Wearable and Automotive Systems for Affect Recognition from Physiology

by

Jennifer A. Healey

Submitted to the Department of Electrical Engineering and
Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

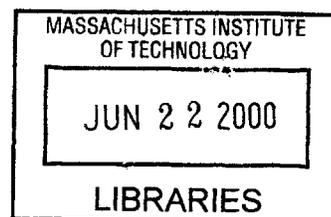MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2000
[June 2000]

Author . . . .
Department of Electrical Engineering and Computer Science
May 18, 2000

Certified by . . . .
Rosalind W. Picard
Professor MIT Media Laboratory
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Wearable and Automotive Systems for Affect Recognition from Physiology

by

Jennifer A. Healey

Abstract

Novel systems and algorithms have been designed and built to recognize affective patterns in physiological signals. Experiments were conducted for evaluation of the new systems and algorithms in three types of settings: a highly constrained laboratory setting, a largely unconstrained ambulatory environment, and a less unconstrained automotive environment. The laboratory experiment was designed to test for the presence of unique physiological patterns in each of eight different emotions given a relatively motionless seated subject, intentionally feeling and expressing these states. This experiment generated a large dataset of physiological signals containing many day-to-day variations, and the proposed features contributed to a success rate of 81% for discriminating all eight emotions and rates of up to 100% for subsets of emotion based on similar emotion qualities. New wearable computer systems and sensors were developed and tested on subjects who walked, jogged, talked, and otherwise went about daily activities. Although in the unconstrained ambulatory setting, physical motion often overwhelmed affective signals, the systems developed in this thesis are currently useful as activity monitors, providing an image diary correlated with physiological signals. Automotive systems were used to detect physiological stress during the natural but physically driving task. This generated a large database of physiological signals covering over 36 hours of driving. Algorithms for detecting driver stress achieved a recognition rates of 96% using stress ratings based on task conditions for validation and 89% accuracy using questionnaires analysis for validation. Further results in which metrics of stress from video tape annotations of the drive were correlated with physiological features showed highly significant correlations (up to $r = .77$ for over 4000 samples). Together, these three experiments show a range of success in recognizing affect from physiology, showing high recognition rates in somewhat constrained conditions and highlighting the need for more automatic context sensing in uncon-

more automatic context sensing in unconstrained conditions. The recognition rates obtained thus far lend support to the hypothesis that many emotional differences can be automatically discriminated in patterns of physiological changes.

Thesis Supervisor: Rosalind W. Picard
Title: Professor MIT Media Laboratory

Contents

List of Figures

11

List of Tables

15

Chapter 1

Introduction

There is a movement in computer science toward developing systems that learn what their users want and that try to model their user's interests and respond in a more adaptive way. Currently, methods of modeling user preferences and frustrations involve active non-social interactions, such as clicking on menus and creating preference lists; however, the natural way people communicate and respond to satisfaction or dissatisfaction is through affective expression. To appear socially intelligent, computers will have to develop a model of their user's emotional state and respond to that state appropriately. This affective intelligence becomes more important as computers become more ubiquitous. A natural, social interaction with a spreadsheet or programming task might seem superfluous, but computers will soon be everywhere, in our homes, assisting with cooking, heating, and room ambiance, in our cars, controlling communication, navigation and music selection and even in our clothing, extending our senses, jogging our memories in appropriate contexts and perhaps broadcasting messages expressing our personality.

One of the main thrusts of engineering has been to model natural phenomena. Since the time of William James in 1890, it has been hypothesized that emotion could be modeled according to the unique patterns of physiological signals. This hypothesis has been challenged and defended many times and most recently it has been put forth that more sophisticated feature combination methods would be the key to finding these patterns[CT90]. The research of this thesis represents a contribution to the

field of electrical engineering by demonstrating a new application of signal processing and pattern recognition techniques to the problem of emotion recognition, and by showing promising results in this analysis. Novel prototype systems for gathering affective data are developed and results show how aspects of a user's affective state can be recognized.

Research in the area of computer emotion understanding is just beginning to develop. Efforts exist focusing different affective channels such as the analysis of voice, facial expression and profiling behavior. This work explores the use of physiological signals for detecting affect. Although the best overall detection method may incorporate several of these modalities, affect detection through physiology offers a private, continuous signal for computer human interaction that is free of the artifact of social masking. Physiological signals have not been traditionally considered as an input channel for a computer, but as computers move beyond desktop applications and become integrated into clothing and vehicles the opportunity for greater physical contact between user and machine increases. This thesis explores the extent to which affect can be detected through physiology in various environments: in the office, while walking around performing daily activities and while driving a car. To study these situations, physiological sensors were integrated into several new prototype systems using wearable computers and an automobile as platforms.

The automotive experiment focused on recognizing the affective state of stress. Reducing stress is an important factor in disease prevention and recovery. Chronic health problems such as back pain and migrane headaches are negatively impacted by stress and reduced stress has been linked to faster recovery from serious diseases such as cancer[Sel56]. Automobile driving provided a natural situation in which stress could be evaluated using task design, driver questionnaires and ground truth annotations from video recordings. The stress detection provided by the automotive system could be used to help manage on board information appliances in situations of high stress or it could be used to track a persons stress level in a long term way by measuring stress on a daily commute over many days.

The systems developed and tested in this thesis show that affect can be recog-

19

nized through physiology, given certain constraints on the situation, providing a new method of interaction which can enrich the effectiveness of computer human interaction especially in new mobile platforms. The following chapters provide a background for this research, a detailed description of the three sets of experiments and the conclusions from this research effort. The chapters are largely self contained and the reader may advance to the sections of greatest interest in any order. Chapter 2, the background chapter, can be referenced at any time for a detailed description of the physiological sensors, an overview of the models of emotion used for labeling and a description of prior related work. Chapter 3 describes the Sentics experiments, designed to recognize eight unique emotional states intentionally felt and expressed by a subject seated in an office environment. For this experiment a recognition rate of 81% was achieved for eight intentionally expressed emotions by researchers in our group using the features presented here and rates of up to 100% were achieved for subsets of the eight emotions sharing similar emotion qualities. Chapter 4 introduces the new prototype designs built for ambulatory physiological monitoring. Systems incorporating cameras and wireless devices are shown to be useful for measuring physical activity. Chapter 5 presents the driver stress detection experiment. The results of this experiment demonstrate that driver stress can be recognized with 96% accuracy using five minute segments of data and a task based metric for validation and that a 89% recognition rate can be achieved using one minute segments of data and a questionnaire based metric for validation. Additional analysis shows that highly significant correlations (up to $r = .77$ for over 4000 samples) exist between physiological features and a metric of observed stressors obtained from a second by second annotation of video tape records of the drives. Chapter 6 discusses conclusions, application and directions for future work. The driver self-report questionnaire and the matlab code used for startle detection and linear classification are included as appendices.

Chapter 2

Background

This background provides a description of the physiological sensors used to collect the data, the models of emotion used to classify the data and presents selected examples of of prior work in detecting emotion from physiology. The first section presents the five different types of sensors used in this research: the electromyogram (EMG) for measuring muscle activity, the galvanic skin response or electrodermal response (GSR) for measuring sweat gland activity, the hall effect sensor for measuring respiration through chest cavity expansion (RESP), the photoplethysmograph for measuring blood volume pulse (BVP) and the electrocardiograph (EKG) for measuring heart rate and heart rate variability. The following section introduces the frameworks used for modeling emotion. These models are used to establish labels for the physiological data and to argue for an association between emotional stress and the axis of emotional arousal.

The final section presents examples of previous experiments designed to detect emotion from physiology. These experiments include laboratory experiments under very controlled conditions and field studies of ambulatory subjects, airplane pilots and automobile drivers.

2.1 Physiological Sensors and Signals

Five sensors were chosen to measure physiological signals that could give a continuous electronic reading to a computer and that were minimally invasive to the user. These sensors measured skin conductance, heart activity, respiration and muscle activity. Two different sensors were used to measure heart activity, the blood volume pressure (BVP) sensor and the electrocardiograph (EKG). Blood pressure and electroencephalogram (EEGs) readings were not used, but are briefly described for potential use in future experiments. More direct methods of autonomic nervous system (ANS) activation such as blood or salivary analysis of adrenalines, corticoids, or ACTH[Sel56] were not used because they require off-line chemical analysis. The goal of the systems designed in this thesis is to provide a continuous digital signal reflecting physiological variables to a computer system for analysis. The sensors described here can all meet the criteria of providing a continuous electronic output while also being minimally inconvenient to the users' normal activities.

2.1.1 Skin Conductance

The skin conductance sensor measures changes in the resistance of the skin caused when glands in the skin produce ionic sweat. The resistance of the skin is usually large, approximately $1M\Omega$; however, momentary changes in the level of the sweat gland activity causes changes in resistance (ΔR up to approximately 950K Ω) that can be measured by passing a small electrical current across two electrodes placed on the surface of the skin[SF90]. This measurement was taken in two locations, on the palm of the hand and the sole of the foot. K-Y Jelly was used as a low-conductivity gel to assure good contact between the skin and the sensor. The gel was applied on the electrode before the electrode was placed on the skin. For the reading of skin conductance across the hand, the electrodes were placed on the middle of the three segments of the first and middle finger, on the side of the palm. For the measurement on the foot the electrodes were placed on the sole of the foot at both ends of the arch of the foot. The skin conductance signal was sampled at 20 samples per second in
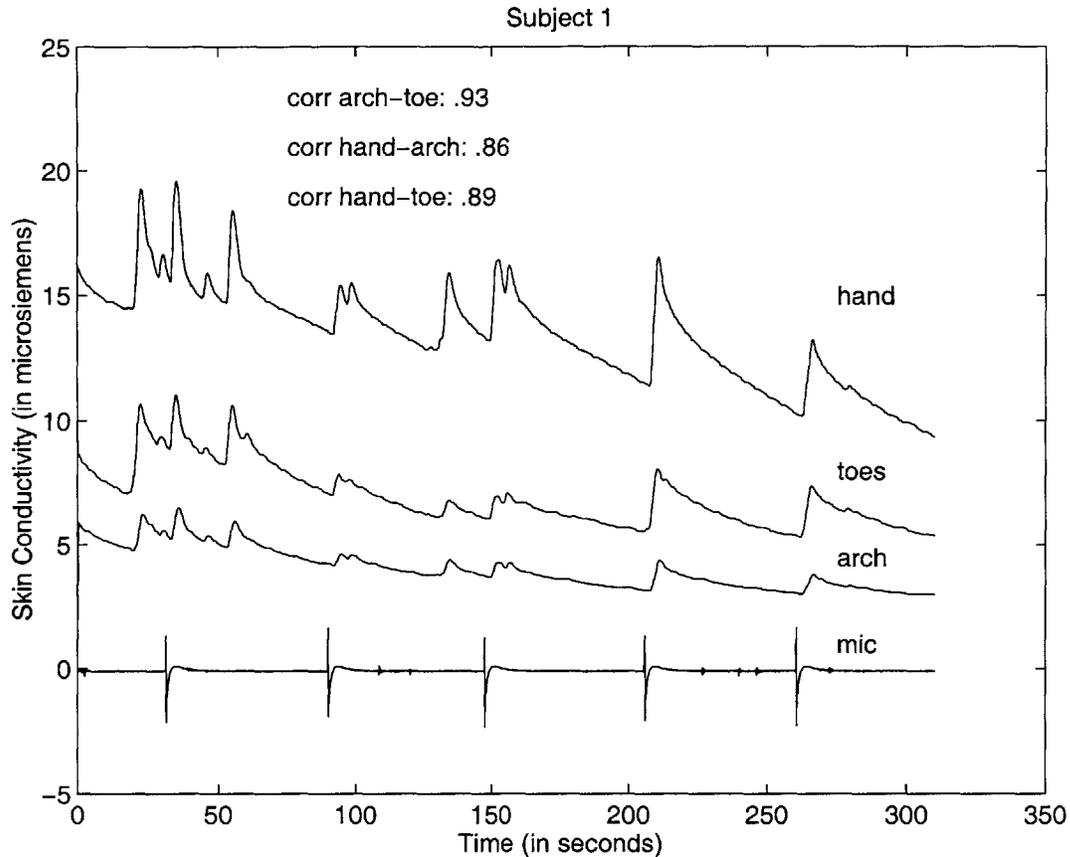
Figure 2-1: Three simultaneous skin conductance readings taken on the hand and the toes and arch of the foot. A noise burst, indicated by the microphone trace was used to stimulate phasic responses. All three traces are highly correlated.

both the eight emotion recognition experiment and the wearables experiments and sampled at 31 samples per second in the experiments for the detection of stress in automobile drivers. The difference in sampling rates is due to the difference in the rates available on the two systems.

This skin conductance reading has two components, a tonic baseline level and short term phasic responses superimposed on the baseline level. The phasic response has many names including the electrodermal response (EDR), the skin conductance startle response, the skin conductance orienting response, skin resistance response (SRR) and skin conductance response (SCR). The terms resistance response and conductance response reflect that the response can be measured either by measuring the resistance or conductance of the skin. The orienting response is a more general

term for the response. The term startle response is typically used to describe a response to more extreme stimuli. This response may occur whenever a person is forced to attend to a change in either their external environment, such as a sudden sound or a change in lighting, or their internal environment, such as when formulating mental plans or when having thoughts of expectation[Dam94].

Figure 2-1 shows examples of skin conductance readings taken at three locations: taken across the hand, across the arch of the foot and off the second and third toes of the foot. Phasic responses are stimulated by 100 ms white noise bursts, which are recorded by a microphone sensor shown in the bottom trace. Unstimulated responses also occur, such as those that occur before the first white noise burst. The magnitude of the responses does not vary consistently with the magnitude of the stimuli. Part of the reason for this variation is the phenomena of habituation in which the subject does not react as strongly to a repeated stimuli as to a novel stimulus.

The electrodermal response is typically scored using a subset of the following features: latency, amplitude, rise time and the half-recovery time. The half recovery time is used because the full recovery time is difficult to determine. The diagram in Figure 2-2 shows an ideal response to a hypothetical stimulus. As this diagram shows, the response occurs a few seconds after the stimulus. The latency is the amount of time between stimulus and the onset of the rise of the response. The amplitude of the response is the difference between the peak of the response and the baseline. The duration of the response is the difference between the time of the response onset and the time of the peak. The half recovery time difference between the time of the peak and the time at which the response decays to one half of the magnitude of the peak. This single response is referred to as a Type 1 response[Bou92].

Scoring ambiguities arise when responses are not well separated, such as when responses can also occur on the recovery (type 2) or rise (type 3) of previous responses. These types of occurrences are shown in the diagram from Boucsein in Figure 2-3 [Bou92]. Three methods of scoring multiple responses, labeled A, B and C are shown. Method A models the decay of the response as an exponential and measures the magnitude of the response from the modeled height of the decay of the first

24
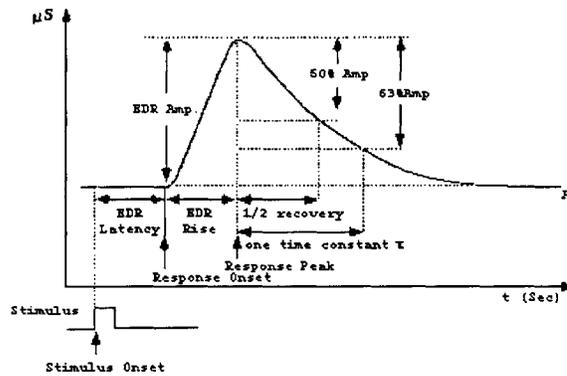
Figure 2-2: The electrodermal response (EDR) shown in this diagram from Boucsein[Bou92], has been measured according to several different features. This diagram shows an ideal response to a hypothetical stimulus. The response occurs a few seconds after the stimulus.

response at the time of the onset of the second response to the peak of the following response. Method B establishes a local baseline at the level of the onset of the second response and measures the distance from that baseline to the following peak. Method C measures all responses as the difference from the local peak to a global baseline. The automatic detection algorithm presented in Chapter 5 is based on Method B.

Skin conductivity has been found to be one of the most robust non-invasive physiological measures of autonomic nervous system activity[CT90]. Laboratory studies, such as those by Lang Ekman[ELF83], Levenson[Lev92] and Winton, Putnam and Kraus[WPK84] have found that skin conductivity response varies linearly with arousal ratings[Lan95]. Skin conductance measurements have been also been used to differentiate between states such as anger and fear[Ax53] and between states of conflict and no conflict[Kah73]. Skin conductance has been used as a measure of stress in anticipatory anxiety studies including studies of feared electric shock and studies of public speaking. This measure has also been studied in tasks such as driving and piloting aircraft[Bou92].

EDR type 2, evaluation method A

EDR type 3, evaluation method C

EDR type 2, evaluation method B

EDR type 3, evaluation method B
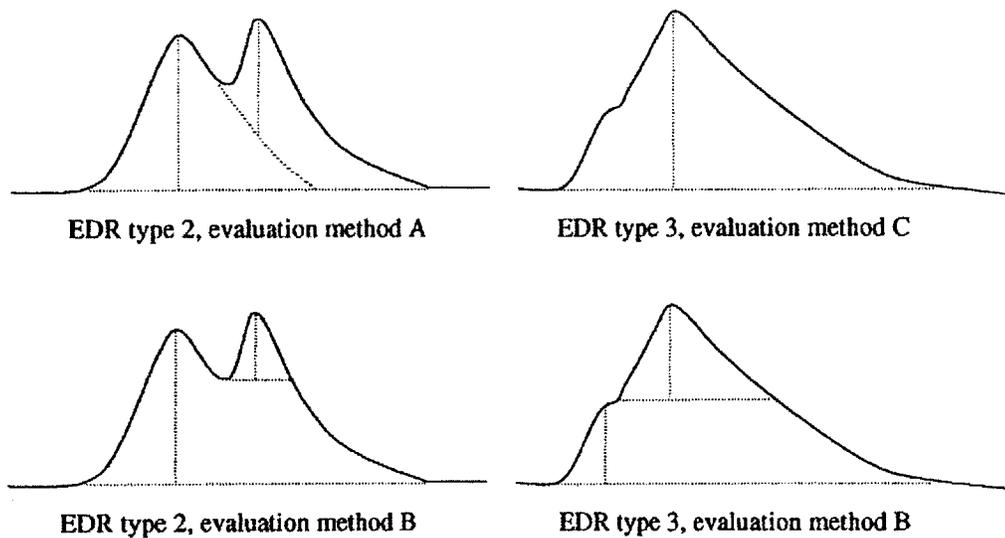
Figure 2-3: Startle responses can also occur on the rise of previous startle responses. Two examples, labeled type 2 and 3 are shown here in this diagram from Boucsein[Bou92]. Three methods of scoring multiple responses, labeled A, B and C are shown. Method A models the decay of the response as an exponential and measures the magnitude of the response from the modeled height of the decay of the first response at the time of the onset of the second response to the peak of the following response. Method B establishes a local baseline at the level of the onset of the second response and measures the distance from that baseline to the following peak. Method C measures all responses as the difference from the local peak to a global baseline.

2.1.2 BVP and Electrocardiograph

Heart rate and heart rate acceleration have been used as measures of overall physical activation [EG88] and changes in heart rate have been reported as indicators of fear[Lev92] and anger[Lev92][Kah73]. Blood volume pulse (BVP) and electrocardiograph (EKG) are both used to measure heart activity. The BVP sensor is a photoplethysmograph which measures light (infra-red or red) reflected from the skin. It is placed on the surface of the skin and does not require adhesives or gels. From the reflectance reading, the BVP can measure heart rate and vasoconstriction, however heart rate measured through this method is not precise enough to use for determining heart rate variability and it is subject to many placement and motion artifacts. The EKG is able to give a precise estimate of instantaneous heart rate by detecting sharp R-wave peaks, however, this sensor requires more effort to apply.

Blood Volume Pulse

A back-scatter photoplethysmograph is used to measure blood volume pulse. This device emits light and measures the amount of light reflected by the surface of the skin. After every heart beat, blood is forced through the blood vessels, producing an engorgement of the peripheral vessels under the light source and modifying the amount of light reflected to the photosensor[Tho94]. Therefore, the reflectance gives a relative measure of the amount of blood in the capillaries from which heart rate (pulse) and vasoconstriction can be derived. An example of the BVP signal is shown in Figure 2-4. The pulse train indicates the heart beats and shape of the envelope indicates the relative constriction of the blood vessel. This example shows twelve pulses and gradual vasoconstriction. Heart rate can be calculated by measuring the distance between successive pulse peaks. The heart rate from BVP was automatically calculated by the ProComp software in these experiments[Tho94]. This signal was sampled at 20 samples per second for all experiments. Vasoconstriction is a defensive reaction[Kah73] in which peripheral blood vessels constrict. This phenomena increases in response to pain, hunger, fear and rage and decreases in response to quiet relaxation[Fri86].

27
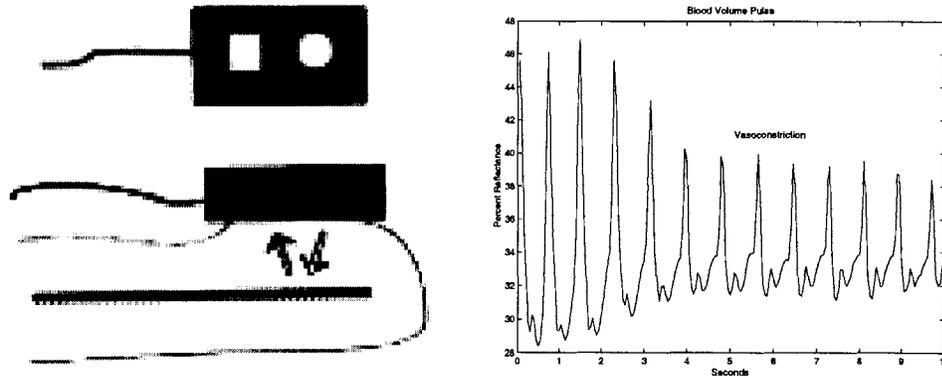
Figure 2-4: Blood volume pulse measures the amount of light reflected by the skin. This gives a measure of both the overall constriction of the blood vessel as determined by the envelope of the signal and a measure of the heart rate as determined by the pulse train. The figure on the left shows the BVP sensor [Tho94]. The example signal from this sensor (right) shows increasing vasoconstriction.

The BVP sensor can be placed anywhere on the body where the capillaries are close to the surface of the skin, but peripheral locations such as the fingers are recommended for studying emotional responses[Tho94]. In the experiments presented in Chapters 3 and 4, the BVP was placed on the tip of the ring finger. The BVP sensor requires no gels or adhesives, however the reading is very sensitive to variations in placement and to motion artifacts.

Electrocardiograph

The electrocardiograph measures heart activity by detecting voltages on the surface of the skin resulting from heart beats. The skin was prepared by using alcohol as a cleanser and Electro-Trace pre-gelled were applied. A modified lead II electrocardiogram was used to minimize motion artifacts and to produce a rhythm trace with sharp R-waves. In this configuration, shown in Figure 2-5 two signal electrodes are placed across the heart. The negative electrode is placed just to the right of the sternum near the base of the clavicle. The positive electrode is placed on one of the floating ribs underneath the left armpit. A third ground electrode is placed at a laterally symmetric position on the right floating ribs. An example of the signal
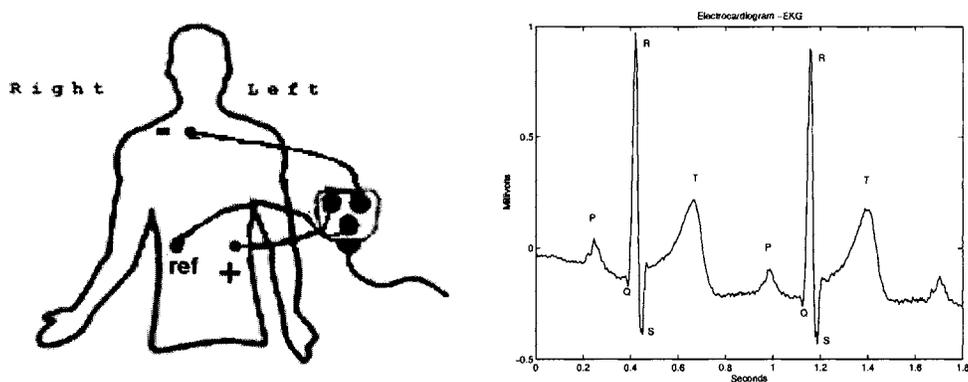
Figure 2-5: An EKG was applied in a modified lead II configuration to minimize motion artifacts and to attain a good record of the "R" wave. The EKG trace, right, shows the P wave, QRS complex and T wave(left). Detection of successive R wave peaks is used to calculate inter-beat intervals.

recorded by this configuration of the electrocardiogram is shown in Figure 2-5 with the characteristic P, Q, R, S and T segments are labeled.

Heart Rate Variability

Heart rate variability (HRV) has been measured in both the time and frequency domains. The term HRV often refers to the time series measure of the standard deviation of heart periods within the recording epoch[Bea97]. This is considered a good measure of short term variation (STV) in the electrocardiograph[vRAKea93]. Other time series measures have been used to assess long term variation of the spectrum including the difference between the maximum and the minimum R-R interval length in the window[vRAKea93] and the percent differences between successive normal R-R intervals that exceed 50 msec and the root mean square successive difference[KF98]. Recently, with the availability of digital recording devices and signal processing algorithms, short-term power spectral density analysis of the heart rate has become more popular as a method for assessing heart rate variability. In the spectral domain, the relative strengths of the sympathetic and parasympathetic influence on HRV can be discriminated. A ratio of the low frequency energy to the high frequency energy in the spectrum represents a ratio of the sympathetic to parasympathetic influence on
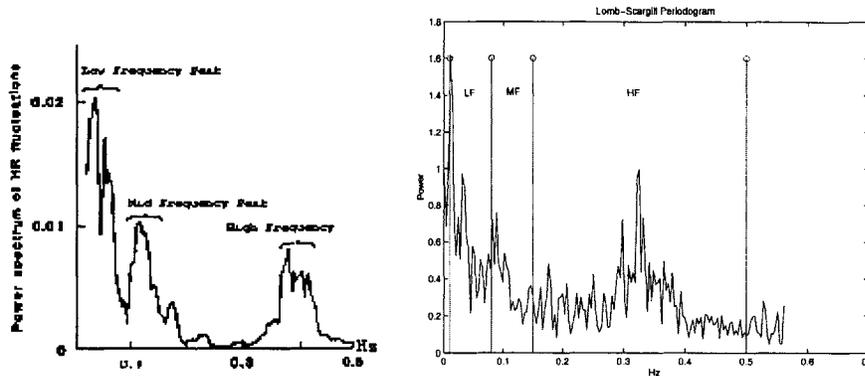
Figure 2-6: The power spectrum of the heart rate has three distinct peaks, one in the lowest frequencies under 0.1Hz, one near 0.1Hz and another in the higher frequencies between 0.3 and 0.5Hz. The diagram left shows the three peaks with the labels given by Akselrod [AGU⁺81] and the right shows an example of a spectrogram calculated using the digital signal from the car experiments.

the heart.

The power spectrum of the heart rate has three distinct peaks, one in the lowest frequencies under 0.1Hz, one near 0.1Hz and another in the higher frequencies between 0.3 and 0.5Hz. [AGU⁺81] [KF98]. The first two peaks are described alternately as very low and low frequency or low and medium frequency. The term low frequency is preferred by Akselrod(0.02-0.08Hz)[AGU⁺81], McCraty (0.01-0.08Hz)[MAea95] and Aasman (0.02 to 0.06Hz)[AMM87] and Very Low Frequency by van Ravenswaaij-Arts(<0.05Hz)[vRAKea93]. Energy in this region of the heart rate spectrum (below 0.05Hz) has been linked to circulation, vasomotor control and temperature control.[KF98]. This region of the spectrum is influenced by both sympathetic and parasympathetic branches of the ANS[Bea97].

The frequency band encompassing the 0.1Hz peak has been described as low frequency by Kamath (0.06Hz-0.15Hz)[KF98], Itoh(0.04-0.15Hz)[ITea95] and vanRavenswaaij-Arts(<0.05Hz)[vRAKea93] and medium frequency by Akselrod(0.1-0.15)[AGU⁺81], McCraty (0.08-0.15Hz)[MAea95] and Aasman (0.06-0.14Hz)[AMM87]. This 0.1Hz peak is associated with baroreceptor-mediated blood pressure control[KF98]. There are synchronous fluctuations in the blood pressure are called Mayer waves at this
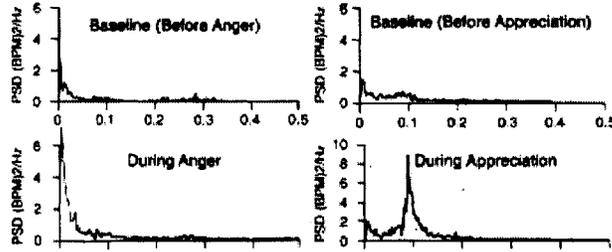
Figure 2-7: Different portions of the low frequency spectrum have been used to differentiate between the emotions anger and appreciation [MAea95].

frequency. They are increased when sympathetic tone is increased[vRAKea93].

The high frequency band, generally agreed to range from 0.15 to at most 0.6 Hz reflects Respiratory Sinus Arrhythmia (RSA). Due to inspiratory inhibition of the vagal tone, the heart rate shows fluctuations with a frequency equal to the respiration rate[vRAKea93]. Respiratory-frequency rhythms in autonomic nerves are translated into changes in discharge frequency of the Sino-Atrial (SA) node. RSA is mediated predominantly by parasympathetic influences on the sinus node and is often employed as an index of vagal control[Bea97].

