[70] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussianmixture speaker models," *IEEE Transactions on Speech and Audio Processing,* Vol. 3, pp. 72-83, Jan 1995.

[71] M. N. Stuttle and M. J. F. Gales, "A mixture of Gaussians front end for speech recognition," In Proceeding of Eurospeech, Denmark, September 2001.

[72] V. Faber, "Clustering and the continuous k-means algorithm," *Technical Report,* 1994.

[73] R. I. Damper and J. E. Higgins, "Improving speaker identification in noise by subband processing and decision fusion," *Pattern Recognition Letters*, Vol. 24, Issue 13, pp. 2167-2173, Sep. 2003.

[74] S. Kajarekar *et al.*, "Speaker recognition using prosodic and lexical features," *IEEE Workshop on Automatic Speech Recognition and Understanding*, Vol. 3, pp. 19-24, Dec. 2003.

[75] R. Cowie, E. Douglas-Cowie , J. G. Taylor , S. Ioannou, M. Wallace, and S. Kollias, "An intelligent system for facial emotion recognition," *IEEE International Conference on Multimedia and Expo*, Amsterdam, July 6-8, 2005.

[76] Y. Wang, "Recognizing human emotional state from audiovisual signals," *Master's thesis*, Ryerson University, Toronto, 2005.