Sympathetic and Parasympathetic Influence

The parasympathetic nervous system is able to modulate heart rate effectively at all frequencies between 0 and 0.5 Hz, whereas the sympathetic system modulates heart rate with significant gain only below 0.1Hz[Bea97]. Sympathetic activity increases heart rate while parasympathetic activity decreases heart rate. The integrated response depends on sympathetic and parasympathetic balance[SSea93].It has been proposed that a ratio of low frequency to high frequency portions be used as measure of sympathovagal balance, however, some researchers suggest using 0.02-0.15Hz [Bea97] as the low frequency band while others suggest using 0.08-0.15Hz as the low frequency band[MAea95]. McCraty suggests "anger" can be identified by the $\frac{LF}{HF}$ ratio and that "appreciation" can be identified using the mid-range frequencies, however the emotions he measured were generated by subjects following special training techniques[MAea95].

31

Confounding Variables

Several variables influence HRV which are not necessarily related to emotion. These include age, posture, level of physical conditioning, breathing frequency[vRAKea93] and circadian cycle[Bea97]. As age increases, heart rate variability decreases. Infants have a high sympathetic activity but this decreases quickly between ages 5 and 10[vRAKea93]. HRV is enhanced in the upright position due to an increased sympathetic tone. RSA is augmented in the supine position[vRAKea93]. Patients with certain medical conditions also have altered HRV. A predominance of sympathetic activity and a reduction in parasympathetic cardiac control has been found in patients with acute myocardial infraction[vRAKea93]. And patients with essential hypertension have less (vagally mediated) respiratory sinus arrhythmia and more baro-reflex related (0.1 Hz) heart rate variability when compared with normotensive controls[vRAKea93].

Meaningful analysis of heart rate variability is dependent on the integrity of the basic cardiac signal input[Bea97]. Analog recording systems can have errors due to inconsistencies in tape speed[vRAKea93] while digital recordings must be sampled at a sufficiently high rate to capture the R-wave peaks. The suggested rate varies from 250Hz [KF98] to 500-1000Hz. In the driver stress detection experiments this signal was sampled at 496 samples per second.

Finally, the statistics of the heart rate time series over which the power spectrum is computed is assumed to be stationary. This assumption is true for long time windows for resting subjects and in this case taking the power spectrum of the heart rate over a long heart rate time series can give a more precise estimation. However, in cases where the subject is actively engaged in a task such as when HRV is measured during arithmetic tasks or during active tasks such as driving, the stationarity of the heart rate time series is not assured. In these cases a shorter time window may provide a more accurate estimation of the spectrogram. Windows as short as 120 seconds are used in such studies[KF98] and some authors suggest using windows as short as 30 seconds if the EKG recording is free of missed beats and motion artifacts[vRAKea93].

Respiration through Chest Cavity Expansion

Breath Cycle
|------|

Percent Stretch

Baseline Stretch

Seconds
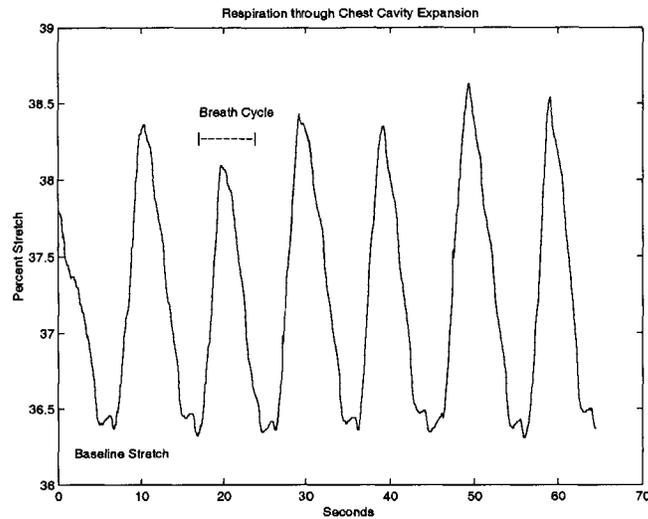
Figure 2-8: The respiration signal shows how the circumference of the diaphragm changes during breathing. The sensor is initially fastened with some degree of baseline stretch, then the inhalation and exhalation cycle causes the magnets to be first further separated then returned to the original position.

2.1.3 Respiration

Respiration is most accurately recorded by measuring the gas exchange of the lungs, however this method inhibits activities such as talking and driving in the natural environment. As an alternate measure, chest cavity expansion is recorded to capture breathing activity. A hall effect sensor comprised of two magnets embedded inside an elastic tube is used to measure inhalation and exhalation. Inhalation stretches the elastic, separating the magnets and creating a current and exhalation allows the sensor to return to the baseline state. An example of steady breathing as measured by this sensor is shown in Figure 2-8, showing breath cycles superimposed on the baseline stretch. This signal was sampled at 20 samples per second in the eight emotion experiment and wearables experiments and 31 samples per second in the automotive system.

Both physical activity and emotional arousal are reported to cause faster and deeper respiration, while peaceful rest and relaxation are reported to lead to slower and shallower respiration[Fri86]. Sudden, intense or startling stimuli can cause a momentary cessation of respiration and negative emotions have been reported to

cause irregularity in respiration patterns[Fri86]. The respiration signal can also be used to assess physical activities such as talking, laughing, sneezing and coughing.

2.1.4 Electromyogram

The electromyograph (EMG) measures muscle activity by detecting surface voltages that occur when a muscle is contracted. Three electrodes are used for this measurement, two are placed along the axis of the muscle of interest and a third ground electrode is placed off axis. Triode electrodes from Thought Technology were used with 10-20 high-conductivity gel in all experiments. This sensor was used to measure jaw clenching in the experiments in Chapter 3 and used to measure upper back (trapezius) tension in the experiments in Chapters 4 and 5. Electromyogram has been used to study facial expression[Ekm90], gestural expression[MP98], emotional valence[Lan95] and emotional stress[CT90][DEM88]. An example of the RMS EMG signal shown in Figure 2-9. This raw EMG signal was sampled at 20 samples per second in the eight emotion experiment and the wearables experiment. In the automotive experiment the RMS EMG signal was saved at 15.5 samples per second using the 0.5 second averaging option provided by the Flexcomp software[Tho94].

2.1.5 Additional Modalities

Many other modalities may be considered to create the more complete picture of physiological state, but these are beyond the scope of this work. Blood pressure and electro-encephalogram (EEG) are mentioned here briefly because they also show potential for capturing important emotion related changes. However, other drawbacks made these modalities impractical for the research of this thesis.

Blood Pressure

Increases in blood pressure have been found to correlate with increases in stress [Sel80] however blood pressure was not used in this research because it was considered too cumbersome at the time of the experimental design. Blood pressure measurements
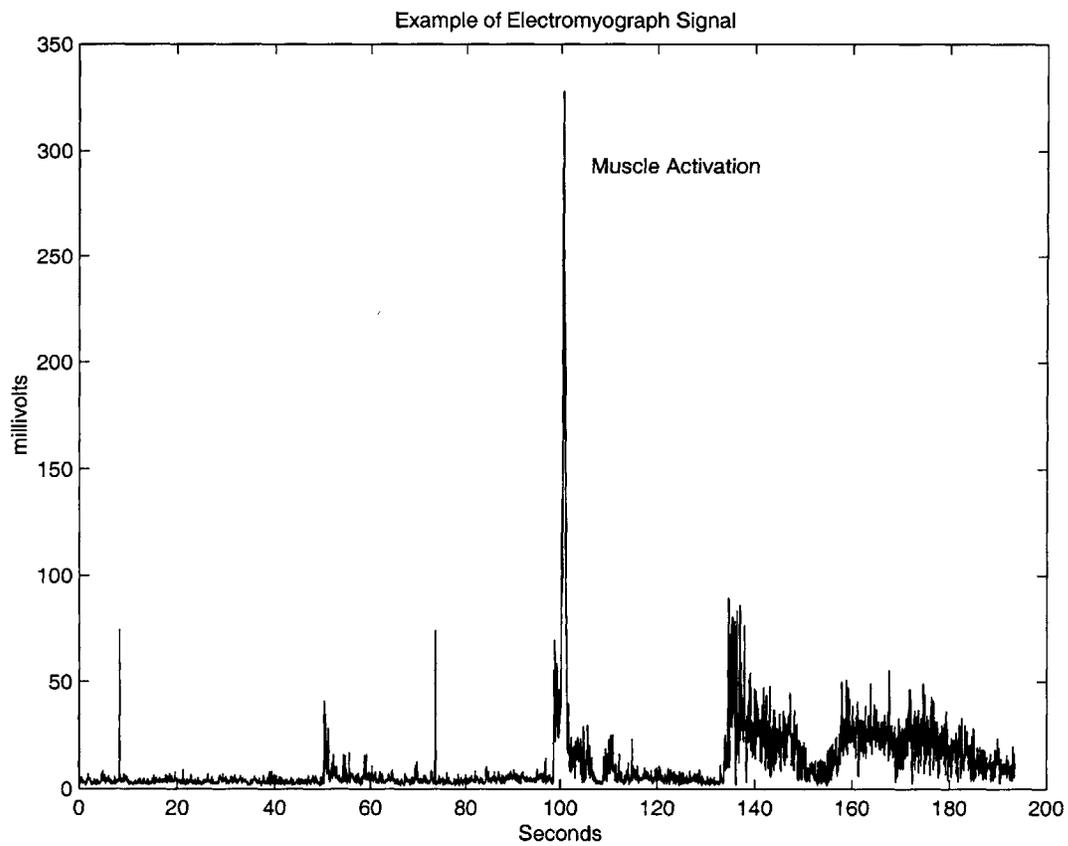
Figure 2-9: The electromyograph records muscle activity. Large activity readings are usually from motor activity

currently require the constriction of a blood vessel to determine the pressure required to restore blood flow in that vessel. To take this reading, a cuff inflates periodically on the subject's arm, a distraction which may have introduced artifacts into sensitive experiments measuring emotion or caused anxiety during the more dangerous driving experiment.

Electroencephalogram

The electroencephalogram (EEG) measures electrical activity of the brain by placing electrodes on the surface of the head. A full electroencephalogram incorporates over 128 electrodes, however simpler metrics using two or four channels are used in biofeedback practice[Tho94]. EEG's have been shown to distinguish between positive and negative emotional valence[Dav94] and different arousal levels[Lev90] under certain conditions. EEG can also be used to detect the orienting response by detecting "alpha blocking." In this phenomenon, alpha waves (8-13 Hz) become extinguished and beta waves (14-26Hz) become dominant when the person experiences a startling event[Lev90].

EEG was excluded from this thesis research because EEG readings are difficult to interpret in the ambulatory environment. The readings are often confounded by muscle activity such as forehead movement and the opening and closing of the eyes. In waking activity, EEG is considered only a crude measure of arousal[Lev90].

2.2 Emotion Classification

Emotion classification is necessary to provide labels to the physiological data in classification algorithms. Although there is no one universally accepted method of dividing the space of emotion, this section provides a description of the models widely used in theory and in practice. Furthermore, definitions of emotional stress are presented and a description of how stress fits into the space of emotion is discussed.

Theorist	Emotion Set
James	rage, fear, grief, love
Ekman	anger, fear, sadness, enjoyment, disgust, (surprise)
Clynes	anger, hate, grief, joy, love, romantic love, reverence, no emotion
Panskepp	rage, fear, panic, expectancy
Plutchik	anger, fear, anticipation, sadness, joy, acceptance, disgust, surprise
Izard	anger, fear, distress, joy, surprise, interest, disgust, contempt, guilt, shame
Frijda	anger, fear, distress, joy, surprise aversion, contempt, pride, shame, desire

Table 2.1: A brief summary of sets of basic emotions proposed by different theorists[OCC88].

2.2.1 Models of Emotion

Several theorists have proposed sets of basic emotions. These sets either span the space of emotion or provide a palette from which all other emotions can be derived. As an example, the basic emotion set proposed by Paul Ekman, includes anger, fear, disgust, sadness, and enjoyment and sometimes surprise[Ekm93]. Other theorists have created similar groupings of fundamental emotions and a brief summary of these is given in Table 2.1. This table is a partial representation of the sets collected by Ortony, Clore and Collins[OCC88]. Differences in these emotion sets may be biased to the particular field of the theorist Ekman's set might be biased to the the set of emotions most "basic" to facial expression while Clynes' set might be more basic to emotions found in piano music.

One of the difficulties of organizations using emotion names is that the relationships between the emotion categories is not clear. There is no structure for measuring the similarity of the emotions or the ways in which they differ. This makes the boundaries of what is included in each category difficult to define. For example, the specifications for separating distress from fear or contempt and disgust are subject to debate. The lack of a descriptive framework also makes it difficult to describe differences between examples of the same emotion. For example, Ekman states that
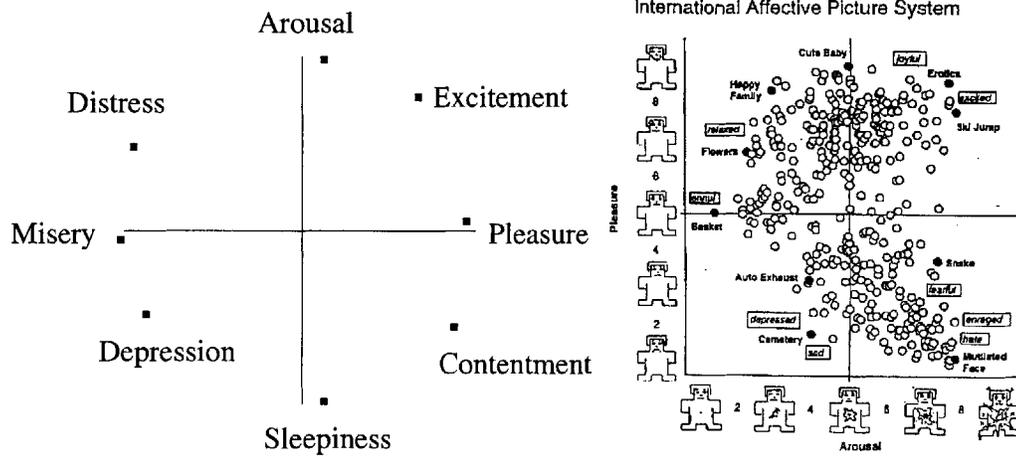
Figure 2-10: The use of emotion qualities to describe a space facilitates creating quantifiable relationships between emotions. Shown left, the circumplex model of emotion[Rus80], shown right, the arousal-valence space with self-assessment manikin (SAM)[LGea93]. In this ratings diagram, the valence axis is labeled pleasure, a more commonly recognized term.

"anger" refers to not one, but a family of over 60 anger expressions[Ekm92b] [EF78]. However the particular qualities which distinguish them do not tell how similar that instance of anger is to any other emotion such as fear or sadness.

In recognition experiments, frameworks which allow differences between emotions to be measured along some metric of similarity provide greater flexibility in designing experiments and offer greater insight into misclassification errors. Frameworks that support similarity metrics include the arousal-valence space used by Lang[LBC] and the circumplex model proposed by Russell[Rus80]. The circumplex model of emotion divides the emotion space into a circle with eight radial axes, spaced at slightly irregular intervals. These include, as opposing pairs, arousal and sleepiness, excitement and depression, pleasure and misery and contentment and distress[Rus80]. The arousal-valence space is similar to the circumplex model, except that only two axes are labeled, an arousal axis, ranging from calm to excited and a valence axis ranging from negative to positive. Both of these models are shown in Figure 2-10. The arousal-sleepiness and misery-pleasure axis in the circumplex model are similar to the arousal and valence axes in Peter Lang's model. It can be seen that if these two axes

38

are aligned across models the resulting spaces are very similar.

The practical benefit of the arousal-valence space is that it allows people to rate how they are feeling in a simple way that is easily quantifiable. To facilitate this rating process, a set of icons called the self-assessment manikins or SAMs were created. The illustrations in the icons help to clarify what is meant by each quantification of the emotion state. SAM is a gender-neutral outline of a person with a simple facial expression. Manikins on the arousal axis have a neutral facial expression and describe the level of arousal by using four icons with an expanding a starburst in the center of the outline. The starburst expands from a dot to an explosion to help the subject quantify arousal. The valence or pleasure axis has a manikin with no starburst and a facial expression varies from a frown to a smile. This rating system has been tested in practise and ratings have been found to be consistent in a large number of trials using emotionally eliciting picture stimuli[Lan95]. The arousal and valence axes may not be entirely independent, although depicted as separated at a 90 degree angle. One indication that the two axes are related is that the low arousal, low valence quadrant shown in Figure 2-10 is sparsely populated. This indicates that there exist few visual images that people find both very unpleasant and not at all arousing (disturbing).

2.2.2 Definitions of Stress

Historically, stress has been defined as a reaction from a calm state to an excited state for the purpose of preserving the integrity of the organism. The idea has been recorded as early as 450BC when Empedocles described stress as a threat to the harmonious balance of an organism's essential elements. Claude Bernard (1850) refined this idea and used the term "milieu interior" to describe the interior environment which held the essential elements. A stress reaction for Bernard is a reaction to protect this *milieu interior*. Walter Cannon (1927) coined the term "homeostasis"[EG88] to refer to the calm or "steady state" of the organism and stress as a transient "fight or flight."[EG88] reaction to preserve homeostasis.

These definitions support the idea of stress as moving from a calm to an aroused state. A negative bias seems to be inherent in the definitions. The reactions are

described by Bernard as protective and by Cannon as "fight or flight" not "fight or joy". Some researchers make a distinction between "eustress" and "distress," where eustress is a good stress, such as joy, or a stress leading to an eventual state which is more beneficial to the organism. The definition of stress in Chapter 5 will only cover "distress" or stress with a negative bias.

Bernard's hypothesis is that as organisms evolve and become more independent of their outer environment, they develop more complex mechanisms to serve the goal of preserving the interior from the exterior[EG88]. For an organism as highly developed and independent of the natural environment as socialized man, most stressors are intellectual, emotional and perceptual[Sel80]. Physical stressors occur far less frequently[Sel80].

The ambiguity of the emotion names is lessened by grouping emotions together in a qualitative space. This also allows discrimination with the metrics of similarity based on the qualitative axes. The arousal-valence space was used to help describe the emotion names used in the experiment presented in Chapter 3, and to help define the stress variable evaluated in the experiment presented in Chapter 5. The stress variable is described as primarily a measure of arousal with a slightly negative valence skew. This agrees with Selye's definition of emotional stress as overall autonomic nervous system arousal[Sel56] and the negative bias is supported by many other historical definitions. The goal of Chapter 5 will be to present a recognition algorithm for these short term emotional stress reactions.

2.3 Finding Physiological Patterns of Emotion

William James (1890)[Jam92] speculated that patterns of physiological response could be used to recognize emotion. This viewpoint has been debated by other theorists who argue that only certain aspects of emotion can be determined from physiology. Researchers have identified physiological quantities that discriminate between emotional states both in the laboratory and in the field, the results of which are reported. There are differences in experimental design and research goals that make some re-

sults hard to compare, but physiological trends are evident. A proposed framework of models and methods of analysis for future research is described.

2.3.1 The Domain of Physiology

Historically, there has been a debate over which aspects of emotion can be recognized, or best recognized, within the domain of physiology. James believed that the emotional physiological reaction *was* the emotion, so that everything which can be known about emotion is inherent in physiological signals. Cannon seriously challenged this viewpoint, arguing emotion was primarily a cognitive event and that autonomic patterns were too slow and non-specific to be unique to each emotion[Can27]. Schacter is said to have resolved the "Cannon-James controversy" by showing that emotional response was both physiological and cognitive[Sch64]. However, arguments were also made against Schacter's experiments and further debate continues on the degree to which emotion can be determined from physiological measures.

2.3.2 Laboratory Studies

Patterns in emotional response have been sought both with respect to emotion name spaces and with respect to the axes of arousal and valence. This section lists a brief summary of the results of previous experiments in the field.

Patterns Using Emotion Name Classification

Many laboratory studies have been performed to find measurable patterns for the emotion descriptions James first proposed. James' descriptions were only qualitative descriptions, for example he described anger as "increased blood flow to hands, increased heart rate, snarling and increases involuntary nervous system arousal" and fear as "a high arousal state, in which a person has a decrease in voluntary muscle activity, a greater number of involuntary muscular contractions and a decrease of circulation in the peripheral blood vessels." Using modern sensors for psychophysiological monitoring, Ekman, Levenson and Friesen have discovered evidence for dis-

Emotion	Physiological response		
	heart rate acceleration	hand temperature	skin conductance
Anger	large positive	up	none
Fright	none	down	big up
Disgust	deceleration	none	big up
Sadness	small positive	none	none
Happiness	none	no effect	small up

Table 2.2: Ekman, Levenson and Friesen noted trends in different physiological variables which could be used to determine differences between some emotion states[ELF83].

tinctive patterns of autonomic nervous (ANS) activity for anger, disgust and possibly sadness [ELF83] [Lev92], as shown in Table 2.2. The authors measured heart rate acceleration, hand temperature and skin conductivity. Using combinations of each of these metrics, it seems that each emotion state could be uniquely recognized, however no such recognition study was performed.

Patterns Using Arousal-Valence Classification

Lang, Winton, Putman and Krauss have reported features that correlate with the arousal and valence axes. Lang found that the magnitude of the skin conductance response varied linearly with arousal ratings[Lan95]. Valence has been reported to correlate with activity in the corrugator muscle, inhibition of the startle reflex[Lan95] and changes in heart rate acceleration. Winton Putnam and Kraus report heart rate acceleration changes in the 10 second window following stimulus which differentiate positive and negative images. Their results are shown in Figure 2-11 [WPK84]. Lang confirmed this result by reporting a correlation between valence and the peak heart rate in the six seconds following picture onset[Lan95].

2.3.3 Ambulatory Monitoring

Recent advances in microprocessors have allowed digital recording of physiological signals in the ambulatory environment. This facilitates processing the recording of
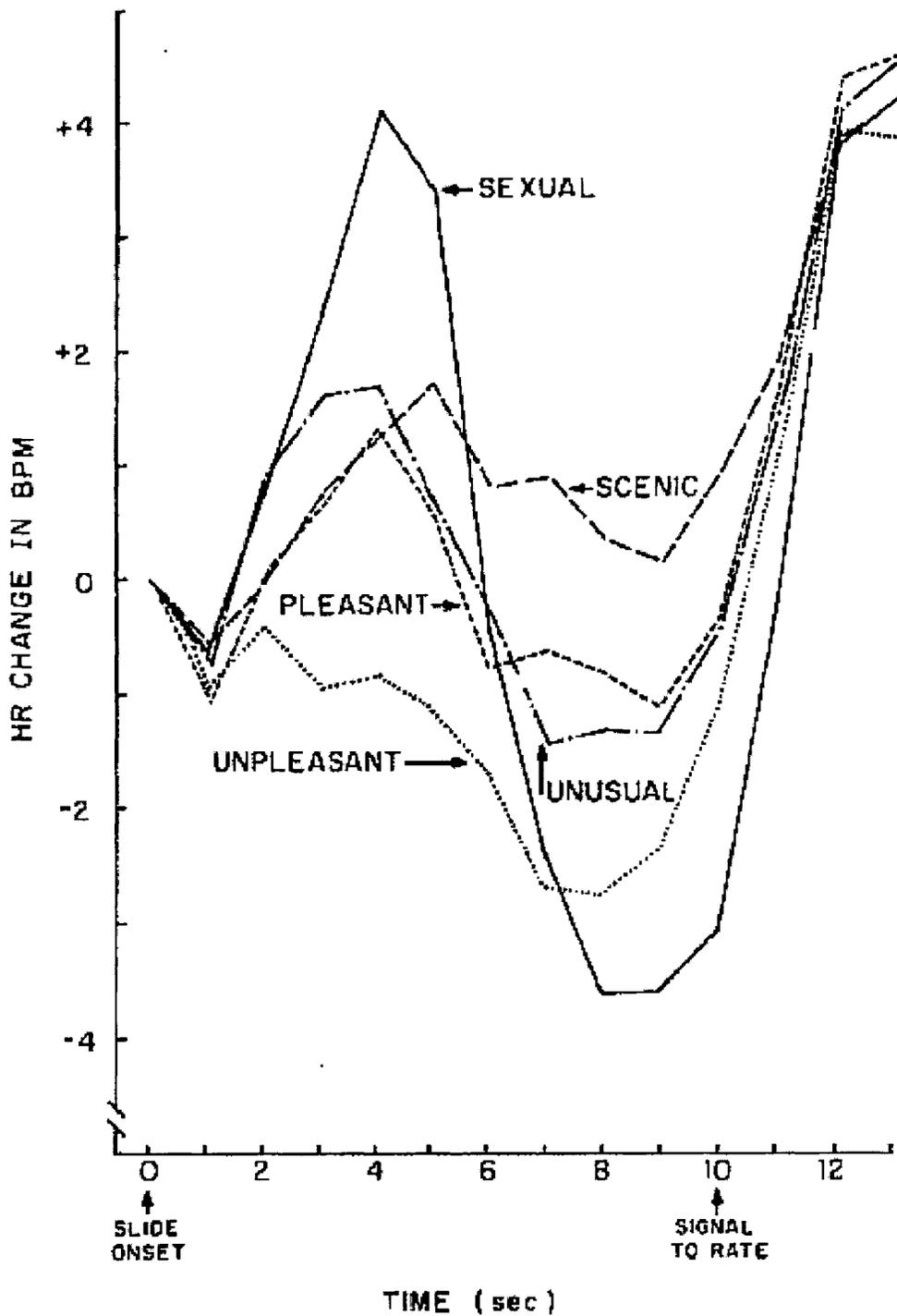
Figure 2-11: Heart rate deceleration, associated with valence is difficult to discern unless the time of stimulus onset is known. Averaging heart rate over the ten second interval here would have obscured the effect. Image from Winton, Putnam and Krauss (1984) [WPK84].

43

physiological signals outside the laboratory and allows computer algorithms to interpret and respond to the signals in the natural environment Such field studies of emotion have a higher degree of "external validity" than experiments measuring emotion in the laboratory. This external validity comes at the cost of less certainty in the validation of the emotion, relying on user compliance to label data and subject self-ratings. Although ambulatory medical monitoring devices exist for cardiac and blood pressure evaluation over 24 hour periods, these tools have not been applied to the problem of emotion monitoring, with the recent exception of monitoring panic attacks[HB96].

One ambulatory experiment was conducted by Buse and Pawlik using a small digital computer and a device for recording heart rate (BVP)[BP96]. The study consisted of two assessment periods of one week in duration, spaced fourteen months apart. The test used active self-report in which each subject was prompted on 71 occasions during the week, with interrogation intervals varying from 30 to 90 minutes (mean 60 minutes). When prompted, the subjects were asked to fill out an electronic survey which included a list of activities, a rating of mental activation and an assessment of emotion along the axis of "euphoria" to "dysphoria." At this time heart rate was also recorded.

The goal of this experiment was to determine measures of consistency for heart rate and mood variables throughout the day. They wished to determine if heart rate and mood at the present time could be predicted using previous measurements. From their results, the authors could not find any significant prediction of heart rate from previously reported heart rate, but found that euphoria and dysphoria had significant correlations with previously reported mood (.48 for euphoria and .33 for dysphoria (p>0.05)). No correlation was found between mood and heart rate or heart rate and previous heart rate. The reason for this was hypothesized to be that the heart rate variable was mainly related to physical motion[BP96].

This experiment encountered several problems with subject self report in the field. The study consisted of 135 subjects originally, but only 104 were willing to complete the study. Also, the compliance rate for completing the surveys when requested by

the prompt was only 61%. This non-compliance was partially due to the subjects being somewhere where filling out the survey would have been distracting, such as a movie theater, at the time for the prompt. To compensate for this the authors asked subjects to fill out a "retrospective protocol of the day" summarizing, by way of an hourly schedule their essential daily activities[BP96]. Their study shows that conducting purely ambulatory monitoring is difficult, both because of the difficulty in assessing the subjects emotional state and because of confounding physical artifacts in the physiological readings.

2.3.4 Measuring Stress in Aircraft Pilots

The psychological workload on pilots in the stress stimulating environment of flight has been a concern for both commercial and military aircraft designers. To measure the amount of stress pilots experience, especially while flying military aircraft, several studies have been performed. These studies provide valuable information regarding the differences between simulated and real experience and between novice and veteran subject groups. As an example, a study that uses heart rate as an indicator of stress in both real and simulated flight of a BA Hawk MK 51[YLL+97] is presented. This study reports the effect of the various observable stressors experienced over the course of a pre-planned flight mission. This study also compares the difference between measures collected from a simulated versus real flight experience. The same flight mission was performed in both a simulator and in a real jet. Five experienced (aged 26-33 yr) and five less experienced (aged 23-25 yr) male military pilots on active flying status participated in the study. In the study, the heart rates for the following observable events were recorded: rest, take-off, initial approach, intermediate approach, final approach, landing tour and landing.

The approximate time of the test was 25 minutes. Heart measurements were continuously measured using a small portable recorder. The R-R intervals were stored and analyzed with an accuracy of 1ms. The different phases of each flight were marked in the data using pre-determined codes. The marking was done either by an investigator (for the simulations) or a pilot in the back seat (for the real flights).

Simulated Flight	Experienced Pilots			Novice Pilots		
Flight Phase	HR	DHR	DSD	HR	DHR	DSD
Rest after seating	74.45	21.65	10.03	85.64	27.04	19.69
Take-off	85.78	32.98	15.41	96.34	37.74	14.17
Initial Approach	82.44	29.64	13.93	94.78	36.18	15.22
Intermediate App.	84.66	31.86	12.65	95.08	36.48	16.12
Final Approach	85.94	33.14	15.15	93.24	34.64	16.13
Landing Tour	80.76	27.96	13.15	89.52	30.92	15.40
Landing	85.92	33.12	13.93	93.56	34.97	15.74
Real Flight	Experienced Pilots			Novice Pilots		
Flight Phase	HR	DHR	DSD	HR	DHR	DSD
Rest after seating	71.62	18.82	6.37	76.10	17.50	11.51
Take-off	89.82	37.02	9.95	89.34	30.74	10.78
Initial Approach	87.50	34.70	8.38	89.84	31.24	12.36
Intermediate App.	88.48	35.68	11.46	91.06	32.46	9.12
Final Approach	95.06	42.26	14.60	92.18	33.58	8.66
Landing Tour	87.52	34.72	11.72	87.60	29.00	7.55
Landing	91.94	39.14	11.84	88.10	29.50	8.68

Table 2.3: The mean heart rates (HR), corresponding mean delta heart rates (DHR) and their standard deviations (DSD) in each real and simulated flight phase

The authors report possible artifacts due to an algorithm that excluded those R-R intervals that differed from previous intervals by more than 60% (to account for missed beats and false positives). For every analysis the baseline heart rate of every pilot was subtracted from his heart rate in each flight phase and during the rest after seating. The heart rates obtained after this subtraction are the delta heart rates (DHR). Both actual (HR) and delta (DHR) heart rates are reported in Table 2.3.

A two way analysis of variance for repeated measures was used for statistical testing of the differences. The results of this study show no significant heart rate differences between real and simulated flight. However, different phases of flight showed significant changes in the heart rate variables in both cases. It is interesting to note however, that the novice and experienced subjects reacted differently. Heart rate was consistently higher for experienced pilots in the real flights and higher for the novice pilots in the simulated flights. Differences in novice vs. veteran subjects were also found in parachute jumpers including an increase in heart rate after the

46

jump is completed[Fri86].

This study also suggests that simulations can be used to gather data similar to that of real situations, although other studies have shown that the results from simulations do not produce the same results as real flight in terms of hormone levels[LLH+95]. Although there is no unambiguous trend, real experience is different from simulated is some situations and a difference between novice and experienced subjects is evident.

2.3.5 Measuring Stress in Automobile Drivers

Previous driving studies have also been done to measure physiological variables in response to driving events. Martin Helander conducted a study in which braking and traffic events were recorded and correlated with multiple physiological signals[Hel78]. The events recorded included: encountering a cyclist or pedestrian, another car merging in front of the driver's car and the driver's car passing another car. In this study, an EKG was used to measure heart rate (HR), skin conductivity was measured across the palm (EDR) and the muscular activity was measured using two EMG, one on the calf (m. tilialis anterior - EMGTA) and one on the anterior hip muscle (EMGAS). The skin conductivity change was predicted to be the best measure of driver stress to fast events because of its fast onset when compared to heart rate changes[Hel78]. The EMG, steering and braking measures were taken to assess the physical workload experienced by the driver since heart rate and skin conductivity changes can be effected by both physical and mental task load[Hel78].

The test vehicle was a Volvo 145 which was driven on a 23.7km stretch of rural road composed of four segments. All tests were performed during daylight non-peak hours and dry road conditions. Sixty test drivers were paid to participate in the study. A Spearman rank order correlation was performed between traffic events and the physiological and sensor measures. The events were recorded as occurring or not occurring on each 10 meter strip of the road, this is why the "no event activity" has such a high rate of occurrence. The results are reported in Table 2.4.

GSR, EMG and EKG were recorded during the driving task, however only the EMG and GSR readings were reported in the analysis. The GSR reading was ex-

		Traffic Events in rank by:				
Traffic Event	n	Brake	EDR	HR	EMGTA	EMGAS
1. Cyclist or pedestrian meets other car	28	1	1	7 (81.2)	2	1
2. Other car merges	47	2	2	4 (81.5)	1	2
3. Multiple events	163	3	3	13 (78.3)	3	3
4. Leading car diverges	207	4	5	8 (80.5)	4	11
5. Cyclist or pedestrian	839	5	7	1 (82.5)	8	5
6. Own car passes and car following	126	6	6	12 (78.4)	13	15
7. Cyclist or pedestrian and car following	65	7	10	3 (81.7)	14	9
8. Car following and meeting other car	353	8	12	14 (76.2)	10	13
9. Meeting other car	1,535	9	9	10 (78.9)	5	8
10. Car following	13,049	10	11	9 (79.3)	11	12
11. Parked car	742	11	15	2 (82.4)	6	6
12. No event	112,630	12	13	11 (78.6)	9	7
13. Other car passes	157	13	8	15 (76.0)	12	4
14. Parked car and car following	64	14	14	5 (81.4)	7	14
15. Own car passes	3,590	15	4	6 (81.3)	15	10

Table 2.4: A previous experiment by Helander analyzed skin conductance, heart rate and EMG signals of drivers. In this table, Brake represents brake pressure, EDR represents the electrodermal response, HR indicates the heart rate in beats per minute, EMGTA represents the activation of the m. tibialis anterior and EMGAS represents the activation of the anterior hip muscle. The magnitude of the responses were measured against brake pressure. The EDR was found to be most highly correlated signal to the braking activity.

48

pressed in decibels: GSR $= 10 log \frac{G}{G_m}$ dB, where $G =$ momentary conductivity in $\mu\Omega$ and $G_m =$ mean conductivity in $\mu\Omega$. The EMG's were scaled for each driver, with the highest response being given a value of 98 and the other responses being scaled accordingly. Measurement values are finally averaged across each 10-meter stretch of road. Braking was taken to be the effective measure of workload and the correlations of the physiological variables with respect to braking are reported in Table 2.4. This study showed the skin conductance was more highly correlated to braking than either heart rate or either of the EMG variables.

2.4   A New Model for Psychology and Physiology

Modern theorists have proposed that the key to finding unique ANS patterns in emotion lies in developing a more advanced model of the relationship between psychology to physiology. This model maps physiological variables to emotion space initially as a many to one relationship, but hypothesizes that by using mathematical combinations of features better discrimination of the states will become possible [CT90]. Cacioppo and Tassinary propose that the psychological domain represented by the greek letter psi, and the physiological domain, represented by the greek letter phi might be related as shown in Figure 2-12. This illustration shows that psychological states such as "relaxation, orienting, startle and defense" can have confused mappings to the physiological domain if features such as skin conductance and accelerating and decelerating heart rate are considered separately. They theorize that if features are combined, these domains can be more unambiguously mapped. They further propose that looking at combinations of features over different time windows will further lead to a unique mapping.

To show how combinations of features can create better discrimination, an example of data points in two feature dimensions is shown in Figure 2-13. They note that if the plotted points were projected onto either the x-axis or the y-axis, representing single features, it would be difficult to create a discrimination boundary that would separate the two classes of points. However, by creating a feature which is a linear combination

PSYCHOLOGICAL DOMAIN   Ψ   Φ   PHYSIOLOGICAL DOMAIN

Relaxation
Orienting
Startle
Defense

SCR

Decelerating HR

Accelerating HR

Relaxation
Orienting
Startle
Defense

No SCR and Decelerating HR

SCR and Decelerating HR

SCR and Accelerating HR

Relaxation
Orienting
Startle
Defense

No SCR and Decelerating HR

SCR and Decelerating HR

Abrupt accelerating HR

Lingering accelerating HR

Figure 2-12: Combining features can create a one to one relationship between events in the psychological domain and events in the physiological domain. From Psychophysiology [CT90]

of these two and projecting the points onto that line, a far better discrimination can be achieved. They suggest the mean and total energy of an EMG reading might generate such a dataset. This idea of combining multiple features and projecting them into an optimally discriminating space has been well studied in the pattern recognition community for years [DH73]. It is the goal of this thesis to apply these techniques to the problem of distinguishing emotion states in physiology as suggested by Cacioppo and Tassinary.

2.5   Summary

Sensing physiological patterns is not a new thing; ambulatory medical devices have been under development for years, recording heart rate and blood pressure for cardiac analysis and helping monitor panic attacks[HB96]. Studies are just beginning to detect affect in the ambulatory environment. These studies face the challenges of
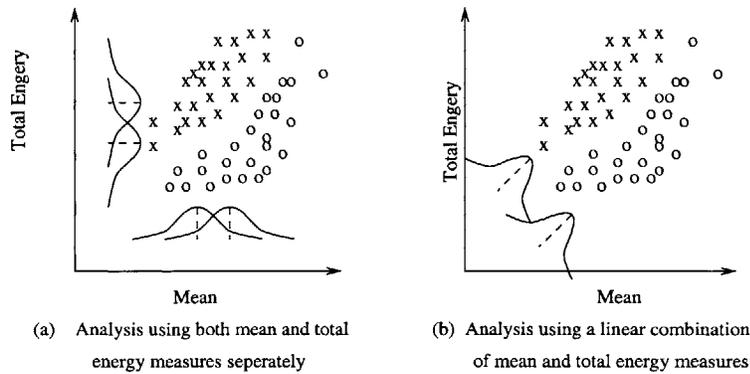
Figure 2-13: Combinations of features can lead to more optimal discrimination of states. If the data points in this two dimensional space were projected on to either the x or the y axis, the distributions that would be formed would not have good separation. By projecting the points onto a line that is formed by the combination of two features, as shown in Figure (b), a much better discrimination can be obtained. This figure taken from Psychophysiology [CT90].

modeling and documenting occurrences of natural affective states and of mediating the impact of physical changes on physiological variables. These challenges may be more easily met by conducting experiments in a vehicle. A vehicle platform allows natural responses to be captured in a platform where motion artifacts are limited and where cameras can be placed to monitor the subject.

Physiological variables have been found to be significantly different during tasks in vehicle studies. Studies in real and simulated flight show that heart rate can be an excellent predictor of stress in extreme circumstances such as take off and landing. In driving studies, galvanic skin response was found to be highly correlated with the stress of various tasks. Cacioppo and Tassinary suggest that by using novel combinations of physiological variables, better discrimination between emotion states might be achieved.

This section introduced the basic tools that will be used for the physiological monitoring performed in Chapters 3,4 and 5. Various frameworks for modeling and quantifying emotion were presented to show how the emotion qualities of arousal and valence can be used to measure differences in emotion states. Results from laboratory and field studies were presented to show how physiological variables have been shown to correlate with affective states. The work of this thesis will be to advance the field

by testing the application of more advanced signal' processing and pattern recognition techniques to the problem of emotion recognition. New features and algorithms for discrimination are presented. Multiple modes of physiological signals are combined to create novel classifiers. In addition, new systems for capturing affect in the natural environment are developed for both ambulatory and driving situations.

Chapter 3

Eight Emotion Experiment

The eight emotion experiment was designed to test for the presence of unique physiological patterns for the emotion set: no emotion, anger, hate, grief, love, romantic love, joy and reverence. This experiment studies the most diverse set of emotions of this thesis; however, the setting is the most constrained. This particular set of emotions was chosen because it has been found to produce a unique set of finger pressure patterns when intentionally expressed[Cly77]. Using the same experimental protocol that generated these results in the finger pressure feature, four additional signals were collected: blood volume pulse, elecromyogram, respiration and skin conductance. The experiment was performed over thirty-two days with the author as the subject. Although the subject was informed, she had no access to the signal readings during the experiment and made every effort to authentically feel and express the emotion states to create a good data set from which to develop features for future use. Twenty perfect datasets in which there were no hardware or software failures were collected over the thirty-two days. The goal of this experiment was to find a discriminating set of features in the physiological domain and to see if an algorithm could be built, using these features, for classification of the eight states.

3.1 Experimental Protocol

Each experimental session required the subject to sit in a chair with a straight back while wearing sensors and to sustain each emotion for a period of three to five minutes at the prompting of the Sentics software. The subject sat with feet placed comfortably apart, with the left hand on the left leg and the right hand on a finger-pressure sensing device called the the Sentograph. The BVP and GSR sensors were placed on the resting left hand, the respiration sensor was placed around the diaphragm, and the EMG was placed on the masseter muscle to measure jaw clenching. The subject wore headphones over which a computer voice announced the start of each emotion episode. The episodes occurred in the same order for each experimental session: no emotion, anger, hate, grief, platonic love, romantic love, joy and reverence. The name of each emotion was first announced, followed by a series of soft metronome clicks. At each click the subject pressed the finger pressure sensor, to mimic the original experiment and to evoke the emotion through physical expression. The software randomly varied the duration of each emotion episode to lessen anticipation effects. The duration of each episode ranged from three to five minutes.

Data was collected over a period of 32 days in an office setting. At each experimental session, approximately thirty minutes of data was recorded. Twenty complete datasets were collected due to various equipment failures, including sensors becoming detached from the subject, the connector from the sensor to the analog-to-digital converter becoming unplugged and loss of electronic files due to software errors. The four analog sensors, GSR, BVP, respiration and EMG were sampled at 20 samples per second using the ProComp unit from Thought Technologies, Ltd[Tho94].

3.2 Emotion Generation

The Sentics software provided a cue to begin each emotion episode and the Sentograph provided a means of physical expression; but the subject was responsible for generation of the emotion. Relying on a guided imagery technique for acting that the

subject learned, the subject picked specific images to use as cues during each emotion episode. The images were chosen which had both a compelling impact on the subject and which matched with the guidelines Clynes laid out for interpreting the emotion states. The goal of these images was to encourage a consistent generation of the emotion across multiple data collecting sessions. During the experiment the subject attempted to feel and express eight affective states for the period specified by the Sentics software (a length of time varying between three to five minutes with random variations). The subjects goal was to authentically feel and express this emotion to the computer as they might strongly express an emotion to another person. Because the goal of this experiment was to truly determine if unique physiological signatures of emotion could be detected, the subject attempted to authentically feel the emotions. She reported that this generation was different from emotion generation for stage action because the expression was not overly exaggerated.

Defining a set of basic emotion states is a difficult task and one that has been the subject of much debate in the psychological community [Ekm92a]. Analysis of any system becomes more problematic when the inputs are unknown or ambiguously classified. Identifying or generating an emotion of a specific class is easier when the class is well defined. Each of the following subsections presents: a summary of the guidelines provided by Clynes, the images used by the subject to elicit the emotion state and an interpretation of the resulting emotion along the axes of arousal and valence. As described earlier in the background chapter, arousal is a measure of the strength of activity associated with the emotion (ranging from calm to very excited) and valence is an axis reflecting the positive or negative aspect of the state of the state (ranging from negative=sad to positive=happy). These generation assessment ratings were used in the analysis to create subsets of emotion that shared similar qualities.

No Emotion

Clynes instructs the subject to just "think of the pressing as just a mechanical action, as if you are pressing on a typewriter key." [Cly77] in his guidelines for the Sentics

experiment. For this state the subject used the image of a blank piece of paper and the image of typing blank characters on the paper using a typewriter with no ribbon. The subject consistently rated the no emotion state as being low arousal and neutral valence. Errors in generation occurred when the subject occasionally became bored and her mind wandered to other topics.

Anger

Clynes did not give specific instructions for expressing anger. However, the subject did not find this emotion difficult to imagine. At the time the subject had a specific person whose image consistently aroused anger. She used the image of this person to generate the emotion. The subject rated this emotion as having the highest arousal level of the eight states, although on a few occasions the subject reported that fatigue affected the generation of this emotion and a high arousal level was not achieved. She also reported the valence as consistently negative. This emotion could also be described as rage.

Hate

Clynes also gave no specific instructions for imagining hate. The subject was asked to name different things she felt hatred towards and to describe the emotion. She chose to use images of injustice (television images of people being treated badly by police forces or seeing people forced to live in extremely bad conditions), school yard bullies, hate crime graffiti, and scenes of war. The subject described these feelings of hate as things she would feel angry about if she felt it was possible to take action against them. This may be similar to the experience of the "cold anger"[Ax53] described in emotion literature. The subject rated this emotion as low arousal and negative valence. The subject reported that the vagueness of this emotion made it difficult to generate.

Grief

Clynes describes the grief state as a transition between negative and positive emotions. From this the subject understood it to be a cathartic emotion, an active form of grieving that would purge negative emotions. The subject used two primary images to generate this emotion. She either imagined the loss of someone she loved, in particular her mother or she remembered a specific picture in which a mother is grieving over her deformed child in a large bathtub. Both images were imagined situations, the first possibly a more direct form of grief and the second perhaps a more empathic form of grief. The subject reported that the spontaneous generation of this emotion was often the most difficult. Either the emotion was not well generated or the emotion was so well generated that it was difficult to contain its expression to the window allowed by the Sentics software. The subject described this emotion as high arousal, negative valence when successfully generated, and as a lower arousal, negative valence state when the generated emotion felt more like depression. The subject reported that this emotion could also be interpreted as a kind of pity, either for herself or for the deformed child in the image.

Platonic Love

Clynes describes platonic love as the experience of motherly or brotherly love as opposed to sexual love. The subject used a childhood image of being with her mother in a hammock at a summer cottage with her family to generate this state. The subject rated this emotion as low arousal and positive valence. Many psychologists disagree with the inclusion of love as an emotion due to the long duration of the state[Ekm92a]; however, the subject reported no difficulty generating this emotion or containing its expression to the short window of the emotion episode.

Romantic Love

Clynes describes romantic love as a love for a member of the opposite sex[Cly77]. To generate this emotion the subject used erotic imagery. The subject reported that this

emotion was often difficult to generate. When properly generated, this emotion had very high arousal and positive valence, otherwise a neutral arousal, neutral valence state was reported. The results of this state were the most highly varied across days.

Joy

Clynes does not give any specific instructions for experiencing joy. The subject immediately associated this emotion with Beethoven's "Ode to Joy" and used the memory of this song rather than a visual image to elicit the emotion. The subject reported feeling joy as a uplifting emotion, an excited form of happiness. Often this feeling was coupled with a feeling of triumph. Joy was rated as a high arousal, positive valence emotion.

Reverence

Clynes describes reverence as "not reverence for a person, but for nature or for God, or the starry skies, i.e. for something larger than yourself"[Cly80]. The subject used images of being in church and reciting prayers to generate this emotion. The subject reported this emotion as having very low arousal and neutral valence. She reported is was easy to generate and was consistently experienced as peaceful.

Emotion Generation Summary

Although the intensity of the emotions sometimes varied, the overall character of each state was consistent. The subject reported that using the finger pressure device as channel for expressing each of these emotions was always helpful. It provided a focus for renewed expression of the emotion every few seconds. A summary of the images used to elicit these emotions and the general ratings they were given in the dimensions of arousal and valence are summarized in Table 3.1.

Emotion	Imagery	Description	Arousal	Valance
No Emotion	blank paper typewriter	boredom vacancy	low	neutral
Anger	people who aroused anger	desire to fight	very high	very negative
Hate	injustice cruelty	passive anger	low	negative
Grief	deformed child loss of mother	loss sadness	high	negative
Platonic Love	family summer	happiness peace	low	positive
Romantic Love	romantic encounters	excitement lust	very high	positive
Joy	Ode to Joy	uplifting happiness	medium high	positive
Reverence	church prayer	calm peace	very low	neutral

Table 3.1: A description of the emotional states intentionally induced and expressed in this experiment

3.3 Feature Extraction

Eleven features were initially proposed for extraction from the four signals. This section described the features and the signals from which they are derived. The signal processing for each sensor, including smoothing, normalization and feature extraction is described in four subsections. The symbols $\mathcal{G}, \mathcal{B}, \mathcal{R}$ and $\mathcal{E}$ represent the GSR, the BVP heart rate, the respiration and the EMG respectively. The heart rate signal, $\mathcal{B}$, is calculated automatically by the ProComp system using the inverse of the inter-beat intervals detected from the BVP sensor[Tho94]. No features were extracted from the raw BVP signal. All signals are analyzed digitally after being sampled at a rate of twenty samples per second by the ProComp system.

Let X designate the samples taken from any one of the eight emotion episodes on any one of the twenty days. The features representing a particular emotion on a particular day are calculated from these segments. Let the bar symbol represent the signal taken over an entire day, across all eight emotion episodes, e.g. $\bar{X}$. These long

data records are used in normalization calculations. Let the tilde symbol represent a normalized signal, e.g. $\tilde{X}$ is a normalized version of X, where different normalization procedures are described for each feature in the following sections. Let the lower case represent a smoothed signal, e.g. x represents the smoothed signal X. All smoothed signals were created by convolution with a 500 pont (25 sec) Hanning window. This window, represented as w, was generated by the Matlab function "hanning." The smoothed signal x is described by $x = X * w$. All signals were sampled at a rate of 20 samples per second, creating a digital version of the signal. Let X_n represent the value of the nth sample of the raw digital signal, where $n = 1, \ldots, N$, with N in the range of 2000 to 5000 for the varied lengths of the emotion episodes.

Features extracted from multiple sensors include the mean, variance and the mean of the first forward difference. These features, represented by the symbols μ_X, $\sigma_X{}^2$ and δ_X or $\tilde{\mu}_X$, $\tilde{\sigma_X}{}^2$ and $\tilde{\delta}_X$ for normalized signals are defined by the Equations 3.1- 3:

1. Mean for Raw and Normalized Signals:

$$\mu_X = \frac{1}{N}\sum_{n=1}^{N} X_n; \qquad \tilde{\mu}_X = \frac{1}{N}\sum_{n=1}^{N} \tilde{X}_n \tag{3.1}$$

2. Variance for Raw and Normalized Signals:

$$\sigma_X^2 = \left(\frac{1}{N-1}\sum_{n=1}^{N}(X_n - \mu_X)^2\right); \quad \tilde{\sigma}_X^2 = \left(\frac{1}{N-1}\sum_{n=1}^{N}\left(\tilde{X}_n - \tilde{\mu}_X\right)^2\right) \tag{3.2}$$

3. First Forward Difference Mean for Smoothed and Normalized Smoothed Signals:

$$\delta_x = \mu(x_{n+1} - x_n) = \frac{1}{N-1}(x_N - x_1); \quad \tilde{\delta}_x = \frac{1}{N-1}(\tilde{x}_N - \tilde{x}_1) \tag{3.3}$$

The δ_x calculation was used to capture a general estimate of the slope. It is performed only on smoothed signals which are less likely to contain outliers in the first and final

EMG on Jaw

Skin Conductivity
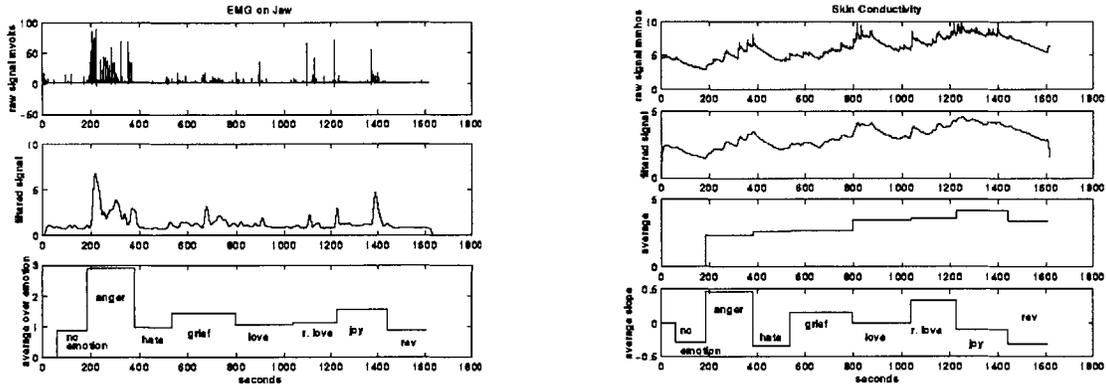
Figure 3-1: An example of a raw EMG signal from which the feature was extracted. An example of a raw GSR signal from which the features of mean and average differential were extracted Each mean feature is labeled with the name of the emotion for the episode during which it was recorded. The line of the mean feature extends over the time period of each emotion.

points of the emotion episode.

EMG

The EMG sensor was used to capture motor activity in the masseter muscle. The signal was neither filtered nor normalized in the feature analysis. One feature was extracted, the mean of the raw signal $\mu_\mathcal{E}$. This measure can reflect jaw clenching, frowning and smiling. Figure 3-1 shows the raw data $\mathcal{E}$, the same data smoothed for better visualization, e and the mean for each emotion segment $\mu_\mathcal{E}$. Each mean feature is labeled with the name of the emotion for the episode during which it was recorded. The line representing of the mean feature extends over the time period of each emotion.

GSR

Two features were extracted from the smoothed and normalized signal, the normalized mean, $\tilde{\mu}_G$ and the normalized first difference mean $\tilde{\delta}_G$. These features were chosen to represent the overall autonomic level during the emotion episode and the change in autonomic activation over the course of the episode. An example of how these features represent the signal is shown in Figure 3-1.

The smoothed signal, g was normalized using a metric proposed by Rose in 1996[LRLM66] and subsequently found to be valuable to other researchers[LV71], [SF90]. The normalization metric is defined by:

$$\tilde{G} = \frac{g - \min(\bar{g})}{\max(\bar{g}) - \min(\bar{g})}.$$ (3.4)

where $\max(g)$ and $\min(g)$ are the maximum and minimum points of the signal for the entire day's session.

BVP Heart Rate

The heart rate signal, B is calculated automatically by the ProComp system using the inverse of inter-beat intervals detected from the raw BVP signal. Two features were extracted from The heart rate signal after smoothing, the the mean,μ_b, and the mean of the first forward difference, δ_b. These features are similar to those extracted from the GSR, except no normalization procedure is applied. The signal B was smoothed using the same window w which was used for the GSR, $b = B * w$.

Respiration

The respiration sensor measures expansion and contraction of the chest cavity using a Hall effect sensor attached around the chest with a velcro band. Six features were extracted from the respiration signal, two in the time domain and four in the frequency domain. In the time domain, the signal was normalized by subtracting off the overall mean of the data for that day. This removes baseline variations due to the initial tightness of the sensor when it is stretched around the diaphragm, an artifact which

is not due to any physiological effect. This normalization is described by:

$$\tilde{\mathcal{R}} = \mathcal{R} - \bar{\mu}_{\mathcal{R}}. \tag{3.5}$$

The mean, $\bar{\mu}_{\mathcal{R}}$ and the variance, $\bar{\sigma}_{\mathcal{R}}^2$ were calculated from this normalized signal. These features reflect resting lung volume and the overall amount of variation in the respiration signal.

Four additional features were calculated from a power spectral density of the signal. For these features, the last 2048 points of data collected for each emotion episode were used. Each 2048 point segment was normalized by subtracting off the mean of the segment. Let R represent the uniform length segments. The power spectrum estimate used the normalized signal:

$$\tilde{R} = R - \mu_R \tag{3.6}$$

From this normalized signal the power spectral density was calculated using the Matlab command PSD. This command was invoked using the command:

$$P_{RR} = PSD(\tilde{R}, 2048, 10); \tag{3.7}$$

where $\tilde{R}$ is the normalized 2048 point segment, 2048 is the length of the FFT and 10 is the Nyquist frequency for the 20Hz sampling rate. The segment is windowed by a hanning window the length of the segment. The spectral estimate is most accurately described by the on-line reference provided by Matlab[Mat99].

> Pxx = PSD(X,NFFT,Fs,WINDOW) estimates the Power Spectral Density of a discrete-time signal vector X using Welch's averaged, modified periodogram method.
>
> X is divided into overlapping sections, each of which is detrended (according to the detrending flag, if specified), then windowed by the WINDOW parameter, then zero-padded to length NFFT. The magnitude squared of the length NFFT DFTs of the sections are averaged to form Pxx. Pxx is length NFFT/2+1 for NFFT even, (NFFT+1)/2 for NFFT odd, or NFFT if the signal X is complex. If you specify a scalar for WINDOW, a Han-

ning window of that length is used. Fs is the sampling frequency which doesn't affect the spectrum estimate but is used for scaling the X-axis of the plots.[Mat99]

The spectrum was then normalized by dividing by the sum of the samples:

$$\tilde{P}_{RR} = \frac{P_{RR}}{\sum_1^{2048} P_{RR}};$$ (3.8)

From this normalized spectral estimate four band characteristics were calculated by summing 10 samples and dividing by the number of samples. Using a 2048 point FFT in Matlab results in a spectrum with 1024 points mirrored across the real axis. The positive 1024 points correspond to a frequency band ranging from zero to the Nyquist rate of the sampling frequency. In this case, the 1024 points represent the frequency range 0-10 Hz. Each sample represents a frequency spaced at intervals of $\frac{10}{1024} = 0.00976$ Hz ≈ 0.01 Hz. Therefore each feature represents a band of 0.1Hz. Let p_N represent the band power feature as defined by Equation 3.9:

$$p_i = \sum_{10*(i-1)+1}^{10*i} \frac{\tilde{P}_{RR}}{10}$$ (3.9)

where i ranges from 1 to 4 for these features. These four band features capture most of the energy in the spectrum, covering the range 0-0.4 Hz. Figure 3-2 shows an example of how these features represent the signal during each of the eight emotion episodes.

3.4   Feature Summary

Sample signals from the four physiological signals are shown in Figure 3-3. The annotations in the top graph indicate the times at which the subject expressed no emotion, anger, hate, grief, love, romantic love, joy and reverence. These signals appear to exhibit trends which differentiate the eight emotion states. The eleven features extracted from these signals ($\mu_{\mathcal{E}}$, $\tilde{\mu}_{\mathcal{G}}$, $\tilde{\delta}_{\mathcal{G}}$, $\mu_{\mathcal{B}}$, $\delta_{\mathcal{B}}$, $\tilde{\mu}_{\mathcal{R}}$, $\tilde{\sigma}_{\mathcal{R}}^2$, p_1, p_2, p_3, p_4) were used to capture the qualities of the signals. A summary of the features is given in
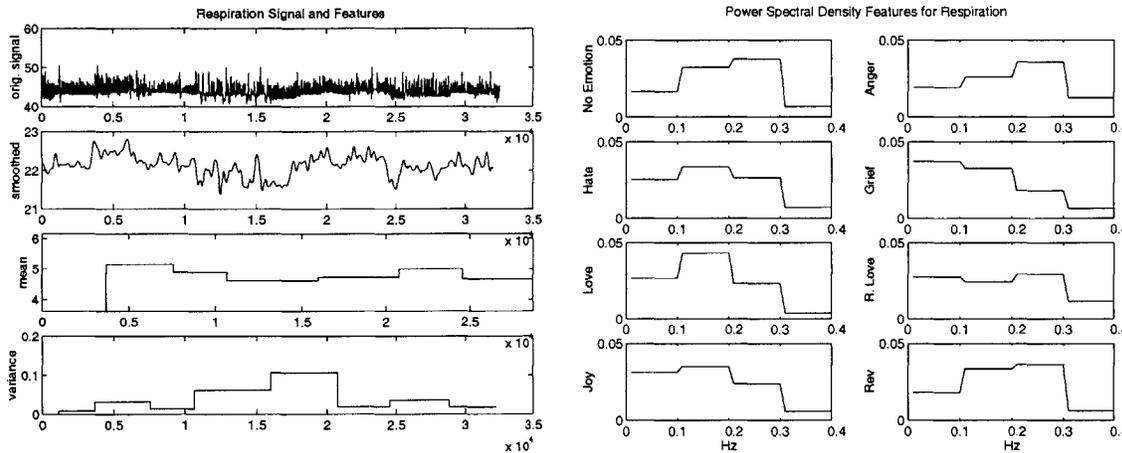
Figure 3-2: An example of how the respiration signal is represented by features in both the time and frequency domain. On the left, the raw and smoothed signals are shown along with the mean and variance features which represent them. On the right, examples of how each emotion episode is represented by features in the frequency domain.

Table 3.2. This table shows the mean value for each emotion across all twenty days. From this a rough estimate of how well each feature distinguishes the emotion states can be seen.

3.5 Pattern Recognition Results

To discriminate sets of emotion states, points in the eleven dimensional feature space representing the 160 emotion segments were projected into a two dimensional space using Fisher projection. In this projected space, emotion sets were modeled as Gaussian distributions and linear and quadratic classifiers were created using discriminant functions. Recognition rates were calculated using leave one out cross validation. These methods are well established in the pattern recognition community. A brief description is provided here to show how these methods were applied to the emotion recognition problem. This implementation is based on a description by Duda and Hart[DH73].

The Fisher projection matrix uses linear combinations of features to find a space

65
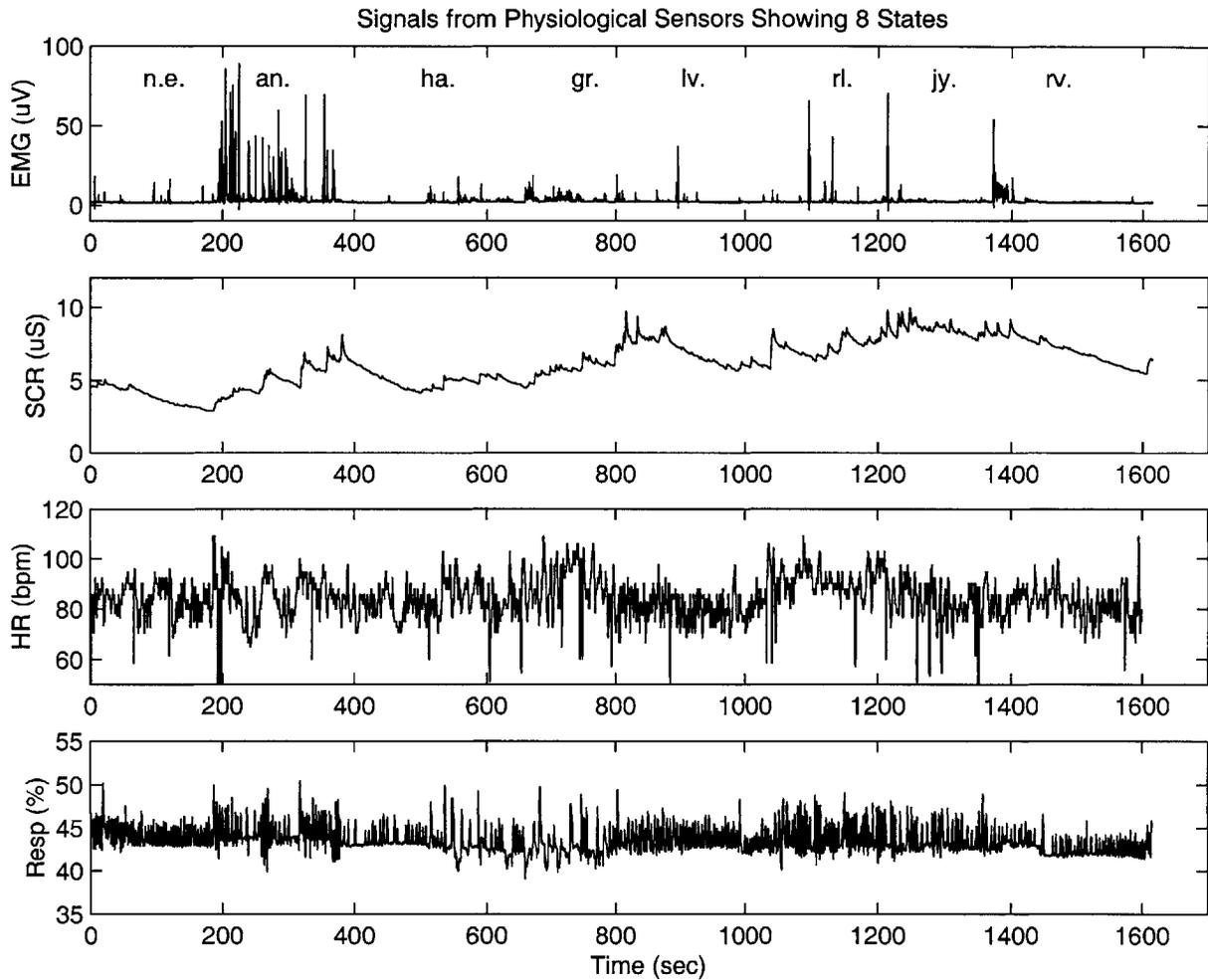
Figure 3-3: An example of a session's data collected from four sensors. Signals from the EMG on the masseter muscle in microvolts (top), the skin conductance waveform (in micro-Siemens), the heart rate (in beats per minute), and the respiration waveform (in % expansion) are shown. The annotations at the top on the EMG waveform indicate the periods during which the subject was asked to express no emotion, anger, hate, grief, love, romantic love, joy and reverence. These periods are the same for all emotions

66

	no emotion	anger	hate	grief	platonic	romantic	joy	reverence
μ_ε	1.0	3.2	1.5	1.9	1.2	1.5	1.7	0.9
$\tilde{\mu}_g$	.56	.68	.60	.54	.54	.56	.60	.57
δ_g	-.36	.76	-.82	.05	-.01	.01	.20	-.44
μ_B	78	86	80	81	80	83	80	78
δ_B	.0020	.0001	.0004	.0000	-.0030	.0001	-.0001	-.0007
$\tilde{\mu}_R$	.10	.14	.02	-.12	-.01	.01	.00	-0.2
$\tilde{\sigma}_R^2$	.02	.02	.02	.08	.02	.02	.06	.01
p_1	.32	.26	.34	.33	.43	.24	.35	.33
p_2	.38	.36	.27	.18	.23	.29	.24	.36
p_3	.07	.12	.07	.07	.04	.11	.06	.06
p_4	.03	.04	.03	.03	.02	.04	.02	.03

Table 3.2: The average values of the eleven features extracted from the data for the eight emotion states. Differences in the means across the features gives a rough estimate of how well the feature distinguishes between emotion states

Three Individual Emotions			Linear Classifier		Quadratic Classifier	
em 1	em 2	em 3	misclassified	correct	misclassified	correct
no em.	joy	reverence	1-7-4	80 %	2-5-4	82 %
anger	hate	romantic	6-4-3	78 %	5-4-3	80 %
anger	hate	reverence	5-4-3	80 %	5-3-3	82 %
anger	grief	reverence	2-6-2	83 %	3-4-1	87 %
grief	platonic	reverence	6-6-4	73 %	5-5-5	75 %
anger	romantic	reverence	4-5-3	80 %	4-6-3	78 %

Table 3.3: Subsets of three individual emotions were projected into the Fisher space and a good discrimination was achieved. These methods did not yield good results for greater numbers of individual emotions.
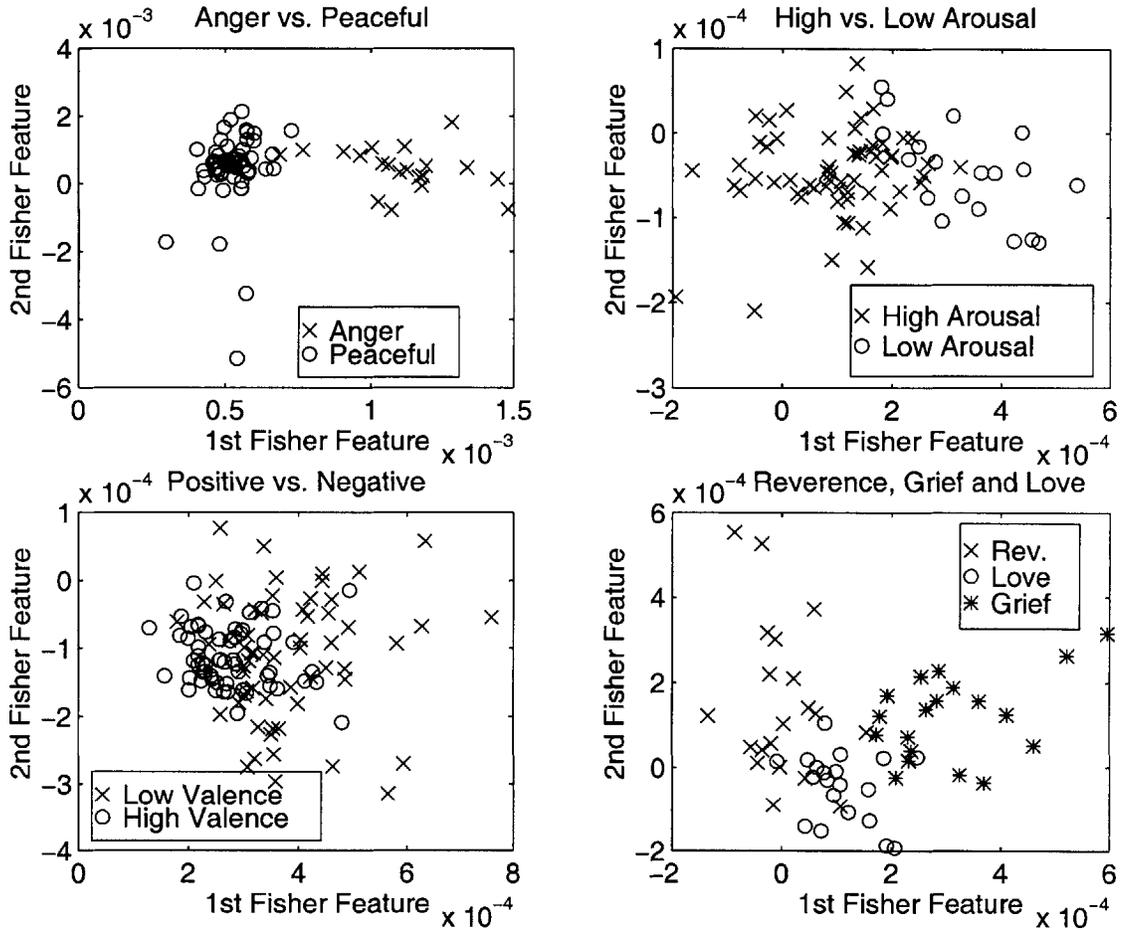
Figure 3-4: Points representing emotion episodes are projected onto the first two Fisher features. These graphs show how sets of emotions are separated in the projected space. The high arousal emotion Anger is well separated from the low arousal Peaceful emotions, however the Positive and Negative valence sets are confused in this space.

Sets of Similar Emotions		Linear Classifier		Quadratic Classifier	
Emotion Set	Set Size	misclassified	correct	misclassified	correct
anger	20	2	90%	0	100%
peaceful	60	1	98%	1	98%
high arousal	80	15	81%	16	80%
low arousal	80	11	86%	10	88%
positive	60	15	75%	11	82%
negative	60	29	53%	30	50%

Table 3.4: This table summarizes the results of discriminating between sets of emotions defined by differences in arousal and valence qualities. Anger was rated as a very high arousal emotion, it was discriminated from a class of "Peaceful" emotions including no emotion, reverence, and platonic love. All eight emotions were divided into classes of high arousal including anger, grief, romantic love, and joy; and a low arousal class containing no emotion, hate, love, and reverence. A positive valence class was created containing platonic love, romantic love, and joy; which was compared to the negative valence class consisting of anger, hate, and grief. Better discrimination was achieved for emotions separated along the arousal axis.

in which the classes are most well separated from each other and in which the distance between points belonging to the same class is minimized. This is a method which would produce the kind of optimal discrimination suggested by Cacioppo and Tassinary and shown in Figure 2-13 of the background chapter.

To find this optimal projection two quantities, the between-class scatter, S_B, to the within class scatter, S_W are defined by Equations 3.10 and 3.11:

$$S_W = \sum_{k=1}^{C} \sum_{x \in \chi_k}^{D} (x - m_k)(x - m_k)^t \qquad (3.10)$$

$$S_B = \sum_{k=1}^{C} D_k (m_k - m)(m_k - m)^t \qquad (3.11)$$

where C is the number of classes, representing the emotion episodes, D is the number of sample vectors in a class, corresponding to the number of days over which the data was taken, m_k is the sample mean for class k, m is the pooled mean over all classes, and x are the 11-dimensional feature vectors in each emotion class χ_k. For a classification involving the entire data set, $C = 8$ (1=no emotion, 2=anger, 3=hate,

4=grief, 5=platonic love, 6=romantic love, 7=joy and 8=reverence) and $D = 20$ representing each of the days in which a complete data record was taken. In this example, there would be 160 feature vectors, x, 20 belonging to each class χ_k. The Fisher projection matrix is found by solving the generalized eigenvector equation:

$$S_W^{-1} S_B w_k = \lambda w_k \tag{3.12}$$

where the w_k corresponding to the largest eigenvalues form the columns of the projection matrix W. This matrix is then used to project the test point onto the classifier space using

$$y = W^T x \tag{3.13}$$

In this analysis the points y exist in a two dimensional space, where linear and quadratic classifiers were calculated. These classifiers use discriminant functions described by the following functions:

1. Quadratic discriminant function

$$g_k(y) = -(y - m_k)^T \mathbf{K}_k^{-1}(y - m_k) - \ln |\mathbf{K}_k| + 2 \ln \Pr[c_k], \tag{3.14}$$

2. Linear discriminant function

$$g_k(y) = -(y - m_k)^T \mathbf{K}^{-1}(y - m_k) - \ln |\mathbf{K}| + 2 \ln \Pr[c_k], \tag{3.15}$$

where for each class k, m_k represents the sample mean for that class and K_k represents the covariance matrix for the class. For the linear classifier K is the sample mean of the individual covariances. $\Pr[c_k]$ is the prior probability of the sample belonging to class k. This prior is based on the number of samples in the class with respect to the total number of samples. In the cross validation procedure, a single point, x, is excluded from the data set and a Fisher projection matrix, W is calculated for remaining members of the set. The excluded point is then projected using that same

W and classified using the maximum discriminant function of both the quadratic and linear classifiers in the standard method described by Therrien [The92] and described by the rule:

$$\text{choose class } c_i \text{ when } g_i(y) = \max_K g_k(y).$$

This procedure was performed on subsets of individual emotions and subsets of groups of emotions. It was found that it was difficult to discriminate all eight emotion states. Some subsets of three emotion states were found to be well discriminated, such as anger, grief and reverence. The results of projecting three emotions into the two dimensional Fisher space and performing recognition with both linear and quadratic classifiers are reported in Table 3.3.

Other discriminations were made by taking sets of 4, 6 and eight emotions and projecting them into a two dimensional space defined by the Fisher features w_i corresponding to the two greatest eigenvectors. Each of the emotions was individually modeled by a gaussian distribution and linear and quadratic classifiers were calculated. Classification was performed on each of the emotion classes individually. If a point belonging to any of the emotion classes in a set was classified as belonging to any emotion class in that set the point was labeled as being correctly classified. If the point was classified as an emotion of the other set it was labeled as being misclassified. These sets were created to determine if emotions described as similar along the axes of arousal and valence could be well discriminated by these features. The results of this classification are shown in Table 3.4. The best discrimination was found between emotions differing along the axis arousal. For example, anger, a very high arousal emotion is well discriminated from a set of more peaceful emotions (no emotion, love and reverence). The entire set of eight emotions was well classified when grouped into sets of sets of high arousal (including anger, grief, romantic love and joy) and low arousal (including no emotion, hate, love and reverence) emotions. Conversely, discrimination was poor when the emotion states were grouped into sets of positive valence (including love, romantic love and joy) and negative valence (including anger, hate and grief), where neutral valence emotions were excluded from the sets. From

Number of Features which formed initial space	Without Day Matrix		
	SFFS (%)	Fisher (%)	SFFS-FP (%)
30: μ_X, σ_X, δ_X, $\tilde{\delta}_X$, γ_X, $\tilde{\gamma}_X$ $X \in (\mathcal{E}, \mathcal{B}, \mathcal{G}, \mathcal{R}, \mathcal{H})$	52.50	56.87	60.00
11: $f_1, f_2, \ldots, f_{10}, \mu_{\mathcal{E}}$	60.62	70.00	70.63
40: all of the above	65.00	77.50	81.25

Table 3.5: Comparative classification rates for eight emotions using combinations of the features used in this original analysis and statistical features derived and evaluated by Elias Vyzas[VP99]. Recognition rates are significantly boosted by the addition of the features described in this section, $f_1, f_2, \ldots, f_{10}$ and $\mu_{\mathcal{E}}$.

these results it can be determined that these features are better for distinguishing between emotion sets which differ along the axis of arousal.

3.6 Additional Results

Additional analysis on this data set and on a subset of the data collected from this experiment was done by Elias Vyzas, a researcher in our group. His analysis used the the eleven features proposed in this thesis plus 29 additional statistical and day-dependent features[VP99]. The results of his analysis using the features described in this section as $f_1 - f_{10}$ and $\mu_{\mathcal{E}}$ is shown in Table 3.5. Using these features along with his statistical features in Jain and Zongker's Sequential Forward Feature Selection Algorithm yielded a recognition discrimination rate of 81.25% for the eight emotions. The confusion matrix for this analysis is shown in Table 3.6 [VP99]. Algorithms were also created to analyze this data as if it were being presented to the computer in real time[VP99].

3.7 Summary

This study has shown that within the constrained conditions of this experiment (seated subject, relatively motionless, intentionally generating and expressing each

	N	A	H	G	P	L	J	R	Total
N	17	0	0	0	3	0	0	0	20
A	0	17	0	0	2	1	0	0	20
H	0	0	14	1	0	0	3	2	20
G	0	0	1	15	0	0	4	0	20
P	0	0	0	0	17	2	1	0	20
L	1	1	0	0	3	14	1	0	20
J	0	0	1	2	0	0	17	0	20
R	0	0	0	1	0	0	0	19	20
Total	18	18	16	19	25	17	26	21	160

Table 3.6: The confusion matrix for the method that gave the best performance in the classification of 8 emotions using both the eleven features presented here plus additional statistical features (81.25%)[Vyz99]. In this matrix an entry's row is the true class, the column how it was classified.

emotion) specific emotion patterns can be automatically recognized by a computer using physiological features. For emotions which were clearly different in their generation along the axis of arousal, 100% recognition was possible. Also subsets of emotions which had clearly different qualities were also well separated with recognition rates of 75-87%. Furthermore, the features used in this analysis were found to contribute significantly to a recognition rate of 81% for the set of all eight individual emotions found using Fisher projection and a Sequential Floating Feature Selection (SFFS) algorithm using a k-nearest neighbor (k-NN) classification criterion. This 81.25% recognition rate may be close to the best possible rate for recognizing the actual emotions generated given that the emotions were not generated with perfect consistency.

Emotions have been studied in the laboratory for decades; however, repeated studies using a single individual are rare. This experiment allowed the best possible circumstances for recognizing a single individual's emotions in a laboratory setting; however, it is not certain how well these acted episodes correlate to real emotional episodes outside the laboratory where their expression is not as encouraged. In real situations, for example, a person may not actively clench their jaw in anger with the same exaggerated expression as when trying to produce the emotion. The respiration

effects during true grief may also differ from the patterns found in self-produced grief. Naturally occurring reactions may be even more distinct and easy to identify, or they may be more subtle. The next two experiments present physiological data from natural situations in which affective states occur.

Chapter 4

Ambulatory Experiments

A wearable computer offers an unprecedented opportunity to collect and analyze large quantities of data in natural situations. Augmented with physiological sensors and the capability of real-time digital processing of these signals, a wearable computer can learn about the user's affective responses as they happen, where they happen and when they happen. Ideally, a wearable system would be as transparent and as easy to wear as a wristwatch, but currently this technology has not been reliably developed. This section presents new devices developed for ambulatory affect detection and the results of some data collection experiments in real world situations. In one experiment, a person wore the ambulatory system for a portion of the day while walking around and performing normal daily tasks. During this time, a digital video camera captured periodic images and the person made annotations of her activities which were logged with the physiological data. Another experiment consisted of a series of in-lab tests designed to determine how the effects of physical activity impacted on physiological signals. Although the range of motion in the ambulatory environment often obscured affective signals the devices here serve as excellent activity monitors and can be used to detect affect in low motion situations. Such a situation, automobile driving is explored in the next chapter.

4.1 Affective Wearables

To determine if natural affective responses could be captured and quantified in the unconstrained ambulatory environment, a wearable system for monitoring physiological signals was designed and tested. The ambulatory monitoring system consisted of a "Lizzy" wearable computer[SMR+97], physiological sensors and an analog-to-digital converter unit. The Lizzy wearable system is slightly bulky and could more accurately be described as "tote-able" rather than wearable, but it provided a comparatively robust base unit that has been used reliably by researchers for the past few years. Experimental prototypes of more lightweight, low power sensing units were also developed. These sensor designs incorporate wireless transceivers and conductive rubber electrodes, and have been embedded into clothing and accessories. Designs for embedded GSR, BVP and respiration sensors are presented.

The Affective Lizzy

The system used to collect the data in the ambulatory experiments was based on Thad Starner's "Lizzy" wearable. This design incorporated a PC104 card stack running Linux, a Private Eye head-mounted-display (HMD) and a Twiddler chordic keyboard. To monitor physiological signals this design was augmented with physiological sensors and a ProComp unit with an analog-to-digital converter [Tho94]. A PalmPilot interface was developed as an input-output device to replace the combination of the head mounted display and chordic keyboard. A digital camera was added to collect periodic snapshots of the wearer's activities throughout the day. Figure 4-1 shows the ambulatory system both with the Private Eye and Twiddler combination and with the PalmPilot. Both images show the ProComp unit and the physiological sensors, including an electromyogram for measuring muscle activity, a Hall-effect sensor for measuring respiration through chest cavity expansion, a photoplethysmograph for measuring pulse, and a skin conductance sensor.

This system is small enough to fit in a satchel and can be carried around unobtrusively. Replacing the head-mounted-display and chordic keyboard with the PalmPilot
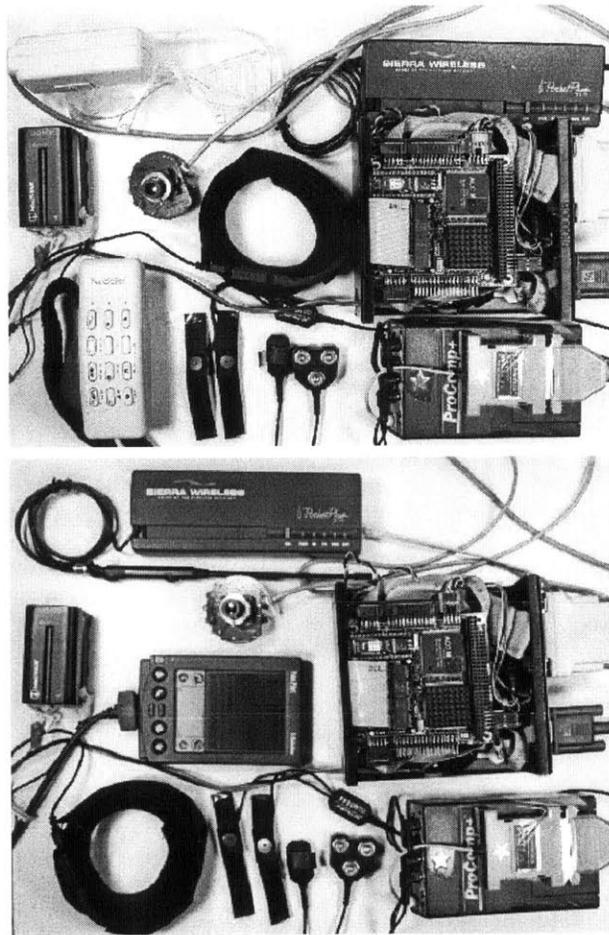
Figure 4-1: Two Lizzy based wearable systems. The top image shows a Lizzy with sensing system using a Private Eye HMD and Twiddler chordic keyboard. The components shown include, clockwise from top left, the Private Eye HMD, CDPD modem, PC104 based wearable [Sta95], the AD converter, EMG, BVP GSR, a chordic keyboard and the battery, and, center the digital camera and the respiration sensor. Another version shown in the bottom image uses the PalmPilot to replace the HMD and chordic keyboard.
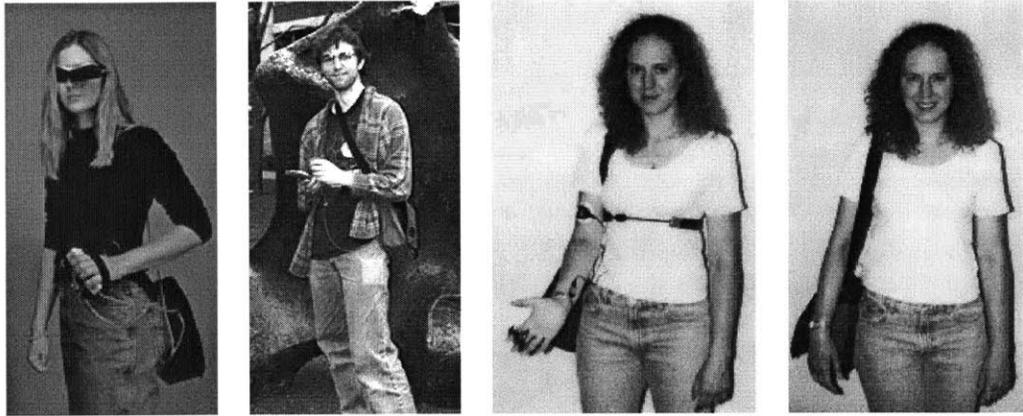
Figure 4-2: To make the wearable monitoring system less noticeable, a PalmPilot interface was developed to replace the head-mounted-display and chordic keyboard. The physiological sensors can be worn underneath clothing.

makes the wearable even less noticeable. The sensors can be worn under clothing to make them minimally obtrusive. Figure 4-2 shows four images of wearable systems. The first two show wearables using either the Private Eye and Twiddler or the PalmPilot interface. The third image shows the author wearing the physiological sensor and the fourth image shows the same system with the sensors hidden. The exposed sensors are shown with the respiration sensor worn around the chest on top of the shirt, an EMG worn on the bicep, a BVP attached to the wrist and a GSR sensor worn on two fingers. The hidden sensors include the respiration sensor worn under the shirt, the EMG sensor on the trapezius (back shoulder) muscle, the GSR sensor on the foot, and the BVP sensor on the side of the neck. The sensors are connected to the wearable computer worn in the satchel.

ADX Microprocessor Units

As an alternative to a bulky general purpose wearable computer with sensors, prototypes of very small, low power, low cost microprocessor based sensors were designed. These units, called ADX for "analog-to-digital converter," were designed to take analog readings of physiological signals at different points on the body, sample the signals using a low-power analog-to-digital converter, and send the digital signal back to a

central processing unit. A PIC series microprocessor controlled the timing of the signal collection, stored data and controlled error checking. The prototypes were primarily designed by Grant Gould, an undergraduate researcher under my supervision. His design consists of a base unit PIC16F84 micro-controller and an LTC1298 two-channel analog-to-digital converter buffered through an Analog Devices OP291 operational amplifier[Gou99]. The micro-controller also controls the transmission of information across a serial bus, an 'infra-red channel (IR) and short distance radio frequency (RF) Personal Area Network (PAN) [Zim96] transmitters. These sensors can operate at up to 9600 bits per second using the RJ11-style serial connection[Gou99]. Wireless transmission methods have more limited bandwidth not exceeding 2400 bits per second.

These ADX boards were coupled with several embedded sensors and tested in the laboratory. Figure 4-3 shows an ADX board attached to a GSR sensor embedded on a wrist-cuff with rubber electrodes and worn on the arm attached to two "ring" electrodes. The wrist-cuff arrangement uses PAN radio frequency wireless transmission while the arm-mount system uses wireless line-of-sight IR transmission. Figure 4-5 shows the ADX board attached to a BVP sensor on the ear. This sensor arrangement uses the RJ11 style serial connection to avoid RF near the head and because drifting strands of hair interrupted the line-of-sight for IR. Testing these sensors showed that signals could be transmitted at the 20Hz sampling rate, but that the prototypes were very fragile. Transmission methods were sensitive to motion artifacts for the embedded GSR and the signal from the embedded BVP sensors was confounded by head motion. Further work would have to be done to develop these embedded sensors into a working system. The prototypes are presented here to illustrate the goals of this experiment and to show the placement of sensors we found to be most useful.

Embedded GSR Sensor

Skin conductivity can be measured effectively from either the palm of the hand or the sole of the foot. The prototype embedded GSR sensor shown in Figure 4-4 is used to take a reading off the sole of the foot. The foot is a good place to take a
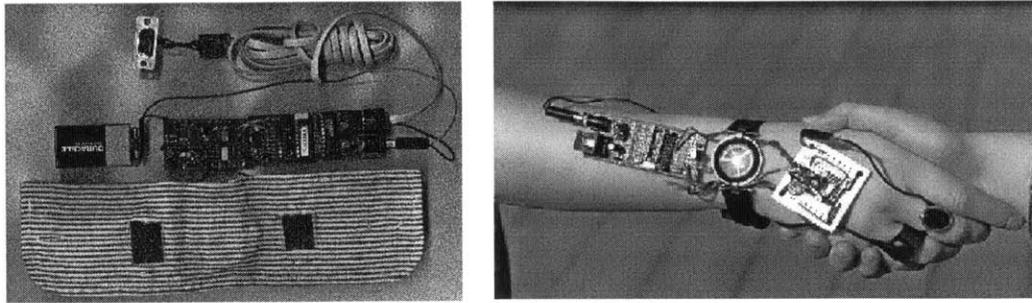
Figure 4-3: Two prototypes for measuring skin conductivity. Left shows a model using a conductive rubber electrode embedded into the cuff of a shirt sleeve attached to an ADX unit with RF transmission capabilities. Right shows two silver rings used as electrodes with an ADX using IR transmission.

reading because hands are used on many daily tasks which create motion artifacts and because hand washing effects the skin conductivity reading. However. the skin conductance reading from the foot is subject to pressure artifacts when the subject is walking.

In this prototype design, Ag-Cl electrodes are shown embedded in the insole of a shoe. The sensors and electronics used to capture the skin conductance reading are embedded in the heel of a shoe. A wire runs through the shoe to attach the sensors to conductive snaps which hold in the insole with the electrodes. Snap in insoles with disposable electrodes allow electrode replacement to be easy and inexpensive.

Embedded BVP Sensor

The pulse can be measured by using a photoplethysmeograph to measure the quantity of blood in the peripheral blood vessels. The sensor has two parts, a light source and a photo-detector to measure the amount of light that is reflected by the surface of the skin. Each time the heart beats, blood swells the capillaries, causing a change in the amount of light detected[Tho94]. By measuring the distance between these peaks in reflectance, heart rate can be calculated.

Both the light emitting device and the photo-detector are very small: in our sensor they are less than one centimeter in diameter. Therefore, this sensor can be embedded
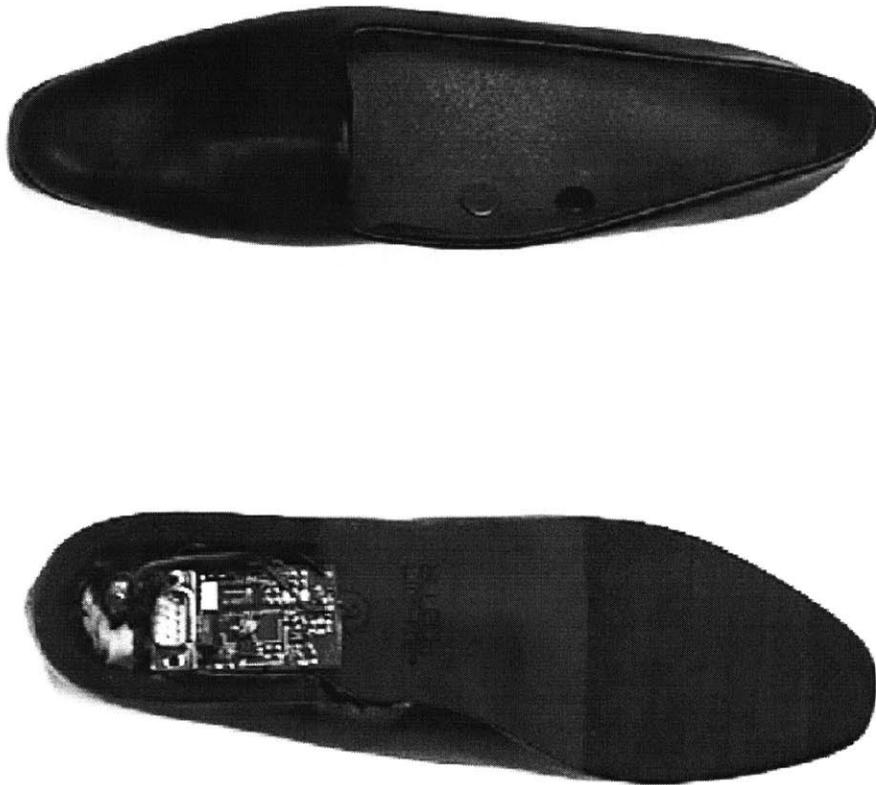
Figure 4-4: A prototype for measuring skin conductivity can be measured from the sole of the foot using electrodes embedded in the insole (top) and a microprocessor unit in the heel of the shoe (bottom).
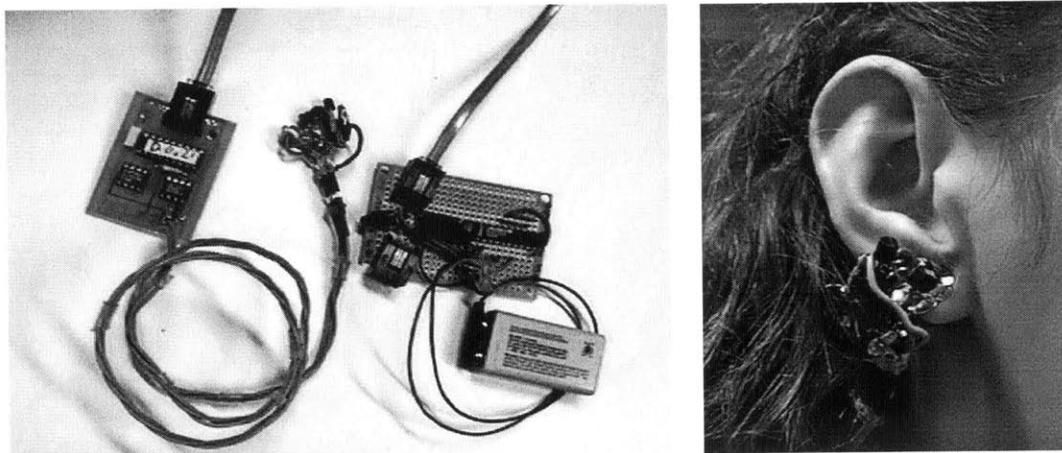
Figure 4-5: A prototype for pulse sensing. A photoplethysmograph is attached to the ADX using a wire instead of RF to avoid the risk of problems associated with having radio frequency emissions constantly near the head. The left image shows the entire system and the sensor as worn is shown right.
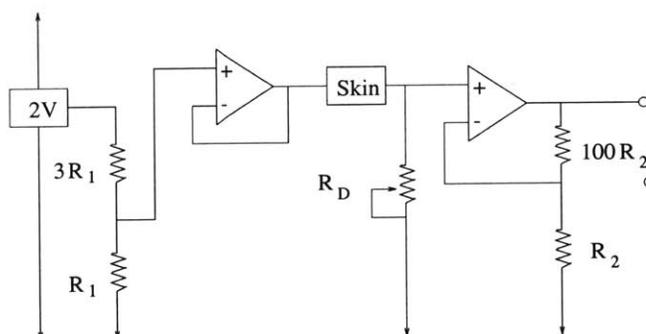
Figure 4-6: The circuit design for the GSR sensor[Gou98]. In this design, $R_1 = 100k\Omega$, $R_D = 5k\Omega$ potentiometer, and $R_2 = 10k\Omega$. Designed by Grant Gould
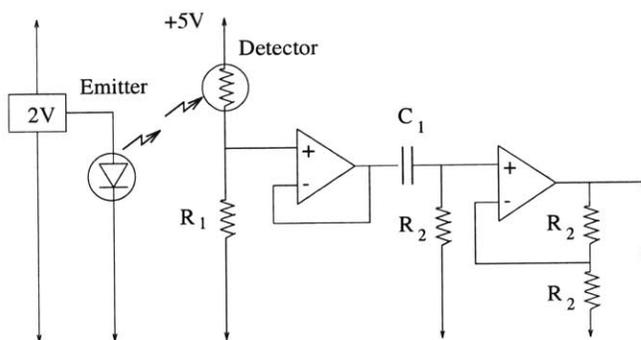
Figure 4-7: The circuit design for the BVP sensor[Gou98]. In this design, $R_1 = R_2 = 100k\Omega$, $R_D = 100k\Omega$ potentiometer. Designed by Grant Gould.
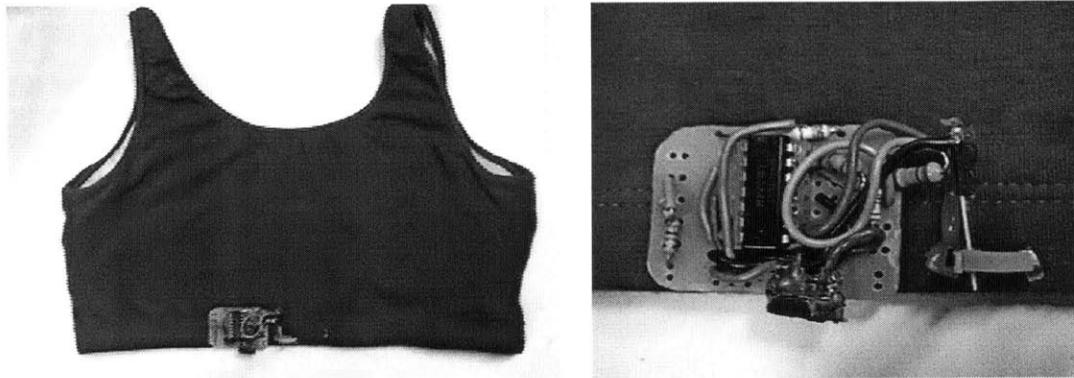
Figure 4-8: A prototype for respiration sensing through chest cavity expansion using a Hall Effect sensor. The left image shows the sensor in the location where it could be embedded into a sports bra. A close up of the sensor in the right image shows the analog components and the constrained magnets. This image does not show the attached ADX board.

into small pieces of jewelry, such as an earring. An example of this is shown in Figure 4-5. The signal from the earring is carried by a wire off the head to the ADX unit. This sensor can be used to measure heart rate as long as the person's head is relatively still. Motion artifacts overwhelm the reading when the person turns their head.

Embedded Respiration Sensor

An embedded respiration sensor was designed using a Hall effect sensor. This sensor detects respiration activity be measuring chest cavity expansion. When the chest cavity expands, the two magnets embedded in an elastic tube are pulled apart, generating an electric current which is measured by the ADX sensor.

This sensor could easily be incorporated into a woman's sports bra. The elastic band for the bra naturally sits at a good location for taking the respiration measurement. Figure 4-8 shows the analog sensor on top of the elastic band of a sports bra (where it would be embedded) and a close-up of the sensor showing two magnets constrained by elastic on the top and bottom. In an embedded version, these magnets would be encased in a tube of elastic.
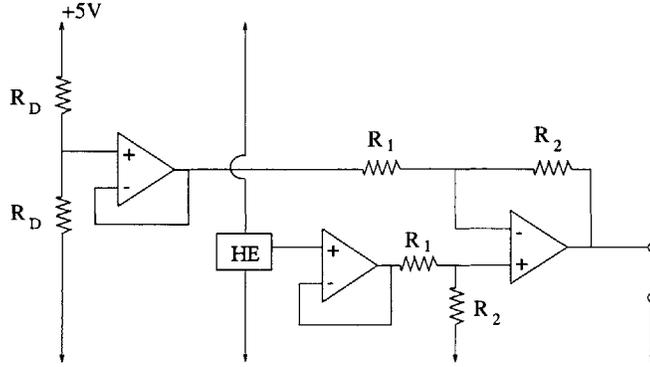
Figure 4-9: The circuit design for the BVP sensor[Gou98]. In this design $R_1 = R_2 = 100k\Omega$, $R_D = 100k\Omega$ potentiometer. Designed by Grant Gould.

4.2 Daily Monitoring

A preliminary experiment was performed to determine the feasibility of conducting ambulatory affect detection. Using the Lizzy wearable system with PalmPilot interface, four physiological signals were collected from the EMG, BVP, respiration and skin conductance sensors over the course of a day while the user went about normal activities. In this experiment, the EMG was placed on the left trapezius muscle, the GSR electrodes were placed on the index and middle finger of the right hand, the BVP sensor was placed on the ring finger of the right hand and the respiration sensor was placed around the diaphragm. Annotations were added manually by the user through the PalmPilot interface. The computer automatically time-stamped the annotations. During the day the subject went through various physical activities and made occasional comments. Figure 4-10 presents the data with the annotations placed where the computer marked them in the log file.

The anecdotal results of this experiment illuminated the main challenges of self-reported ambulatory affect detection: the presence of motion artifacts and the difficulty of capturing and coding physical and emotional events. These challenges are illustrated by the data in Figure 4-10. The EMG record, for example, shows far more activity in the morning than later in the day. Although this muscle tension might be attributed to an emotional episode, it is far more likely that this tension indicates motor activity. The subject was carrying the wearable computer on the left shoulder

in the morning and the recorded tension was most likely the muscle action necessary to support the wearable. The decrease in activity at 9:40 after the "coffee" incident was probably due to the subject's supporting the computer in her lap. The following events explain the difference in the left shoulder motor activity in the afternoon. During the time labeled "thesis," the computer was resting on the subject's desk and later in the afternoon the subject carried the computer on her right shoulder. The activity of lifting the device would have to be noted if muscle tension due to emotional episodes is to be discerned from muscle activity due to these motor actions.

The GSR data record is also sensitive to motion artifacts, such as pressure and electrode motion and it is difficult to know when these artifacts are present because of an insufficiently rich labeling system. The GSR measures conductivity along the path from one finger through the palm of the hand to the other finger. When pressure is exerted on the electrodes, greater contact occurs between the skin and the electrode and conductivity increases. During typing activities such as "thesis" a high variance motion artifact occurs as the fingers move rapidly. The problem of insufficient labeling makes these artifacts difficult to discern. For example at 9:30, a spike occurs in the data at the time labeled "coffee." This spike may be due to an emotional episode, physical activity, caffeine intake, or pressure from holding the coffee cup. Further, the record shows a significant yet unlabeled degree of activity at 4:30. This precedes the annotation "eat dinner" and is probably due to the subject's getting up, walking to the cafeteria and climbing two flights of stairs. These activities appear in the image record, described later, but were not noted by the subject.

The BVP data was very susceptible to motion artifacts, both in terms of sensor motion with respect to the skin and changes in the capillary blood volume due to the motion of the hand. The magnitude of the reflectance varied whenever the photo-detector placement changed with respect to the skin. The sensor could not be affixed very tightly, because doing so would cut off the pulse. This requirement and the placement of the sensor on the finger created situations in which the photo-detector moved with respect to the skin. The second type of motion artifact involved motion of the hand with respect to the body, which actually changed the blood volume in the

capillary. When the hands were raised above the body or when the arms swung freely during walking, identifying a definite pulse train was extremely difficult. Incidents of sensor failure also occurred. For example, at 11:45, during the coughing episode the signal vacillates from 0 to 100%, which usually indicates a loose connection between the sensor and the computer unit. To be used effectively for heart rate detection in an ambulatory situation, the BVP sensor should be placed at a better location or a sensor less susceptible to motion artifacts should be used.

The respiration sensor was well placed to avoid motion artifacts; however, the sensor did occasionally slip and required readjustment. Such an adjustment probably occurred at 4:45 PM when the baseline of the recording changes significantly. This record also suffers from a lack of consistent labeling of respiratory events. A significant coughing episode occurred at 11:45 and was labeled. Similar disturbances occurred at 1:30PM, 2:15PM, 2:45PM, 3:30PM, 4:20PM and 5:20PM, but the cause of these are uncertain. A more complete record of events that effect respiration would help distinguish these events from emotion related signals.

Although the results of this experiment are anecdotal, they illustrate the significant challenge of reliably detecting physiological signatures of emotion in ambulatory data. All of the user's motions and activities must be taken into account, including eating, talking, getting up, sitting down, carrying objects, and walking. There must also be a more complete and reliable method of annotating events, one that does not rely on the user for input. The data recording system also used a computer controlled camera to capture images at a rate of two per minute. Figure 4-11 shows an example collection of these images. This record can give a good overview of the subject's activities, helping them recall and annotate the day's events. However, for better judgment of affective states a continuous video record which also assessed the facial expression of the wearer would be preferred.
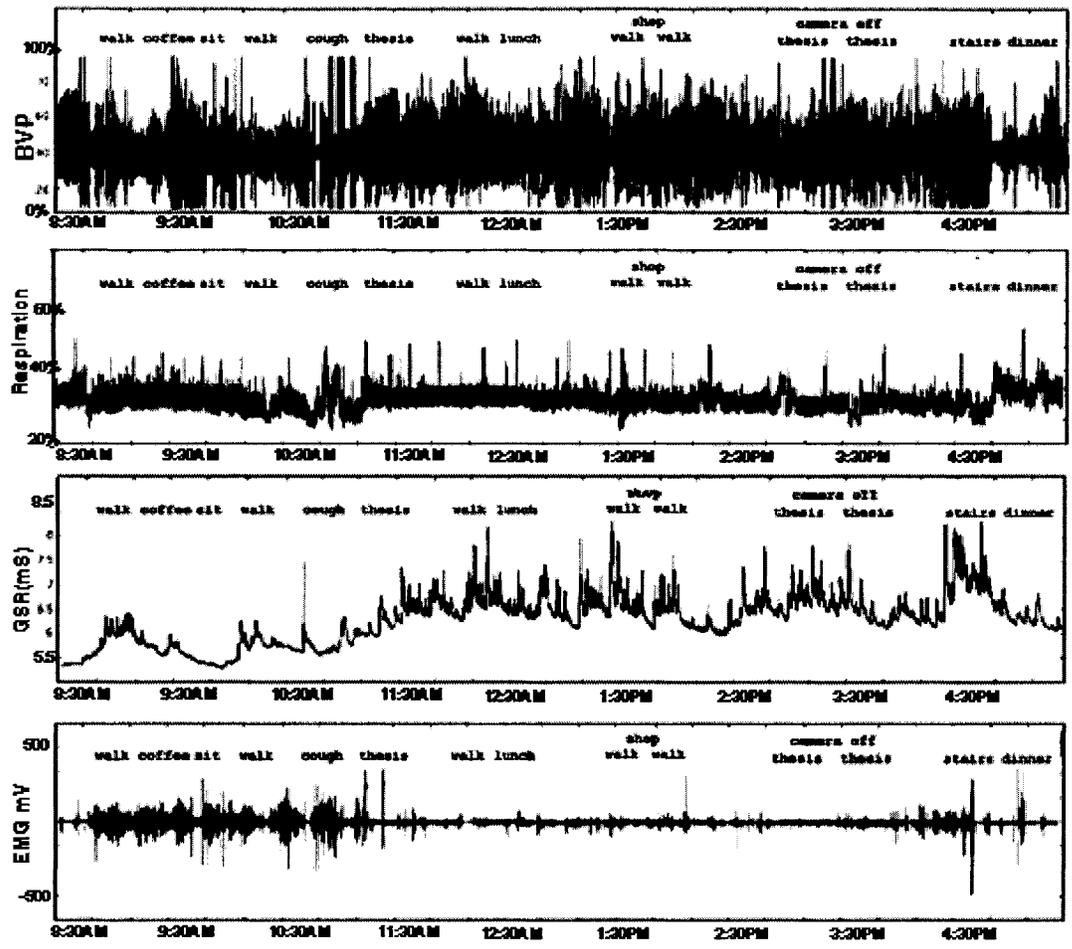
Figure 4-10: A physiological record from four sensors taken during a normal day's activities. Shown here are the traces from the BVP sensor, the respiration sensor, the GSR sensor and the EMG sensor. The included annotations were entered by the wearer at the times recorded by the computer. This record shows increased EMG activity in the morning when the subject was carrying the device on the shoulder measuring EMG activity. In this record physical activity was more readily captured than affect and emotion episodes were sparse and not recognized by the subject.

Figure 4-11: The digital camera automatically took snapshots once every two minutes. The camera was worn on the strap of the wearable satchel. Without any annotation from the user, the activity of the user can be determined. These images show the wearer first eating a snack while working at the computer, then getting up and walking to the kitchen, washing a dish, replacing some milk, then going outdoors.

88

4.3 Physical Activity

To better understand how physical activity affected the sensor readings, a series of controlled experiments was conducted in the laboratory. In this experiment, five subjects wore a respiration sensor, a BVP, and two GSR sensors, one on the hand and one on the arch of the foot. The subjects were then asked to perform the following tasks: sit in a chair for one minute, stand up and sit down twice, walk around the room for one minute, sit in a chair normally for two minutes, stand up and walk around the room for one minute, sit normally for another minute, jog in place for one minute, sit for another minute and finally intentionally cough repeatedly. An experimenter remained in the room with the subjects at all times during the experiment to instruct the subjects. Subjects sometimes had questions during the experiment, and this was noted in the experimental record. The times for the tasks were marked with a special sensor which made a spike in the recording data when depressed. Figure 4-12 shows an example of data collected from this experiments.

The data for this experiment was saved at 16 samples per second. Two features were calculated: the change in heart rate and the change in the skin conductance on both the hand and the arch of the foot. The change in the skin conductivity (GSR) was calculated as the difference between the signal taken at the time each task was begun and the first significant local maximum. The heart rate was calculated from the peak-to-peak intervals of the BVP. To calculate the change in heart rate due to the activity, the average heart rate over the ten second period preceding the activity was subtracted from the average of the heart rate over the ten second period following the activity. Table 4.1 tabulates the results of this experiment, showing that even common physical activities such as standing and walking can increase heart rate and skin conductivity readings significantly. The change in skin conductance in the hand and the foot showed different magnitudes for different subjects, perhaps because of individual differences or differences in sensor placement. The change in the foot's skin conductance is increased in many of these activities because of increased contact between the electrode and the skin from the pressure of standing. The hand skin

Activity	Increase in HR (bpm)				
	S1	S2	S3	S4	S5
stand	16.0	15.8	19.7	22.0	15.2
walk	26.4	18.9	27.7	24.6	19.9
jog	68.2	60.8	74.0	80.9	87.4
cough	22.0	22.2	18.8	53.7	14.7
Activity	Δ GSR Hand (μSiemens)				
	S1	S2	S3	S4	S5
stand	0.5	10.6	0.2	1.7	N/A
walk	2.0	14.3	0.8	2.8	N/A
jog	5.9	16.1	3.5	2.2	N/A
cough	5.4	11.0	2.1	2.7	N/A
Activity	Δ GSR Foot (μSiemens)				
	S1	S2	S3	S4	S5
stand	0.5	12.6	0.3	4.1	3.7
walk	1.5	10.4	1.7	6.9	3.4
jog	11.5	11.0	5.0	5.0	6.4
cough	6.9	9.9	2.4	5.9	3.2

Table 4.1: A wearer's activities can cause large changes in physiological signals. These changes need to be understood so that they can be taken into account by the system trying to recognize affect.

conductivity data for Subject 5 (S5) was not available because the sensor became detached during experiment.

Physiological emotion signals are often secondary in strength to physical signals. In psychological studies by Lang [LGea93] and Winton, Putnam and Krauss [WPK84], the greatest changes in skin conductance and heart rate in their study were 0.6 micro-Siemens, and 8 beats per minute respectively. The results presented in Table 4.1 show that simple physical activities such as standing can create physiological changes which exceed these limits.

4.4 Summary

Wearable computers augmented with physiological sensors offer a new tool with which to study emotion in natural situations. Through the use of this system a wealth of
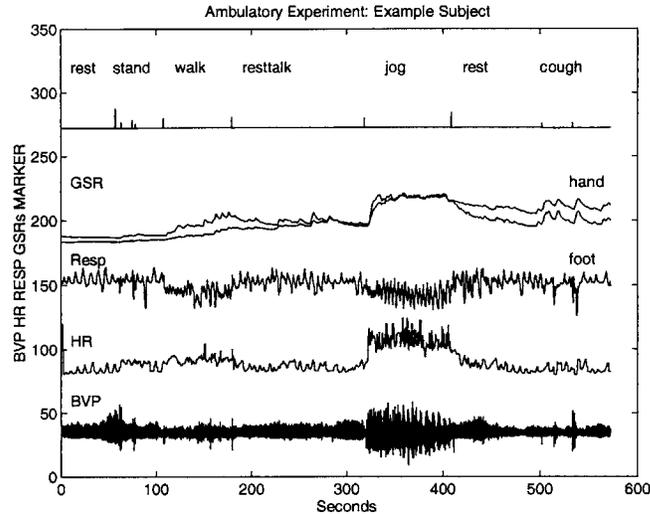
Figure 4-12: An example of data collected from the ambulatory experiment. Data is scaled and offset for show the relationships between the signals.

understanding about how emotions actually occur in daily life could potentially be gained. A wearable platform allows laboratory monitoring outside the laboratory. Now the physiology of emotion can be studied in the real world where emotions naturally occur. In the future, new experiments and a better methods of automatically capturing and codifying context may help reduce confounding artifacts. The hardware systems to create an affective wearable are within the scope of current technologies. Several of the prototype models could be developed into more robust wearable sensing systems. Multiple EMG sensors could be used to better understand and quantify motion. Tiny digital video cameras could replace the digital snapshot record and record action from multiple viewpoints. In the future, vision algorithms might be able to automatically interpret these video streams to provide automatic labeling to aid in coding affect states. As the system currently stands, it provides an excellent activity monitor.

Capturing the affective state to properly label physiological data is one of the greatest challenges of this research. To lessen the effects of motion artifacts and to limit the scope of the subjects actions, a new experiment was designed using a car as a test-bed. In this new framework multiple video cameras captured context and facial expression and were time synchronized with the physiological data stream to

provide a new dimension of ground truth labeling. This experiment is presented in Chapter 5.

Chapter 5

Stress Recognition in Automobile Drivers

Stress recognition in drivers is of growing concern because of the increased availability of on-board electronic appliances (e.g. cell phones and navigation aids) which can distract from a driver's attention. In driving situations when there is a high task demand and when the driver is experiencing a high degree of stress, automatic management of these appliances may be desired. The automobile also offers an excellent platform for measuring long term changes in a persons overall stress level. A regular commute to work provides a relatively constant sequence of events over which to compare relative stress levels from day to day and the physical structure of the car provides a large amount of space in which to embed computers, cameras and sensing systems and also a power supply for these devices.

The design of the stress recognition experiment built on the findings of both the eight emotion recognition experiment and wearables experiments. From the wearables experiments, it was determined that natural situations often made the detection of affective signals difficult. Difficulties included the confounding influence of motion, the problem of the subject not being able to identify specific emotions as they were experienced on a moment to moment basis and the lack of any objective fine grain documentation of the subjects context or reactions to events. Automobile driving is a natural situation in which ambulatory artifacts are limited and the context of
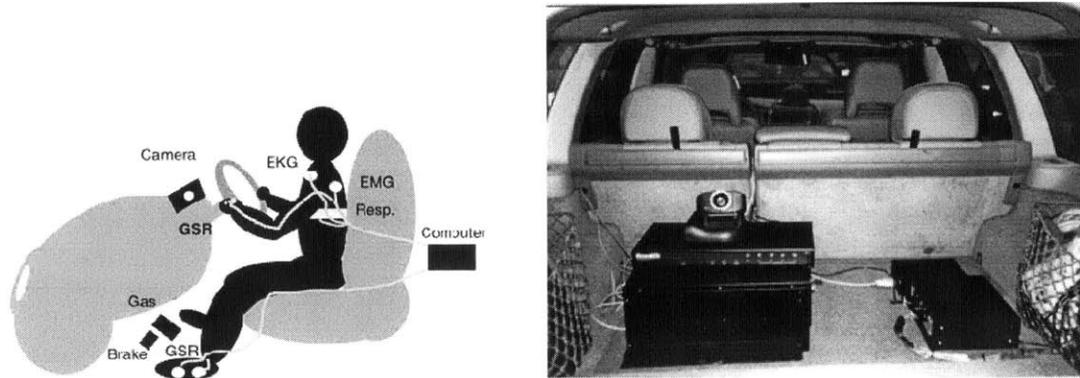
Figure 5-1: The subject wears physiological sensors as shown in the left image. These sensors are connected to a computer in the rear of the car, shown under the camera in the right image.

the drivers experience was captured by multiple video cameras and a microphone. The eight emotion identification experiment showed that emotions differing along the axis of emotional arousal were most easily identified using the features and methods presented here. Given the difficulties of natural emotion detection, the scope of this experiment was limited to the detection of only one emotion state, that of emotional stress, an axis closely linked to emotional arousal. To classify incidents of driver stress, three metrics were used: task design, questionnaire analysis and a second by second video tape annotation of events which might reflect the driver's stress level.

Two new types of features were also used in this analysis, features of the skin conductivity orienting response, and measures of autonomic balance as calculated from ratios of the short term power spectrum of the heart rate. These features were found to be very useful at discriminating stress according to all three metrics used. The results of this analysis suggest the possible use of this system for the management information applications and for the purpose of evaluating automobile and road designs.

5.1 The Automotive System

The automotive system was designed using a Volvo S70 series station-wagon. To create the sensing system, I installed a DOS based operating system in the car's

94

on-board computer and a laptop to remotely control the car's computer, a DSP board for data acquisition, a set of five physiological sensors, three video cameras, a microphone, a quad-splitter and a Hi-8 tape recorder. The physiological sensors were attached to the computer in the rear of the car through a fiber-optic cable that kept subject optically isolated from the car's electrical system, preventing any danger of electric shock. During the driving runs, the observer used the laptop to control the car's computer and monitor the physiological signals as they were being recorded. The video output of the laptop was collected with the output of the three video cameras and the microphone onto a single Hi-8 tape using the quad-splitter. This synchronized the physiological signals with the record of driving events. Figure 5-1 shows the on-board computer, bolted to the rear of the station wagon, in the lower left, the video quad-splitter and a digital camera are stacked on top of the computer. The power supply for the computer, video equipment and observer's laptop are shown in the lower right of this image. Figure 5-2 shows frame from the composite record from the quad-splitter. This record was later independently coded to create a metric of stress for labeling and validation.

During the experiment the subject wore four types of sensors at five locations as shown in Figure 5.1. For the purpose of illustration, the sensors are all shown worn on the left side. The EMG and GSR on the hand were actually worn on the right side, symmetrically identical to their placement in the diagram. The electromyogram was placed on the right trapezius muscle, on the upper back, near the shoulder. The electrocardiograph was placed on the torso, underneath the subjects clothing. The respiration sensor was placed around the subjects diaphragm on top of their shirt. The two GSR sensors were placed on the hand and on the foot, with the wires secured so that driving was not impeded. A more detailed description of sensor application is provided in Chapter 2.

The sensors were attached to a FlexComp sampling unit (identical in appearance to the ProComp unit shown in Figure 4-1) and data was transmitted via a single fiber optic cable to the car's on-board computer in the rear of the vehicle. The FlexComp unit was rated to be able to sample eight channels at up to 1984 samples

Figure 5-2: A sample frame from the video collected during the experiment. Views of the driver facial expression (upper left), body movement (upper right) and road conditions (lower left) are combined with the microphone input and a visual record of the physiological signals (lower right)
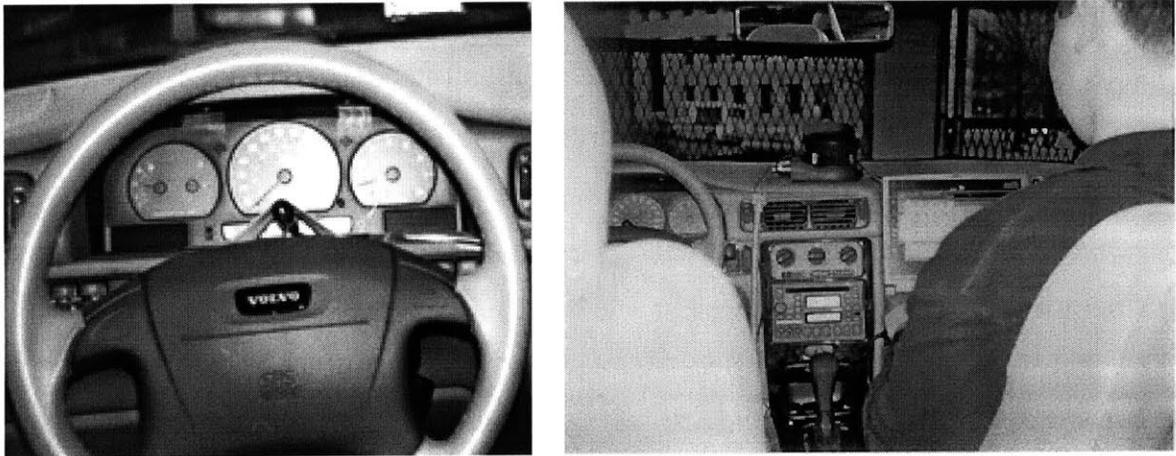
Figure 5-3: A small camera is placed on the steering column to capture facial expression while driving. A second camera monitors driver body motion and allows the observer to mark events using index cards.

per second [Tho94]; however, it was found by the author that sampling above 496 samples per second caused signal dropout across more than one channel. Also, the author discovered that the "16 samples per second" setting actually saved data at a rate of 15.5 samples per second, probably due to the fact that the unit is designed to save samples only at rates in multiples of 31 samples per second. Caution is advised in taking any of the system's reported specifications at face value. The type cumulative errors in time synchronization which can occur due to dropped bits and averaged sampling rates can become quite significant during a ninety minute drive; for example, a signal event recorded at the 15.5 second rate would be over 87 seconds off from the actual time the event occurred if the assumed 16 sample per second rate were used.

Three video cameras captured the context of the drivers' stress, one mounted on the steering wheel to capture facial expression, another mounted on the dashboard to capture road conditions and a third in the rear seat to captured body motion and indicate the roughness of the road. The camera positions are shown in Figure 5-3. The camera on the steering wheel was placed behind the airbag release point to ensure the driver's safety and a wide angle lens (.42) on the dashboard camera to capture a wider view of the road. The third camera was originally intended solely to be used by the observer to mark events using index cards; however, information from this camera was also valuable to determine the driver's body motion and the roughness of the road from the amount of jitter in the video image (it was the only camera mounted on a soft surface - the rear car seat). A small microphone was placed on one side of the driver's headrest to record the driver's voice and the noise from the road.

An observer was also present at all times during the experiment. The observer was seated in the right rear seat of the station-wagon, diagonally in back of the driver. The observer monitored the physiological signals from the driver using a laptop and was also available to answer the subject's questions. The laptop was a Toshiba Tecra780 that remotely controlled the on-board computer in the rear of the vehicle. This laptop had an RCA video output which was used to record the screen's output onto the Hi-8 tape through the video quad-splitter.

5.2 Experimental Design

This experiment was designed to capture naturally occurring emotions, and therefore was subject to far more uncertainty in terms of when such emotions were going to occur. The stressors were not controlled, but the driving protocol was designed to take the driver through situations where stressors were more or less likely to occur. Major road types and obstacles that might typically be encountered in a daily commute were designed into the task with the addition of two rest periods at the beginning and end of the drive. The assumptions of the design were that the resting periods would provide the least opportunity for stress, the uninterrupted highway driving periods would provide a low opportunity for stress, that the city driving periods would provide a higher opportunity for stress and that incidental encounters such as the garage exit, tolls, the exits for the turnaround and the two lane merge would provide the highest opportunity for stress. These assumptions were mainly supported by the subjective report of the drivers, however some exceptions arose as explained in the section on Questionnaire Analysis.

The session began with a fifteen minute resting period while parked in MIT's East Garage. During this period, the motor was running and the drivers were asked to rest with their eyes closed. When the rest period was over, the subject put the car into reverse and exited the garage. This exit was presumed to be stressful because the exit ramp is narrow and spirals down six floors. Following the exit, the subjects were instructed to drive through the city of Cambridge on Massachusetts Avenue. The subjects then took a left onto River Street and crossed the River St. Bridge.

The subjects encountered a toll booth before beginning the highway driving period along the Massachusetts Turnpike(Interstate 90). After the toll booth the drivers were instructed to stay in the right hand lane and maintain the speed limit, 55mph. They were cautioned to avoid exiting the turnpike before the second toll at Route 95. The highway driving was designed to provide a relatively uninterrupted driving task. After the second toll at Route 95, the drivers were to exit the Mass Pike going west and turn around through a series of exits to get on to the Mass Pike going east.
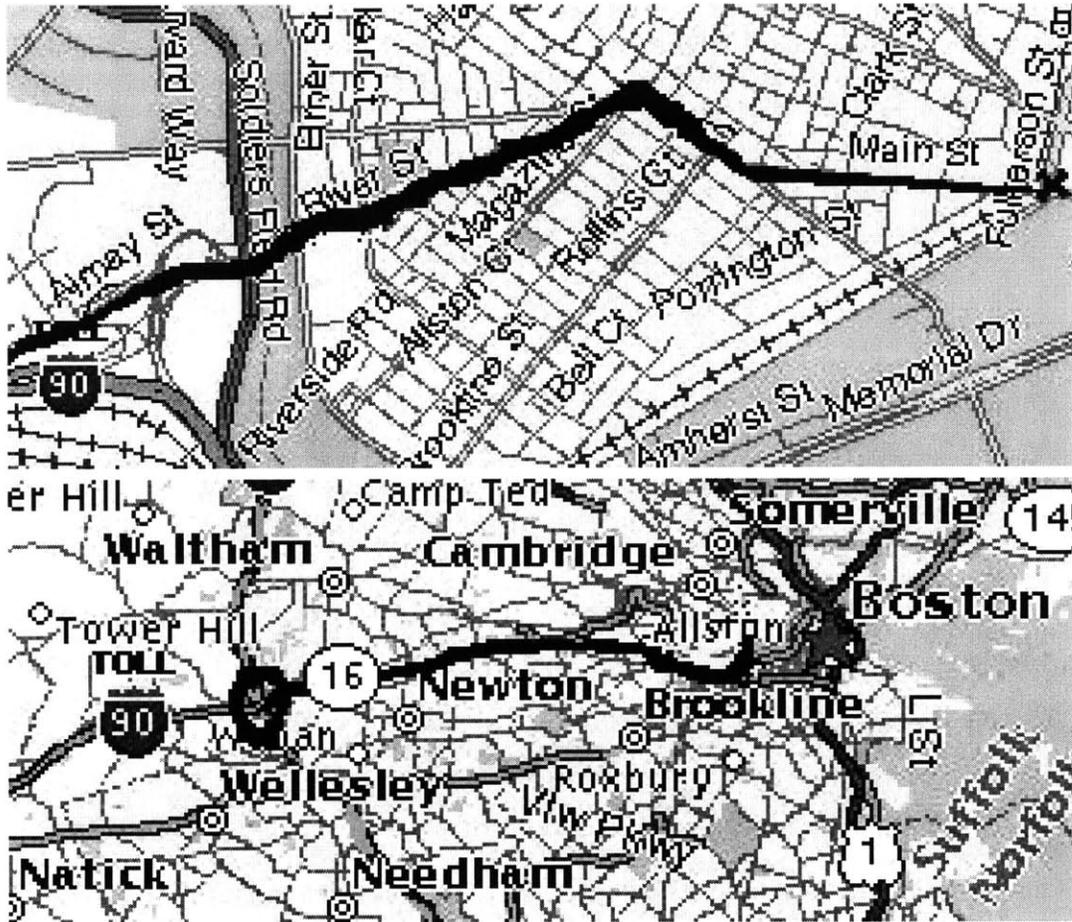
Figure 5-4: The driving route contained periods of city and highway driving. The route began on Main St., continuing to Mass Ave and then to River Street shown in the top map. From River Street, drivers continued along Interstate 90 (the Mass Pike) to Interstate 95, shown in the bottom map then turned around and returned along the same route.

Event Num.	Event Description	Stress Rating
1.	Beginning stationary period (rest1)	1
2.	Garage Exit	2
3.	City Road (city1)	4
4.	Toll Booth (toll1)	3
5.	Highway driving period (hwy1)	3
6.	Toll Booth (toll2)	3
7.	Exit Ramp Turnaround (exit)	5
8.	Toll Booth (toll3)	3
9.	Highway driving period (hwy2)	3
10.	Two Lane Merge (merge)	5
11.	Toll Booth (toll4)	3
12.	Bridge crossing (bridge)	4
13.	City Road (city2)	4
14.	Enter Garage	1
15.	End stationary period (rest2)	1

Table 5.1: A summary of driving events and the median stress rating from the ten questionnaires

They encountered a third toll before entering the Mass Pike East. On the return trip, they were asked to drive in the second to right lane until they saw the signs for the Allston-Cambridge exit. The second to right lane was chosen to provide the best opportunity for an uninterrupted highway drive since the far right lane diverged into an exit.

The merge across two lanes to the Allston-Cambridge exit was intended to induce stress in the drivers. The drivers were instructed not to begin the merge until they saw the sign for the Allston-Cambridge exit to achieve the longest uninterrupted highway run on the return trip. Following the exit, they encountered the fourth toll and an exit ramp leading to a bridge. After crossing the bridge, the drivers followed the reverse path through Cambridge down Massachusetts Avenue and Main Street and re-entered the MIT East Garage. The route of the experiment, including rest periods, was completed in approximately an hour and a half. The maps of the actual driving route are shown in Figure 5-4 and a summary of the driving events, in order of occurrence, is given in Table 5.1.

5.3 Subject Pool

This experiment was originally designed to test differences between individual drivers and within individual drivers. For this purpose the subject pool was comprised of three subjects who repeated the experiment multiple times and six subjects who completed the drive only once. In total 27 driving runs were attempted, however portions of some drives were not used due tolost data and deviations proscribed driving route. Appendix A gives a listing of all the drives comprising the database for this experiment and the errors which occurred.

Subject 1 was a male undergraduate with three years of driving experience who had not driven regularly for the past three years. Subject 1 also attempted seven driving runs. In the first driving run, the skin conductivity sensor became detached, creating an incomplete record, in the second run, the subject took a wrong turn and altered the driving route and in the third run the resting data was accidentally not recorded. For Subject 2, perceived stress was positively correlated with experience with a correlation coefficient of 0.41, i.e., perceived stress increased with successive driving runs. The sum of the stress ratings was 86, higher than subject 1. This number is unusually high. The average of the stress ratings for the six subjects who were all novice drivers was 63.

Subject 2 was an undergraduate male student with over four years of driving experience. He had not driven a month previous to the experiment and the first drive in the experimental vehicle was his second driving experience in Boston. Subject 2 attempted seven driving runs, for the first run there is no record of the driving events and the second run was incomplete due to a minor accident and was excluded from the questionnaire analysis. Overall, Subject 2 did not find driving to be a stressful experience. His highest rating on the absolute stress scale was a single "3" for the first experience with merges and exits. For Subject 1, the sum of the absolute ratings was negatively correlated with experience with a correlation coefficient of -.84, i.e. perceived stress decreased with successive runs. Sum of stress ratings was 41.

Subject 3 was a female undergraduate with eight years of driving experience.

Data was lost on the turnaround of the first driving run, the driver got off the driving route on the second driving run and on the sixth day the driver was forced to take an alternate route because Mass Ave was closed.

Six remaining subjects participated in the study. Of these subjects, the second subject had an unusual driving experience because she took a wrong turn on the highway and was unusually agitated second rest period; the fourth subject had the hand GSR become detached and the fifth subject had a faulty EKG signal.

5.4 Creating A Stress Metric

Three metrics were used to assess driver stress level: a task based metric, a questionnaire based metric and a metric based on second by second annotations of the video tape record for perceived stressors. The task metric was based on three major conditions designed into the experiment: two resting periods that were assumed to be low stress, two highway driving periods that were assumed to be of medium stress and two city driving periods that were assumed to be of high stress. A second metric used questionnaire analysis to establish four stress categories based on ratings of different events during the drive. Finally, a metric was developed from the annotations made by independent coders of the video records. To compare these categories a mean rating from the video tape metric created for each of the task based and questionnaire based categories.

5.4.1 Questionnaire Analysis

To validate the assumptions of the experimental design and to create a metric of perceived stress, a questionnaire was administered to drivers following the completion of the driving course. A copy of this questionnaire is included as an Appendix C. The questionnaire contained three sections: an information section that recorded the driver's name, age, driving experience and perceived starting stress level; a section that rated types of driving events on an absolute scale and a section that created a comparative rating of driving events using a forced scale.

The absolute scale rated five types of driving events: rest periods, city driving, highway driving, tolls and merges and exits. The subjects rated these categories on an absolute scale of 1 to 5 (1 =no stress and 5=high stress). The median results for each of the subjects and the total median are shown in Table 5.2. This scale has little dynamic range. Many times subjects rated all events using only two numbers (e.g. 1 1 1 2 2).

The comparative scale was designed to force a finer grain evaluation of driving events. On this scale, drivers were asked to rate fifteen sequential driving events on a comparative scale of 1 to 7 where they were to use "1" to mark the least stressful event and "7" to mark the most stressful event. They were also informed that they could use any rating number, including 1 and 7 as often as they liked, however, the one and the seven both must be used. The median results of this section of the questionnaire are listed in Table 5.3.

From the comparative ratings four stress categories were originally created: very high (events 7,10), high (events 3, 12, 13), neutral (events 4,5,6,8,9) and low stress (events 1,2 14,15). The toll events were later excluded from analysis because of an excess of motion artifacts due to the subject searching for change, obtaining a receipt and storing the receipt and also respiratory artifacts as the subject had to talk to the toll booth attendant to obtain the receipt. These artifacts were consistently present and were considered to obscure the desired readings of physiology for stress. The event garage exit was also reassessed due to a misunderstanding. This category was originally supposed to refer to the exit ramp out of East Garage which is a six story narrow descending spiral path, assumed to be high stress, however there was some confusion about this both because one of the experimenters did not understand this correctly and did not convey this to subjects and because occasionally the experiment was started on a lower floor and the full exit was not used. Subjects most often interpreted "garage exit" to refer to the short portion of flat driveway leading past the garage entrance booth.

	Median Scores				
Subject Pool	Rest	Highway	City	Tolls	Merges and Exits
Subject 1 (S1)	1	1	1	1	2
Subject 2 (S2)	2	3	3	3	4
Subject 3 (S3)	1	1	1.5	1	2
Single Runs (R)	1	2	2	2	3
All (S1+S2+S3+R)	1	2	2	2	2.5

Table 5.2: The median scores of perceived stress for different driving terrains on an absolute scale for various pools of drivers.

	Median Scores														
Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
S1	2.5	1.5	4	3	1.5	3	6.5	3.5	2	5	3	5	4.5	1.5	1.5
S2	1	1.5	3	3	3	3	6	3	3	6.5	4	3	3	2	1
S3	1	1	2	1	1	2	1.5	2	1.5	2	1.5	2	1.5	1	1
R	1	3	3.5	2	3	2	3.5	3	2.5	3	3	2.5	2	1	1
All	1	2	4	3	3	3	5	3	3	5	3	4	4	1	1

Table 5.3: The median scores of the forced 1-7 relative stress rating scale for fifteen driving events. The results for each of the returning subjects (S1-S3) and for the pool of single day drivers (R) are given. The collective median of all the correctly returned questionnaires (on which events were rated on the 1-7 scale) is shown in the last row (All).

Time	Eyes	Head	Body	Talk-D	Talk-E	Other	Comments
15:13:55	0	0	0	0	0	0	
15:13:56	0	0	0	0	0	0	
15:13:57	0	0	0	0	0	0	
15:13:58	0	0	0	0	0	0	
15:13:59	0	0	0	0	0	0	
15:14:00	0	0	0	1	0	0	
15:14:01	1	0	0	1	0	0	
15:14:02	1	0	0	0	0	0	
15:14:03	1	0	0	0	0	0	
15:14:04	0	0	0	0	0	0	

Table 5.4: An example from the exported Excel worksheet from the video coders for a portion of the rest period. Very little activity is typical of the rest period.

5.4.2 Video Coding

As a separate method of validation, the video tapes were viewed by independent researchers who were not involved with the signal processing or pattern recognition aspects of the project and coded on a second by second basis to mark observed stressors. The record was recorded on an Excel Spreadsheet in which rows represented each second of the ninety minute drive and the columns represented different events that were coded. In this rating scheme, the columns were marked with a "1" if the event occurred during that second and a "0" otherwise. Examples of two spreadsheet records are shown in Tables 5.4 and 5.5. The spreadsheet for scoring the resting segments did not include active driving events such as turning, stopping or road events such as bumps and tolls to facilitate coding. These events were assumed to be zero in the stationary vehicle.

The marked stressors were intended to represent deviations from a relaxing experience. The spreadsheet for the two rest periods included the following categories: Eyes, Head, Body, Talk-D, Talk-E and Other. Although there was some discretion on the part of the coder in rating these events the following heuristics were used: Eyes was marked as occurring whenever the subject opened their eyes, Head was marked whenever the subject moved their head, Body was marked whenever the sub-

105

Time	Bump	Stop	Turn	Talk-D	Talk-E	Eyes	Head	Toll	O	Cmts
15:49:11	0	0	1	0	0	1	0	0	0	
15:49:12	0	0	1	0	0	1	0	0	0	
15:49:13	0	0	1	0	0	1	0	0	0	
15:49:14	0	0	1	0	0	1	0	0	0	
15:49:15	0	0	0	0	0	1	0	0	0	
15:49:16	0	0	0	0	0	1	0	0	0	
15:49:17	1	0	0	0	0	1	0	0	0	
15:49:18	0	0	0	0	0	1	0	0	0	
15:49:19	0	0	0	0	0	1	1	0	0	
15:49:20	0	0	0	0	0	1	0	0	1	sit up
15:49:21	0	0	0	0	0	1	0	0	1	

Table 5.5: A sample from the exported Excel worksheet from the video coders for a portion of the city drive. More activity is typical of the city driving period.

ject moved any other part of their body, such as the arm or when they readjusted their position, Talk-D was marked whenever the driver spoke, Talk-E was marked whenever the experimenter spoke and Other was marked whenever the coder judged that some other stressor had occurred. For the driving segments, the additional categories Bump, Stop, Turn and Toll were added. For these categories: Bump was marked whenever the car went over a noticeable bump in the road, Stop was marked whenever the car braked to stop and Toll was marked whenever the driver conducted some extra action associated with passing through the toll such as looking for change. The category "Bump" was added in so that large bumps in the road that shook the car could be included as stressors. This was a category that was determined to be stressful by several drivers in pilot experiments. The category "Stop" replaced two original categories, "Stop" and "Start." The toll category was created to be used for all actions which were associated with approaching a toll, including getting money and putting the window down. This was created as a separate category because so many events happened preceding a tool that the raters often wanted to include more than one unclassified event in the "other" category. This category was created essentially to allow a value of up to "2" to be accumulated in the "other" category preceding a toll, which was the only place where it was deemed necessary. To test

106

reliability between coders, correlations were taken between two coders for the same drive. These results showed that the stress metrics were highly correlated (r=.99 for the rest segment, r=.84 for the city segment and r=.92 for the highway segment).

To create a signal with which to correlate the features on a second by second basis a time series was created by taking the sum of the nine events in each row of the spreadsheet ΣV. This time series was then smoothed by convolution with a 100 second hanning window to create the video stress metric $\Sigma \mathcal{V}$

$$\Sigma \mathcal{V} = \Sigma V * \mathcal{H}(100) \tag{5.1}$$

This smoothing allows stressors just before and after the second of interest to contribute to the stress metric.

5.4.3 Discussion

The video code ratings were compared to the ratings given by the driving tasks by experimental design and to the questionnaire ratings for twelve drives for which both the data segmentation marks and the video code data was available. After the toll and garage exit categories were eliminated, the driving task segmentation and the questionnaire were identical in two categories such that rest was equivalent to "low stress" and highway was equivalent to the "neutral stress." The city task category was included as a subset of the "high stress" category and the "very high stress" events were not considered in the first analysis. To see how the task categories and questionnaire categories compared to the scores from the coded video, a score was created for each segment by summing signal $\Sigma \mathcal{V}$ over the entire segment and dividing by the number of minutes in the segment. The score presented in Table 5.6 is the mean score for each category across for all days which were scored.

The results of this comparison show that the video code supports the assumptions of the task design and the questionnaire analysis, except for the "high stress" category. One difficulty may be that the very high stress category contains events that are very short in duration, such as merges, which may not span an entire minute. Therefore

107

Segment Label	Video Code Rating	Segment Label	Video Code Rating
Rest	13.6	Low Stress	13.6
Highway	61.4	Neutral Stress	61.4
City	87.7	High Stress	87.0
		Very High Stress	64.3

Table 5.6: A comparison of the video code metric to the task based stress categories and the questionnaire based categories. In this table the score presented for each category is the average value of the signal $\Sigma\mathcal{V}$ per minute across all segments for which video ratings and data segmentations were available. The relative ranking of the stress categories agrees with the task based assumptions and most of the questionnaire based assumptions, the exception being the rating for the very high stress category.

the minute segment which includes the merge may actually be mainly highway. A short event of very high stress, such as turning the car in a merge would be rated as less stressful by the video code metric than the same event taking a longer time, since the video is coded as events happening or not happening at any particular second. The intensity of the event is not measured.

5.5   New Features

Two new types of features were extracted from the physiological signals: features reflecting autonomic balance from the short term power spectrum of the heart rate and measures of the skin conductance orienting response. The power spectrum of the heart rate is used as a measure of heart rate variability (HRV). The energy in different portions of the spectrum reflects changes in heart rate mediated by the sympathetic and parasympathetic nervous systems. By taking ratios of the energy in the spectrum, a measure of sympathovagal balance can be calculated. A more detailed description of the influences on the heart rate spectrum is provided in Chapter 2. An algorithm for automatically detecting features of the skin conductance orienting response was also used in this analysis. This method detects the occurrence of the orienting response and measures their amplitude and rise time.

5.5.1 Heart Rate Variability

Features of short term power spectrum heart rate variability were derived from the EKG rhythm trace. A software package developed by George Moody was used (available to researchers http://ecg.mit.edu/) to detect the inter-beat R-R intervals using an autocorrelation technique. Heart rate was derived by using a time series representing the time between successive detected beats (inter-beat intervals). Outlying samples were eliminated from the time series. No method of estimation was used to replace the samples. Instead, the Lomb-Scargill technique, a method of spectral estimation which allows for missing data was used to estimate the spectral density. Unlike the FFT transform, the Lomb-Scargill method does not require the data to be resampled at an even sampling rate. From this spectrum features representing the high frequency, low frequency and middle frequency range were extracted. Ratios of these features were used in the feature vector for identifying periods of stress.

To derive the features used for heart rate variability, a time series of the heart rate was derived using the command "ihr" with standard defaults from the WAVE cardiac analysis software. The data was then partitioned into segments of 5 minute durations and a lomb spectrogram was performed on the data. In this spectrogram the total power below the Nyquist frequency (defined as the reciprocal of the mean interval between consecutive input samples) is normalized so that it is equal to the variance of the input. This normalization is used so that the Lomb periodograms can be compared directly to similarly-normalized FFT or AR spectra.

Three variables were derived from the Lomb-Spectrogram, the sum of the energy in the low, middle and high frequency regions:

$$LF = \sum_{0Hz}^{0.08Hz} lombPSD(ihr(ecg.dat)) \qquad (5.2)$$

$$MF = \sum_{0.08Hz}^{0.15Hz} lombPSD(ihr(ecg.dat)) \qquad (5.3)$$
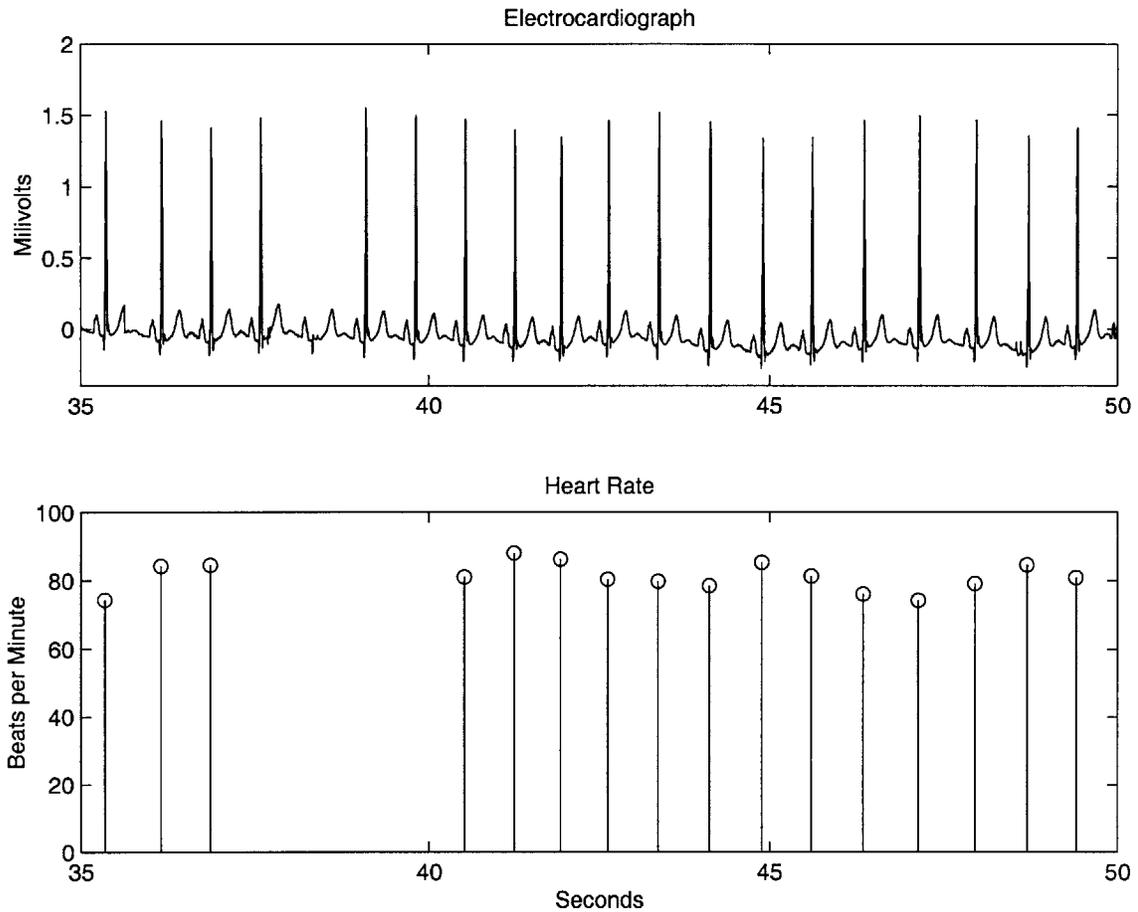
Figure 5-5: The digital electrocardiogram was not perfect. Absences of beats occasionally occurred. To help correct for this the time series record only records the inter-beat intervals of the beats it detects. Even beats bordering the missed beat are not recorded. The beats are not interpolated. The Lomb Scargill method was used to calculate the power spectrum of this series due to this kind of missing data.

$$HF = \sum_{0.15Hz}^{0.5Hz} lombPSD(ihr(ecg.dat)) \qquad (5.4)$$

These features are used in two ratios, $\frac{LF}{HF}$ and $\frac{LF+MF}{HF}$ to assess autonomic balance. Two different length of window were used to create the spectrum, a 100 second window and a 300 second window. In the two analysis of driving conditions the 300 second window was used and in the correlation analysis a continuous variable describing the autonomic balance is created by sliding both a 100 and 300 second window along the heart rate series and calculating both ratios at one second intervals as the window is advanced by one second.

5.5.2 Skin Conductance Orienting Response

The automatic startle detection algorithm developed and used in this thesis work implements the method of scoring the startle response described as method "B" in Chapter 2 (see Figure 2-3). The algorithm establishes a local baseline at the point where the derivative exceeds a certain threshold. The algorithm measures two features directly from the response, the magnitude (S_M) and the duration of the rise time (S_D). From this information two derived features were calculated, the frequency of occurrence (S_F) as the sum of the startle durations per minute and the area of the responses (S_A) as approximated by a triangular model of the rise time $(S_A = \frac{1}{2}S_M * S_D$. An illustration of these features as derived by the detection algorithm is shown in Figure 5-6.

The code for the detection algorithm is included in the appendix as sdetect.m. It is a Matlab function which takes as input a signal segment and the sampling frequency and gives as output the magnitudes, durations of the startle responses and the frequency of occurrence of the responses. The signal segment is first passed through a low pass filter with a cutoff of 4Hz to eliminate high frequency noise. This filter was implemented with the matlab command:

[b,a]=ellip(4,0.1,40,4*2/Fs);

yielding the filter coefficients of the system function:

111
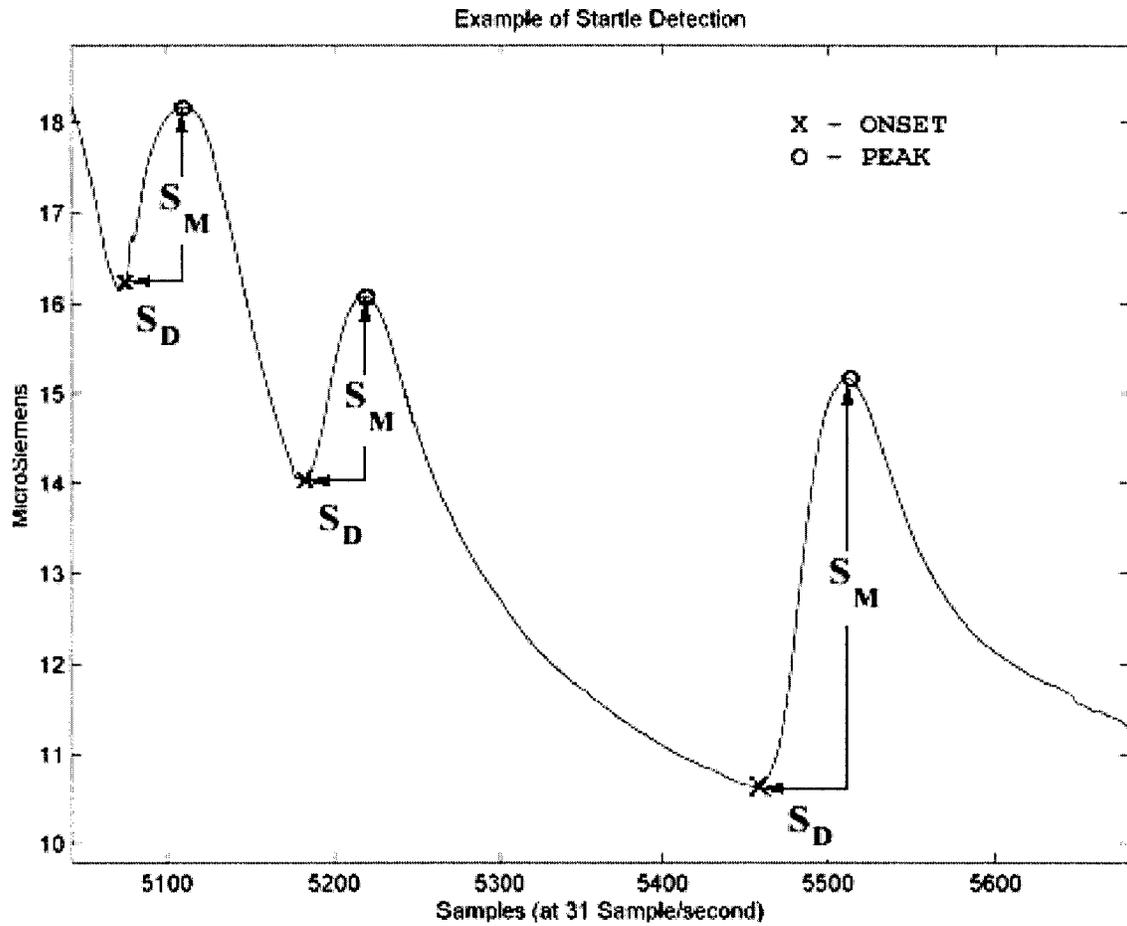
Figure 5-6: An example of the startle responses occurring in a one minute segment and the results of the algorithm showing onset "X" and peak "O" detection. The features S_M and S_D are calculated as shown.

$$H(z) = \frac{0.0375 + 0.0367z^{-1} + 0.0641z^{-2} + 0.0367z^{-3} + 0.0375z^{-4}}{-2.0792z^{-1} + 2.1493z^{-2} - 1.0979z^{-3} + 0.2426z^{-4}} \qquad (5.5)$$

The smoothing effect of this filter eliminated many false positive results. The derivative of the low pass filtered signal was then calculated by taking the first forward difference:

$$\delta_x[n] = x[n] - x[n-1]. \qquad (5.6)$$

The detection algorithm then identified all occurrences of when the first derivative exceeded a certain threshold. This threshold was empirically determined to 0.003 micro-Siemens per sample or 0.093 micro-Siemens per second. A filtering algorithm was then used to determine if this change was a new startle response by determining if it was more than one second away from other responses. This eliminated both high frequency noise artifacts and very small responses of less than one second duration. Once the response was detected, the zero-crossings of the derivative preceding and following the response were identified as the onset and peak of the response respectively. The amplitude and rise time of the response were then simply defined as:

$$S_M = t_{peak} - t_{onset} \qquad (5.7)$$

$$S_D = x_{peak} - x_{onset} \qquad (5.8)$$

$$S_A = \frac{1}{2} * S_M * S_D \qquad (5.9)$$

These features representing the skin conductance orienting response were used in both the task metric analysis and the questionnaire analysis.

113

5.6 Feature Summary

Different subsets of features were used in each of the three analyses. The definitions of the features not described in this section can be found in Chapter 3. In the task metric analysis, 22 features were used and derived for data segments 300 seconds in duration belonging to each of the three task categories: rest, highway driving and city driving. AB as $\frac{LF}{HF}$ for the entire segment, $\tilde{\mu}_{\mathcal{K}}$ and $\tilde{\sigma}^2_{\mathcal{K}}$ for the mean and variance of the heart rate derived from the EKG signal using the tach method of the WAVE software, S_M, S_D, S_A , S_F $\tilde{\mu}_{\mathcal{G}}$ and $\tilde{\sigma}^2_{\mathcal{G}}$ as derived from both the hand and the foot skin conductance signal, $\tilde{\mu}_{\mathcal{R}}$, $\tilde{\sigma}^2_{\mathcal{R}}$, p_1, p_2, p_3, p_4 from the respiration signal and $\tilde{\mu}_{\mathcal{E}}$ from the electromyogram on the trapezius muscle.

In the analysis using the questionnaire metric, twelve features were used. The AB feature was calculated using a 300 second window, while the signals: $\mu_{\mathcal{K}}$, S_M, S_D, S_A, S_F, $\tilde{\mu}_{\mathcal{G}}$ and $\tilde{\sigma}^2_{\mathcal{G}}$ from the hand skin conductance signal, $\tilde{\mu}_{\mathcal{R}}$, $\tilde{\sigma}^2_{R}$ and $\tilde{\mu}_{\mathcal{E}}$ and $\tilde{\sigma}^2_{\mathcal{E}}$ were calculated using one minute segments of data.

In the video code metric analysis, eleven features are used. These include six statistical features: $\tilde{\mu}_{\mathcal{E}}$, $\tilde{\sigma}^2_{\mathcal{E}}$, $\tilde{\mu}_{\mathcal{G}}$, $\tilde{\sigma}^2_{\mathcal{G}}$, $\mu_{\mathcal{R}}$ and $\sigma^2_{\mathcal{R}}$ calculated for each second of the data. Additionally, HR was calculated using the WAVE method ihr which excludes outliers. The correlation was then performed on a point by point basis between each detected R-wave and the video rating at that time. In addition, four functions representing autonomic balance at each second of the drive were calculated represented by AB1, AB2, AB3, AB4. These represent respectively, the $\frac{LF}{HF}$ ratio for the sliding 100 second window, the $\frac{LF}{HF}$ ratio for the sliding 300 second window, the $\frac{LF+MF}{HF}$ ratio for the sliding 100 second window, the $\frac{LF+MF}{HF}$ ratio for the sliding 300 second window.

5.7 Data Analysis

Three types of analysis were used to assess the performance of these physiological features in recognizing driver stress: an analysis of 22 features in discriminating stress according to the three major driving tasks (rest, city driving and highway
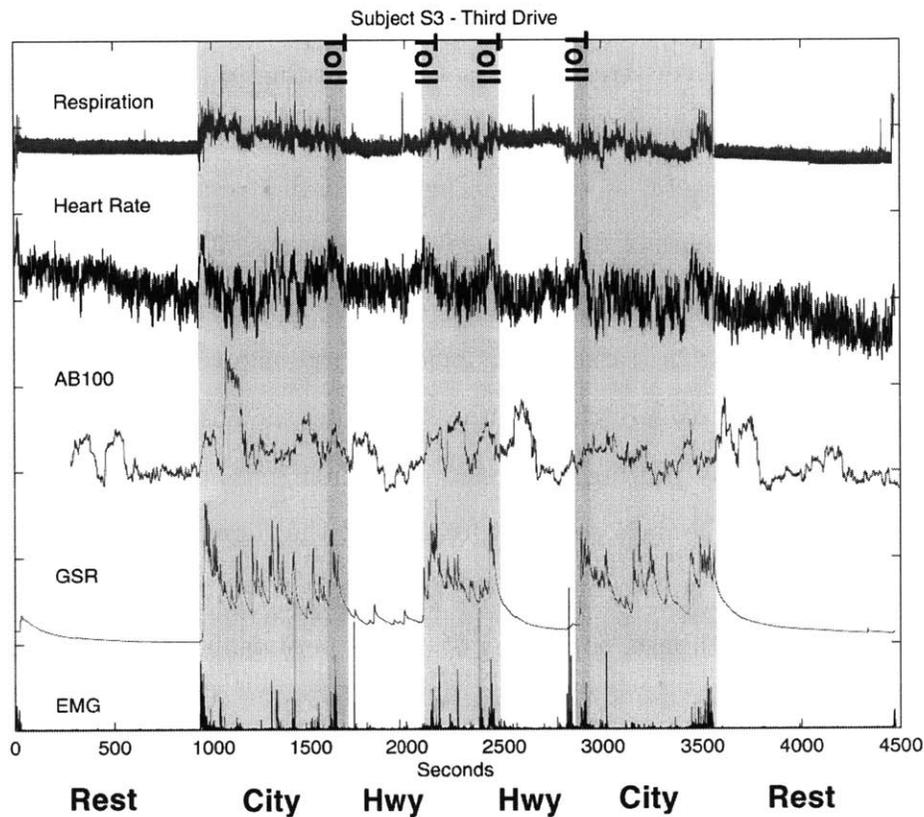
Figure 5-7: An example of the physiological data record collected from the drive. Signals are shown to emphasize how they change during the different parts of the drive. This record is from a day in which no unusual events occurred.

driving) using five minute windows; an analysis of 12 features at discriminating four stress levels as defined by the median questionnaire ratings using only one minute segments and an analysis of how eleven features of the physiological data correlated with the video code metric on a second by second basis. The goal of these studies was to prove the value of these features for discriminating stress levels and to push the limits of the assessment to shorter and shorter time segments to determine how quickly an automatic system could response to driver stress using these recognition algorithms. An example of the signals collected and analyzed on a typical day's drive along with markings showing the rest, highway driving and city driving periods is shown in Figure 5-7.

5.7.1 Task Metric Analysis

The first analysis of driver stress was assessed according to task. This metric is one of the most strongly supported metrics because the categories are observably different. These categories are useful at determining an overall stress level. This analysis also used the longest segments of data to derive the features, 300 seconds. The long time of these segments allowed the segments to be statistically similar over the window even if many variations in traffic conditions occurred on a second by second basis. To determine the start of each of these major segments, the observer made a mark in the physiological record by pulling a second respiration sensor. This created a spike in the data recording. The data segment for each of the three tasks began at this mark and included the next 300 seconds of data. A total of 112 five minute segments were extracted from the database, consisting of 36 rest periods, 38 city driving periods and 36 highway driving periods. As indicated by Appendix A, drives S1-1, S1-2, S1-3, S1-4, S1-6, S1-7, S2-1, S2-3, S2-7, S3-2, S3-3, S3-4, S3-5, S3-6, S3-7, R1-1, R2-1 and R3-1 were initially used in this analysis. On one of the days the rest period was lost and on another day the second rest period did not occur.

The 22 features calculated for each of the 300 second segments were initially grouped into six classes, representing the first and second three task types for each drive. Each of these classes was modeled be a gaussian distribution mean (m_k equal to the sample mean of the class, where k=1,2,3 for the rest, city and highway classes) the covariance K was the pooled covariance. A linear classifier was implemented by assigning each test sample $\hat{y}$ to the class k for which the function: g_k was maximum[The89] where:

$$g_k(\hat{y}) = 2\mathbf{m_k^T K^{-1}}\hat{y} - \mathbf{m_k^T K^{-1} m_k} + 2ln(Pr[w_k]); \qquad (5.10)$$

and the a priori probability of belonging to class k, $Pr[w_k] = \frac{1}{n_k}$ and n_k is the number of members in class k. For the results in Table 5.7 leave one out and test cross-validation was used where first m_k and K were calculated using all but the feature for one minute, then classifying the excluded feature according to the maximum g_k.

	Rest	City	Highway
Rest	36	0	0
City	0	37	1
Highway	0	2	36

Table 5.7: The confusion matrix for the linear discriminant after using the leave one out and test method.

The results were that these 22 physiological features can discriminate driver stress at 96% accuracy based on task. These tasks were confirmed to be of three distinct stress levels: low, medium and high by both the driver perceived stress questionnaires and the video code analysis. The confusion between the highway and city tasks may actually be due to certain drives through the city being of a lower stress level and certain drives on the highway being more stressful (such as when the driver tried to pass other cars on the highway).

5.7.2 Questionnaire Metric Analysis

A second analysis was performed using a metric of four categories defined by the perceived stress rated in the comparative analysis of the driver questionnaires. As indicated by the Table A.1 in Appendix A, thirteen data sets were used in this analysis. These were chosen at the time because the videos had been coded to indicate the start and stop times of the fifteen different driving events. The 12 features listed in the Feature Summary: μ_K, S_M, S_D, S_A, S_F, $\tilde{\mu}_G$ and $\tilde{\sigma}_G^2$ from the hand skin conductance signal, $\tilde{\mu}_R$, $\tilde{\sigma}_R^2$ and $\tilde{\mu}_\varepsilon$ and $\tilde{\sigma}_\varepsilon^2$ were calculated for 545 one minute segments to create a feature vector. This vector was assigned a label according to one of the four stress categories: low, neutral, high and very high stress ($k = 1, 2, 3, 4$) based on the median questionnaire rating of the minute from which it was extracted.

The features were evaluated and ranked individually, then combinations of features were evaluated and the best performances were reported. The individual features were tested on their ability to discriminate four stress classes using the linear classifier and leave one out and test cross validation described in the previous subsection on

117

Feature	Rank	Correct	Feature	Rank	Correct
$\tilde{\mu}_\mathcal{R}$	1	62.2%	$\tilde{\sigma}^2_\mathcal{E}$	6	53.5 %
$\tilde{\mu}_\mathcal{G}$	2	62.0%	ΣS_A	7	53.0 %
ΣS_D	3	58.5%	HR	8	52.6 %
$\tilde{\mu}_\mathcal{E}$	4	58.3%	AB	9	52.5 %
S_F	5	57.6%	$\tilde{\sigma}^2_\mathcal{R}$	10	50.2 %
ΣS_M	5	57.6%	$\tilde{\sigma}^2_\mathcal{R}$	11	48.3 %

Table 5.8: A ranking of each individual feature in the four stress level recognition task

Feature Set	Recognition Rate
$\tilde{\mu}_\mathcal{R}$ (best individual feature)	62.2 %
$\tilde{\mu}_\mathcal{R}$ $\tilde{\mu}_\mathcal{G}$ S_D $\tilde{\mu}_\mathcal{E}$ S_F S_M	78.5 %
AB $\tilde{\mu}_\mathcal{R}$ $\tilde{\mu}_\mathcal{G}$ S_D $\tilde{\mu}_\mathcal{E}$ S_F S_M	80.9 %
All 12	78.9 %

Table 5.9: Multiple feature combinations out-perform all single features. Combinations of six and seven features performs similarly to using all twelve features indicating that some features contribute no additionally useful information.

task metric analysis. The percent of segments correctly classified by this method is reported in Table 5.8.

This analysis shows that individual features perform better than random, but not very well, at discriminating the four stress categories. To test the performance of combinations of features the same linear classification using leave one out and test cross-validation was performed on various subsets of the data. The results presented in Table 5.9 show the results of the best single feature compared to combinations of the top six highest ranked individual features, the entire feature set and a set containing the top six individually ranked features with the addition of the AB variable, adding a representation of heart activity to the set.

These results show that sets of multiple features out-perform all single features. The best combination presented here achieved a recognition rate of 80.9%, a significant improvement over the 62% result achieved by the best single feature. This analysis also shows that some combinations of features perform similarly; however, in

118

Feature Set	Recognition Rate
AB HR S_D $\mu_{\mathcal{R}}$ $\mu_{\mathcal{E}}$ S_M $\sigma_{\mathcal{E}}^2$	88.6 %

Table 5.10: The results of the SFFS algorithm using the k-nearest neighbor classifier for the four stress classes.

some combinations, adding one features might contributes more noise while adding a different feature might improve performance. To determine which features might be best for stress recognition, a second analysis was performed using Jain and Zongker's sequential forward floating selection (SFFS) algorithm [JZ97] to evaluate set of features. This algorithm was run using a k-nearest neighbor classifier and leave one out cross validation. Table 5.10 shows the resulting optimal features (seven were chosen as an optimal set out of all possible sets) selected by the algorithm and the improved recognition rate of 88.6%.

The questionnaire based four stress level analysis shows that for even one minute segments of data which are subject to a greater variety of statistical variation within the segment due to second by second changes in the driving task demand, high recognition rates of perceived driver stress can be achieved using the features described here.

5.7.3 Video Code Metric Correlations

A final analysis was performed by comparing eleven features against the video code metric $\Sigma\mathcal{V}$ on a second by second basis. This analysis uses the most fine grain description of the ground truth collected during the drive to assess the stress level of each individual driver on each individual drive, whereas previous metrics assigned a stress category label to a drive segment based on either assumptions about the road conditions or on the median of statistics collected for all drives. For each drive, the video code metric $\Sigma\mathcal{V}$ was correlated with each of the eleven features: AB1, AB2, AB3, AB4, HR, $\mu_{\mathcal{E}}$, $\sigma_{\mathcal{E}}^2$, $\mu_{\mathcal{G}}$, $\sigma_{\mathcal{G}}^2$ and $\mu_{\mathcal{R}}$, $\sigma_{\mathcal{R}}^2$ and also with a randomly generated white noise signal w ($\mu = 0$, $\sigma^2 = 1$). The correlation with the random signal was to establish that the correlation coefficients found for the physiological features were

119

Day	AB1	AB2	AB3	AB4	HR	$\mu_\mathcal{E}$	$\sigma^2_\mathcal{E}$	$\mu_\mathcal{G}$	$\sigma^2_\mathcal{G}$	$\mu_\mathcal{R}$	$\sigma^2_\mathcal{R}$	w
S1-2	.53	.61	.53	.64	.34	.22	.01	.75	.09	-.53	.04	.01
S1-3	.45	.45	.44	.42	.35	.04	.01	.77	.08	-.49	.04	.00
S1-4	.45	.58	.47	.60	.53	.14	.06	.71	.18	-.33	.26	.01
S1-5	.41	.35	.22	.09	.46	.30	.08	.85	.22	-.22	.15	.01
S1-6	.62	.62	.59	.62	.31	.32	.09	.74	.00	-.56	.16	.01
S1-7	.46	.36	.41	.31	.52	.28	.04	.77	.23	-.23	.16	.01
S2-2	.49	.66	.55	.69	.49	.02	.03	.13	.00	-.24	.15	-.01
S2-4	.22	.29	.13	.17	.41	.27	.01	.59	.12	.12	.18	.00
S3-2	.74	.73	.75	.74	.44	.20	.06	.78	.20	.17	.25	-.01
S3-4	.46	.41	.48	.48	.38	.16	.06	.77	.15	.59	.19	.01
S3-5	.41	.51	.44	.50	.35	.09	.00	.81	.20	.21	.01	-.02
S3-6	.44	.53	.44	.51	.40	.20	.04	.73	.14	.67	.24	.03
S3-7	.35	.35	.39	.35	.29	.22	.08	.78	.16	.44	.12	-.01
R2-1	.41	.58	.39	.54	.30	.20	.06	.47	.06	.10	.03	.00
R3-1	.32	.42	.35	.41	.30	.16	.13	.45	.08	.03	.10	.01
R4-1	.49	.55	-.08	-.19	.76	.37	.09	-.07	.03	-.28	.22	-.03
Sum	7.3	8.0	6.5	6.9	6.6	3.2	0.8	10.0	2.0	-0.6	2.3	0.02

Table 5.11: Correlation coefficients between the stress metric created from the video and variables from the sensors. This coefficient shows how closely the sensor feature varies with the detected stressors on a second by second variable. As a control a set of random numbers was correlated with the video metric for each day, to assure that this correlation was close to zero.

significantly different, in most cases, from random. The results of these correlations are shown in Table 5.11 for each driver (S1 for Subject 1, S2 for Subject 2, S3 for Subject 3 and R for the remaining subjects who came for only one drive) and for each drive (S2-4 represents Subject 2's fourth drive). Some drives were excluded because either the video record had not been coded or the sensors had failed for a major portion of the drive.

The last line of Table 5.11 shows the sum of the correlation coefficients for each feature to provide a general metric for seeing how closely each of the features correlates with the video metric. These results show that the one second mean of the skin conductance signal and the autonomic balance variables and heart rate most closely track the stress metric provided by the video coders, suggesting their use in recognition tasks which required tracking stress on a second by second basis.

EMG Only	Rest	City	Highway
Rest	27	2	7
City	0	22	16
Hwy	3	7	28

Table 5.12: The confusion matrix for the task based linear discriminant using only the two features of the EMG, $\tilde{\mu}_{\mathcal{E}}$ and $\sigma_{\mathcal{E}}^2$

5.8 Discussion of Confounding Variables

Conducting experiments in the natural environment often entails accepting the occurrence of unexpected events and the presence of confounding variables such as motion, for often motion will co-occur with stressful events. This section presents a discussion of some possible confounding variables in this study: motion, subject variation and day variations.

5.8.1 Motion

The effect of motion can undoubtedly impact the autonomic variables used in this analysis. The design of this experiment has attempted to minimize the impact of motion in several ways. The driving task constrains the motion of the subject to a large extent such that the changes in variables found for standing, walking and jogging in Chapter 4 are not of concern. One of the goals of a previous study of driver physiology conducted by Helander [Hel78] (presented in the Chapter 2) was to determine to what extent driver muscle actions contributed to changes in skin conductivity and heart rate. His study particularly focused on the impact of braking activity. He found that heart rate was not well correlated with this muscle activity and the correlation coefficient between the skin conductance and brake pressure was over twice that of the correlation between skin conductance and muscle activity. From this he concluded that it "seems clear that EDR can be used to measure the mental difficulty of traffic events[Hel78]."

There is no doubt, however that muscle activity and stressful events co-occur. It

Optimal Selected Feature Set	SFFS kNN
AB, HR, ΣS_D, ΣS_M, $\mu_\mathcal{E}$, $\mu_\mathcal{R}$, $\sigma_\mathcal{R}^2$	88.6 %
AB, HR, ΣS_D, ΣS_M, S_F, $\mu_\mathcal{G}$, $\sigma_\mathcal{G}^2$	88.4 %

Table 5.13: Recognition rates achieved using Jain and Zongker's FS-SFFS algorithm with a k nearest neighbor classifier. No significant drop in performance occurs when $\mathcal{E}$ and $\mathcal{R}$ are eliminated from the initial pool.

may also be that a single EMG as used in this study or only two EMGs as used in Helander's study are sufficient to capture the impact of motion on the other physiological signals. For the sake of comparison, an analysis was performed using the mean EMG ($\tilde{\mu}_\mathcal{E}$) and EMG variance ($\tilde{\sigma}_E^2$) calculated in the experiment to evaluate stress based on driving task for discrimination. The same linear discriminant method used to generate the 96% correct result reported in Table 5.7 was used to create a discriminant using only EMG variables. The resulting confusion matrix is reported in Table 5.12. This Table shows that an accuracy of only 69% can be achieved using EMG alone, significantly less than the result achieved using all the features.

A second test was performed to see how well a classifier could do without features from sensors which might mainly represent motion. In this test, features derived from both the EMG sensor and the respiration sensor were eliminated from the initial pool of features and Jain and Zongker's FS-SFFS algorithm with a k nearest neighbor classifier was applied to the remaining features. From this pool the algorithm found that using the features AB, HR, ΣS_D, ΣS_M, S_F, $\mu_\mathcal{G}$, $\sigma_\mathcal{G}^2$, a recognition rate of 88.4% resulted. This rate is not significantly different from that achieved by the best combination of sensor using all features. The result of this experiment indicates both that motion is not necessarily a strong contributor to the effects measured as stress in this research and that good systems for recognizing driver stress can be developed using as few as two sensors, EKG and GSR.

Train				Test			
S2+S3+R	Rest	City	Hwy	S1	Rest	City	Hwy
Rest	23	0	0	Rest	10	1	0
City	0	23	1	City	0	12	0
Hwy	0	1	23	Hwy	1	1	12
S1+S3+R	Rest	City	Hwy	S2	Rest	City	Hwy
Rest	25	0	0	Rest	9	0	0
City	0	26	0	City	0	10	0
Hwy	0	1	25	Hwy	0	1	9
S1+S2+R	Rest	City	Hwy	S3	Rest	City	Hwy
Rest	23	1	0	Rest	10	0	0
City	0	25	1	City	1	7	2
Hwy	0	1	25	Hwy	0	0	10
S1+S2+S3	Rest	City	Hwy	R	Rest	City	Hwy
Rest	30	0	0	Rest	4	0	0
City	0	31	1	City	0	4	0
Hwy	0	0	32	Hwy	0	1	3

Table 5.14: This table shows the results of leaving approximately a quarter of the data set out of the training set by excluding sets that represent different portions of the data. In order from top to bottom, Subject 1 (S1) is first excluded from the data set, then Subject2 (S2), Subject 3 (S3) and finally the remaining single drive subjects (R). The confusion matrices are shown for both the training and test data. These results are comparable to those found using the leave one out cross validation, suggesting that the data is not overtrained.

5.8.2 Overtraining and Subject Variation

Additional tests were performed using the task metric data to determine if leaving one subject out of the training pool would strongly effect the results. By excluding one of the first three subjects who returned for multiple drives, approximately one quarter of the data was removed from the test pool. To make a more fair comparison the remaining subjects who drove only once were collected as a single subject. Although this last pool varied in its composition of individuals, they all shared the characteristic of being first time drivers. In this analysis, the set of features from the task based analysis was used.

Table 5.14 shows the results classification for the remaining quarter of the data

after excluding each subject and the amount of confusion inherent in the original training set. In order from top to bottom, subject drives S1 S2 S3 and R were excluded from the training data and then tested. The results achieved through this testing metric yield successful classifications of 92%, 97%, 90% 92%. These results are comparable to the 96% recognition rate using the entire data set. The comparable results from leaving one quarter of the data out of the training set and testing on the remaining quarter argues against the hypothesis that the original classifier was not overtrained.

5.8.3 Drive Variation

Conducting experiments in the natural environment allows for many unexpected incidents to occur. For example, Figure 5-8 shows a day in which two unexpected events occurred. During the first of the two highway driving segments, the subject took an unexpected exit and had to get back on to the highway. Additionally, during the second rest period the subject was agitated because they needed to use the restroom and had difficulty resting. The metric most susceptible to deviations from normal assumptions in these analyses is the questionnaire metric. The task metric has a relatively long time window such that a data period that is partially inconsistent with the assumption still had enough data in line with that assumption to make the correct classification. The correlation analysis with the video code does not rely on assumptions about what the stress level should be for a certain type of driving event, but rather assesses stress for each moment of each drive individually. Any single minute of the detour or the second rest period might have been misclassified by the questionnaire metric.

5.9 Summary

The driving task provides a minimally constrained environment in which naturally occurring stresses occur with reliable frequency and in which motion artifacts have a limited impact. This experiment was performed on the open roads and was limited
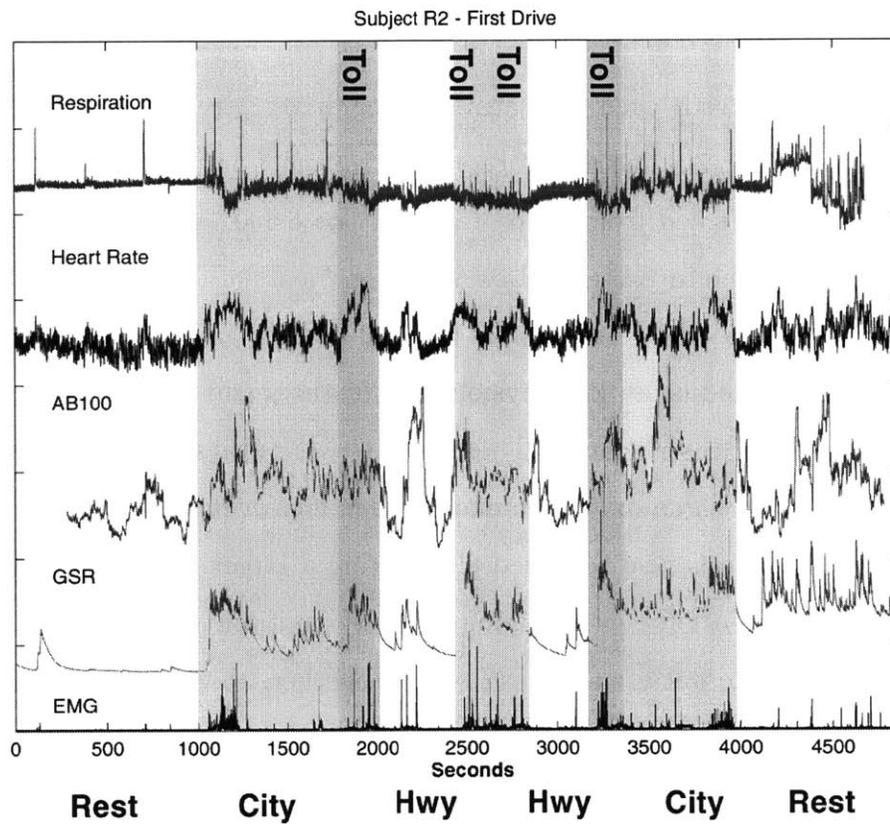
124

Figure 5-8: A second example of the physiological data record collected from one of the drives. This data record shows a record in which the subject was took an unexpected detour during the first highway session and in which the subject was unusually agitated during the second rest period.

only by a specific route the drivers were to follow, as would generally occurs on a daily commute. Analysis for both five minute segments representing the three driving conditions (rest, city and highway) and one minute segments (low, neutral, high or very high stress) show that driver stress can be recognized with 96% and 89% accuracy respectively. An evaluation of features of the four physiological signals on a second by second basis showed highly significant correlations, especially in the mean skin conductivity and autonomic balance variables (up to $r = .77$ for over 4000 samples).

Three methods were used for rating driver driver stress: experimental design, subject self-report and a metric of observed stressors from independent annotations of ground truth collected from video tapes of the drive. These metrics were compared and found to be in agreement on which segments of the drive represented low, medium, and high driver stress. Results from all three analyses show that these stress levels can be identified using the physiological features and pattern recognition algorithms presented in this research. These results were evaluated for their robustness to excluding signals which may have mainly been measuring motion and found to be robust. Evaluation was also repeated eliminating a larger portion of the data from the training set to test for overtraining. The susceptibility of the metrics to some unusual events in the data is also discussed. Together, these experiments show that systems for automatically detecting driver stress can be designed using the algorithms presented here.

Chapter 6

Conclusions

Novel systems and algorithms were designed and built to recognize affective patterns in physiological for the work of this thesis. Results show that in variously constrained situations affective states can be recognized by a computer using continuous physiological signals obtained through electromyogram, electrocardiogram, skin conductance, respiration and blood volume pressure sensors. These sensors were tested in the laboratory and embedded into wearable and automotive systems to measure affective signals in the natural ambulatory environment.

The most diverse set of emotions was studied in the most constrained laboratory setting. This experiment was designed to test for the presence of unique physiological patterns for the emotions set: no emotion, anger, hate, grief, love, romantic love, joy and reverence for a single subject over many days. New features of respiratory spectral analysis and the slope of the skin conductance were extracted from the physiological data. These features were combined with statistical features in an analysis of the data that yielded a recognition rate of 81% across all eight emotions. Higher recognition rates of up to 100% were achieved for subsets of these emotions clustered around similar emotion qualities.

To capture affective responses in the natural ambulatory environment, several prototype systems were developed using wearable computers. These systems included low-power, micro-processor based systems embedded into clothing and a versatile wearable computer system with desktop processing power and memory, augmented

127

with a digital camera and a wireless ethernet connection. A new interface was developed using a PalmPilot to facilitate data annotation in the field. These systems were found to provide a novel method monitoring physical activity, providing a time-stamped image diary synchronized with the physiological data record. Laboratory tests showed significant changes in signal features as subjects walked, jogged, stood up and sat down. Unfortunately, these effects were far greater than the modulations of physiology due to affect and it was determined that more sophisticated models of activity would be necessary to detect emotion in this fully unconstrained setting.

Finally, an automotive system was developed to detect driver stress. The design of the stress recognition experiment built upon the findings of both the eight emotion recognition experiment and wearables experiments. Automobile driving presented a natural situation in which ambulatory artifacts were limited and in which the negative arousal emotion of stress occurred with reliable frequency. The automotive system also allows the context of the drivers experience to be captured by multiple video cameras and a microphone and provided a strong metric of ground truth assessment. This analysis incorporates several previously developed features as well as two new types of features, an automatic assessment of the skin conductivity orienting response and a continuous metric of autonomic balance through short term power spectrum of the heart rate variability. Two recognition based performance tests were conducted, showing that given five minute segments of data three stress levels represented by driving conditions could be with recognized 96% accuracy, and that given only one minute segments of data four stress levels determined by driver self-report could be recognized with 89% accuracy. An evaluation of features of the four physiological signals on a second by second basis, highly significant correlations were found (up to $r = .77$ for over 4000 samples) with the skin conductivity signal and autonomic balance variables. Results from all three analyses show that driver stress levels as confirmed by three separate metrics can be identified using the physiological features and pattern recognition algorithms presented in this research. Additional analysis shows that highly significant correlations (up to $r = .77$ for over 4000 samples) exist between physiological features and a metric of observed stressors obtained from a

second by second annotation of video tape records of the drives. Results from all three analyses show that driver stress levels as confirmed by three separate metrics can be identified using the physiological features and pattern recognition algorithms presented in this research. The reliability of this algorithm might make it suitable for use in non-critical applications such as automatic management of information appliances and as an evaluation tool for road and vehicle designers.

Together, the three main experiments of this thesis show a range of success in recognizing affect from physiology. The recognition rates obtained thus far lend support to the hypothesis that many emotional differences can be automatically discriminated in patterns of physiological changes. The systems and algorithms developed in this thesis work open a new channel of computer human interaction which will become more viable as computers begin to be with us everywhere and as more automatic context sensing systems become integrated into these new computers. This work shows that affect detection through physiological signals is beginning to be possible. These findings contribute toward progress in developing machines which can respond intelligently to human affect.

Appendix A

Driving Database

Comments indicate problems in the data. In "Phys Data," hgsr and ekg represent that that these signals were lost for this run, rest means that no data was recorded during the rest period, marks means that there were no observer coded markers on the data record, lost indicates that a part of all the data was lost and sound indicated that some of the sound was lost in the recording.

In "Route" comments are made to indicate deviations from the expected route. Turn indicates that the subject was lost on the turnaround, crash indicated the day that the car was bumped into from behind, toll indicates that there was a problem at the toll, mass indicates the day Mass Ave. was closed, hwy indicates the day there was an accident on the highway that seriously altered traffic conditions and bath indicates the day the subject was particularly agitated and had to use the bathroom. OK indicates that the day was not marked as unusual by the observers, but further analysis of the video code might reveal some route discrepancies.

In the table:

```
Phys Data = Physiological Data Records
Route = Driving Route
3Task = This drive was used in the Task Metric Analysis
4Quest = This drive was used in the Questionnaire Metric Analysis
V-Code = The video is coded at the time of this thesis writing
V-ext = The video record exists for this day
```

Sub	Day	Phys Data	Route	Task	Quest	V-Code	V-Ext
S1-1	jul14	hgsr	OK	Yes	No	No	Yes -froze
S1-2	jul19	OK	turn	Yes	Yes	Yes	Yes
S1-3	jul21	rest	OK	Yes-r2	No	Yes	Yes
S1-4	jul23	OK	OK	Yes	No	Yes	Yes
S1-5	jul26	OK	OK	No	No	Yes	Yes
S1-6	jul28	OK	OK	Yes	Yes	Yes	Yes
S1-7	aug04	OK	OK	Yes	Yes	Yes	Yes
S2-1	jul15	OK	OK	Yes	Yes	No	Yes
S2-2	jul29	OK	crash	Yes-r2	No	Yes	Yes
S2-3	aug05b	OK	OK	Yes	Yes	No	Yes
S2-4	aug09a	OK	OK	No	Yes	Yes	Yes
S2-5	aug10a	OK	OK	No	Yes	No	Yes
S2-6	aug12	marks	OK	No	Yes	No	Yes
S2-7	aug13	OK	OK	Yes	No	No	Yes
S3-1	aug01	lost	toll	No	No	No	Yes
S3-2	aug02	OK	turn	Yes	No	Yes	Yes
S3-3	aug05a	sound	OK	-	No	No	Yes
S3-4	aug06	OK	OK	Yes	Yes	Yes	Yes
S3-5	aug07	OK	OK	Yes	No	Yes	Yes
S3-6	aug08	OK	mass	Yes	No	Yes	Yes
S3-7	aug09b	OK	hwy	Yes	Yes	Yes	Yes
R1-1	jul16	OK	OK	Yes	Yes	No	Yes
R2-1	jul22	OK	bath	Yes	Yes	Yes	Yes
R3-1	jul25	OK	OK	Yes	No	Yes	Yes
R4-1	jul27	hgsr	OK	No	No	Yes	Yes
R5-1	aug10b	ekg	OK	-	No	No	Yes
R6-1	aug19	sound	OK	No	Yes	No	Yes

Table A.1: Database for the Driver Stress Detection Experiment. In this Table, each driver is listed as a row, and each column shows the data that exists for that day.

Appendix B

Driving Experiment Subject Instructions

Sorry, you cannot participate if you do not have a valid and current driver's license

:(

Here is an outline of how things will go:

The experimenter will take you out to the MIT East Parking Garage behind building 68. You will get into the driver's seat of the 1998 Volvo V70 XC. The experimenter will be turning on the cameras and biofeedback recording equipment. Sensors will be placed on your skin that will monitor EKG (chest), EMG (shoulder), and GSR (fingers and foot) waveforms. You will also wear a belt around your chest to monitor respiration. Be aware that the tape which holds these sensors in place may hurt slightly when they are removed at the end of the experiment.

Your driving experience should include the following events:

1. A period of stationary monitoring (Stationary I)

2. Exiting the garage

3. A period of city driving down Massachusetts Avenue (City I)

4. A toll booth onto the Mass Pike (Toll I)

5. A period of highway driving out to Western Massachusetts (Highway I)

6. A toll booth onto Route 95 (Toll II)

7. A turn back onto the Mass Pike

8. A toll booth back onto the Mass Pike (Toll III)

9. A period of highway driving back to Boston (Highway II)

10. A three-lane merge onto the Allston/Cambridge exit

11. A toll booth onto the Allston/Cambridge exit (Toll IV)

12. A bridge crossing the Charles River back to Massachusetts Avenue

13. A period of city driving back to MIT (City II)

14. Parking in the East Garage

15. A stationary period at the end of the drive (Stationary II)

Please remember to observe all posted speed limits. When we get back, you will be asked to fill out a questionnaire asking about your driving experience today and your driving habits and history. The whole process should take about two and a half hours. Drive safely and remember to buckle up!

Appendix C

Driving Questionnaire

Subject Name: _____   Subject Number: _____

Session Number: _____   Date: _____   Experimenter: _____

This questionnaire is designed to help us assign labels to the data we have collected during your drive. You will be asked to rate the stress level of certain driving events and epochs and to rate certain episodes in relation to other episodes.

I. Background Questions:

1. How long have you had your driver's license?

2. How often do you usually drive?

   a. every day

   b. a few times a week

   c. a few times a month

   d. a few times a year

   e. never drive

3. Do you own a car or have a car you can use frequently?

   Yes   No   Other (explain)

If so, what kind of car is it?

4. Do you usually find driving a stressful experience?

    Yes   No

5. Are there any recent event in your life that you feel may have affected your driving experience today?

    Yes   No

6. If so, do you feel that you were more or less bothered by driving stressors than you would usually be?

    More   Less   Same

7. In general, do you feel you are more stressed than others, less stressed than others, or at about the same stress level as others?

    More   Less   Same

II. Today's Driving Experience

Overall, how would you describe the following events using the five point scale listed below:

```
1= no stress
2=a little stress
3=average stress
4= very stressful
5=extremely stressful
```

```
A.   Stationary periods:        _____

B.   City driving periods:      _____

C.   Highway driving periods:   _____

D.   Tolls:                     _____

E.   Merges and exits:          _____
```

III. General Questions

1. Was there a significant difference between your stress levels during the two stationary periods? Yes No

   If yes, which was more stressful? (I) before (II) after

   Why?

2. Was there a significant difference between your stress levels during the two city driving periods? Yes No

   If yes, which was more stressful? (I) (II)

   Why?

3. Was there a significant difference between your stress levels during the two highway driving periods? Yes No

   If yes, which was more stressful? (I) (II)

   Why?

4. Was there a significant difference between your stress levels during the toll encounters? Yes No

   If yes, which were more/less stressful? (I) (II) (III) (IV)

   Why?

On a scale of 1 to 7 with 1 representing the least stressful event and 7 representing the most stressful event, please rate the fifteen driving events. Feel free to use the same number more than once, even the one and the seven.

1. Period of stationary monitoring before driving (Stationary I)

2. Exiting the garage

3. Period of city driving down Mass. Ave. (City I)

4. Toll booth on the Mass. Pike (Toll I)

5. Period of highway driving out to Western Mass. (Highway I)

6. Toll booth onto Route 95 (Toll II)

7. Turn around back onto the Mass. Pike

8. Toll booth back onto the Mass Pike (Toll III)

9. Period of highway driving back to Boston (Highway II)

10. Three-lane merge onto the Allston/Cambridge exit

11. A toll booth onto the Allston/Cambridge exit (Toll IV)

12. Crossing back to Mass. Ave.

13. City driving back to MIT (City II)

14. Parking in the garage

15. Period of stationary monitoring after driving (Stationary II)

Please feel free to make any additional comments:

Thank you for participating in this experiment!

Appendix D

Matlab Code for Affect Analysis

D.1 Startle Detection

```
%The latest startle detection program
%Programmer: Jennifer Healey, Dec 13, 1999
%
% Usage: this is a script so you dont' have to pass a vector.
% You can easily make it into a function if you prefer.
%
% The signals is "s"  the sampling frequency is "Fs"
%
% the results are s_freq, s_mags and s_dur
%


% Edit all here or comment out for use in a giant script
Fs=31;


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%Filter out the high frequency noise%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
lgsr=length(s);

lgsr2=lgsr/2;

t=(1:lgsr)/Fs;


[b,a]=ellip(4,0.1,40,4*2/Fs);

[H,w]=freqz(b,a,lgsr);

sf=filter(b,a,s);


sf_prime=diff(sf);


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%Find Significant Startles%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


%There is ringing in the signal so the first 35 points are excluded


l35=length(sf_prime);


sf_prime35=sf_prime(35:l35);


%Set a threshhold to define significant startle


%thresh=0.005;

thresh=0.003;


vector=sf_prime35;


overthresh=find(vector>thresh);


%overthresh is the values at which the segment is over the threshold
```

```
overthresh35=overthresh+35;


%the true values of the segment


gaps=diff(overthresh35);


big_gaps=find(gaps>31);


%big_gaps returns the indices of gaps that exceed 31
%
% eg - big_gaps=[60 92 132 168....]
%
%  gaps(60)=245; gaps(92)=205 ...
%
% overthresh35(58:62)=  [346    347    348    593    594]
%
% so overthresh(61) is where the startle starts (ish)
%
% is overthresh (60) where the startle ends?



%check the results


iend=[];
ibegin=[];


for i=1:length(big_gaps)
   iend=[iend overthresh35(big_gaps(i))];
```

```matlab
    ibegin=[ibegin overthresh35(big_gaps(i)+1)];
end;




%%%%%%%%%%%%
%%Fine Tuning%%
%%%%%%%%%%%%%


% The idea being this is to find the zero crossing closet to where it goes
% over threshold

% find all zero crossings



overzero=find(sf_prime>0);
zerogaps=diff(overzero);
z_gaps=find(zerogaps>1);


iup=[];
idown=[];


for i=1:length(z_gaps)
  idown=[idown overzero(z_gaps(i))];
  iup=[iup overzero(z_gaps(i)+1)];
end;


% find up crossing closest to ibegin


new_begin=[];
for i=1:length(ibegin)
```

```
      temp=find(iup<ibegin(i));
      choice=temp(length(temp));
      new_begin(i)=iup(choice);
end;


% to find the end of the startle, find the maximum between startle
% beginnings


new_end=[];


for i=1:(length(new_begin)-1)
   startit=new_begin(i);
   endit=new_begin(i+1);
   [val, loc]=max(s(startit:endit));
   new_end(i)=startit+loc;
end;


if (length(new_begin)>0)
   last_begin=new_begin(length(new_begin));
   [lastval,lastloc]=max(s(last_begin:length(s)-1));
   new_end(length(new_begin))=new_begin(length(new_begin))+lastloc;
end;



s_mag=[];   %initialize a vector of startle magnitudes
s_dur=[];    %initialize a vector of startle durations


for i=1:length(new_end)
   s_dur(i)=new_end(i)-new_begin(i);
   s_mag(i)=s(new_end(i))-s(new_begin(i));
```

```
end;


s_freq=length(ibegin);
```

D.2 Linear Discriminant

```
%the_matrix is a matrix that holds all the features for all the classes
% it should be formatted as follows:


%e.g
%          f1 f2 f3 f4 f5
% c1       4  5  5   5  5
%          6  7  7   4  3
%          1  3  3   2  6
% c2       1  5  6   6  6
%          2  3  3   5  6
% c3       3  3  3   3  3
%          3  2  2   2  2


%the variable "class_index_vector" keeps track of where the classes
%are in "the_matrix"


%the_matrix replaces old class variables
%class1=the_matrix(1:class_index_vector(1),:);
%class2=the_matrix(class_index_vector(1)+1:class_index_vector(2),:);


%e.g class_index_vector=[3 5 7]



the_matrix=[];                    %initialize the_matrix
class_index_vector=[];            %initialize class_index_vector
g=[];                             %initialize the liklihood ratio vector
```

```
%%%%%ALL EDITING SHOULD BE DONE IN THESE TWO LINES%%%%%%%%%%%%%

the_matrix=[class_r1_all; class_r2_all;

class_c1_all; class_c2_all; class_h1_all; class_h2_all];

%class_index_vector=[19 36 55 74 93 112];

class_index_vector=[36 74 112];


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


c_i_v=class_index_vector;


[number_of_values number_of_features]=size(the_matrix);

number_of_classes=length(class_index_vector);


%replaces old n1 n2 n3....

% now nv=[n1 n2 n3 ...]

nv(1)=c_i_v(1);

for i=2:number_of_classes

  nv(i)=c_i_v(i)-c_i_v(i-1);

end;


lin_conf=zeros(number_of_classes,number_of_classes);

%Initializes the confusion matrix


n =sum(nv);


% mean vector is now "mv"

% replaces old m1 m2 m3....

% now mv=[m1 m2 m3 ...]


mv=[];
```

```
mv(1,:)=mean(the_matrix(1:c_i_v(1),:));

for i=2:number_of_classes

  mv(i,:)=mean(the_matrix(c_i_v(i-1)+1:c_i_v(i),:));

end;


%pooled mean

%replaces old

%m=(n1/n*m1 + n2/n*m2 + n3/n*m3 + n4/n*m4 +n5/n*m5 +n6/n*m6);


m=0;    %initialize the pooled mean

for i=1:number_of_classes

  m=nv(i)/(n*mv(i));

end;


%The individual covariance matricies in each class

%replaces old K1 = KV(1:n_o_f,:) K2,K3


KV=[];


KV(1:number_of_features,:)=cov(the_matrix(1:c_i_v(1),:));

n_o_f=number_of_features;

for i=2:number_of_classes

  i1=((i-1)*n_o_f)+1;

  i2=i1+n_o_f-1;

  KV(i1:i2,:)=cov(the_matrix(c_i_v(i-1)+1:c_i_v(i),:));

end;


%prior probability

Pw=[];
```

```
Pw(1)=nv(1)/n;

for i=2:number_of_classes

  Pw(i)=nv(i)/n;

end;


%average covariance matrix across all different classes

%pooled covariance

%replaces old

%K= n1/n*K1 + n2/n*K2 + n3/n*K3 + n4/n*K4 + n5/n*K5 + n6/n*K6;



K=[];  %initiatize K

K=nv(1)/n*KV(1:n_o_f,:);  %



for i=2:number_of_classes

  i1=((i-1)*n_o_f)+1;

  i2=i1+n_o_f-1;

  K=K+ (nv(i)/n*KV(i1:i2,:))

end;


%optimized by leaving out log det K


for i=1:number_of_classes

  C(i) = -mv(i,:)*inv(K)*mv(i,:)' + 2*log(Pw(i));

end;


%initialize the number of failures

fail=0;
```

```matlab
%for c=1 to the number of elements in class 1


findex=[zeros(number_of_classes, max(nv))];


lfv=[];   %initialize linear fail vector


j=1;
for c=1:nv(1)
  y=[the_matrix(c,:)]';
  %liklihood ratios
  for i=1:number_of_classes
    g(i)=2*mv(i,:)*inv(K)*y + C(i);
  end;
  for i=1:number_of_classes
    if(max(g)==g(i))
      lin_conf(j,i)=lin_conf(j,i) + 1;
      %This builds the confusion matrix
    end;
  end;
  %this finds the max

  if(max(g)~=g(1))
    fail=fail+1;
    findex(j,c)=1;
  end;
end;


lfv(1)=fail;


for j=2:number_of_classes
```

```
fail=0;
for c=c_i_v(j-1)+1:c_i_v(j)
    y=[the_matrix(c,:)]';


    for i=1:number_of_classes
        g(i)=2*mv(i,:)*inv(K)*y + C(i);
    end;
    for i=1:number_of_classes
        if(max(g)==g(i))
            lin_conf(j,i)=lin_conf(j,i) + 1;
            %This builds the confusion matrix
        end;
    end;


    %this finds the max
    if(max(g)~=g(j))
        fail=fail+1;
        findex(j,c)=1;
    end;
end;
lfv(j)=fail;
end;
```

Bibliography

[AGU+81]   S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. C. Barger, and R. J. Cohen. Power spectrum analysis of heart rate fluctuation: A quantitative probe or beat-to-beat cardiovascular control. *Science*, pages 213–220, 1981.

[AMM87]   Jans Aasman, Gijsbertus Mulder, and Lambertus Mulder. Operator effort and the measurement of heart rate variability. *Human Factors*, 29(2):161–170, 1987.

[Ax53]   Albert F. Ax. The phsyiological differentiation between fear and anger in humans. *Psychosomatic Medicine*, 55(5):433–442, 1953.

[Bea97]   Gary G. Berntson and J. Thomas Bigger. Jr. et al. Heart rate variability: Origins, methods and interpretive caveats. *Psychophysiology*, 34:623–647, 1997.

[Bou92]   Wolfram Boucsein. *Electrodermal Activity*. Plenum Press, New York, NY, 1992.

[BP96]   Lothar Buse and Kurt Pawlik. Ambulatory behavioral assessment and in-field performance testing. In Jochen Fahrenberg and Micheal Myrtek, editors, *Ambulatory Assessment*, Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies, chapter 3, pages 29–50. Hogrefe and Huber, Seattle, first edition, 1996.

[Can27]   Walter B. Cannon. The James-Lange theory of emotions: A critical examination and an alternative theory. *American Journal of Psychology*, 39:106–124, 1927.

[Cly77]   Dr. M. Clynes. *Sentics: The Touch of the Emotions*. Anchor Press/Doubleday, 1977.

[Cly80]   Dr. M. Clynes. *The Sentic Cycle Experiments*. A binder I got with the Sentograph, 1980.

[CT90]   John T. Cacioppo and Louis G. Tassinary. Inferring psychological significance from physiological signals. *American Psychologist*, 45(1):16–28, Jan. 1990.

[Dam94]   A. R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Gosset/Putnam Press, New York, NY, 1994.

[Dav94]   Richard J. Davidson. Temperament, affective style, and frontal lobe asymmetry. In Geraldine Dawson and Kurt W. Fischer, editors, *Human Behavior and the Developing Brain*, pages 518–536. Guilford Press, New York, 1994.

[DEM88]   M. Davis, E. R. Eshelman, and M. McKay. *The Relaxation & Stress Reduction Workbook*, chapter Biofeedback, pages 203–210. New Harbinger Publications, Inc., third edition, 1988.

[DH73]   R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience, 1973.

[EF78]   P Ekman and W Friesen. *Facial Action Coding System: A technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[EG88]   I. J. Kopin G. Eisenhofer and D. Goldstien. Sympathoadrenal medullary system and stress. In *Mechanisms of Physical and Emotional Stress*. Plenum Press, 1988.

[Ekm90]   P. Ekman. Commentaries: Duchenne and facial expression of emotion. In R. Andrew Cuthbertson, editor, *The Mechanism of Human Facial Expression*, pages 270–284. Cambridge University Press, 1990.

[Ekm92a]   P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3/4):169–200, 1992.

[Ekm92b]   Paul Ekman. Are there basic emotions? *Psychological Review*, 99(3):550–553, 1992.

[Ekm93]   Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, 1993.

[ELF83]   Paul Ekman, Robert W. Levenson, and Wallace V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221:1208–1210, Sep. 1983.

[Fri86]   Nico H. Frijda. *The Emotions*, chapter Physiology of Emotion, pages 124–175. Studies in Emotion and Social Interaction. Cambridge University Press, Cambridge, 1986.

[Gou98]   Grant Gould. Affective computing lab notebook. Master's thesis, Massachusetts Institute of Technology, 1998.

[Gou99]   Grant Gould. Vbus: Vehcile instrumentation for affective computing. Master's thesis, Massachusetts Institute of Technology, 1999.

[HB96]   S. G. Hofmann and D. H. Barlow. Ambulatory psychophysiological monitoring: A potentially useful tool when treating panic relapse. *Cognitive and Behavioral Practice*, 3:53–61, 1996.

[Hel78]   M. Helander. Applicability of drivers' electrodermal response to the design of the traffic environment. *Journal of Applied Psychology*, 63(4):481–488, 1978.

[ITea95]   Hiroshi Itoh, Kazuo Takeda, and Kazue Nakamura et al. Young border-
            line hypertensives are hyperreactive to mental arithmetic stress: Spec-
            tral analysis of r-r intervals. *Journal of the Autonomic Nervous System*,
            54:155–162, 1995.

[Jam92]   William James. *William James: Writings 1878-1899*, chapter on Emo-
            tion, pages 350–365. The Library of America, 1992. Originally published
            in 1890.

[JZ97]   A. Jain and D. Zongker. Feature-selection: Evaluation, application, and
            small sample performance. *PAMI*, 19(2):153–158, February 1997.

[Kah73]   Daniel Kahneman. Arousal and attention. In *Attention and Effort*,
            pages 28–49. Prentice-Hall, Englewood Cliffs, N.J., 1973.

[KF98]   Markad V. Kamath and Ernest L. Fallen. Heart rate variability: Indica-
            tor of user state as an aid to human-computer interaction. *Proceedings
            of CHI98*, 34:480–487, 1998.

[Lan95]   P. J. Lang. The emotion probe: Studies of motivation and attention.
            American Psychologist, 50(5):372–385, 1995.

[LBC]   P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective
            picture system (IAPS): Technical manual and affective ratings. NIMH
            Center for the Study of Emotion and Attention, University of Florida.

[Lev90]   Charles F. Leventhal. *Introduction to Physiological Psychology*. Prentice
            Hall, Englewood Cliffs, NJ, 1990.

[Lev92]   R. W. Levenson. Autonomic nervous system differences among emo-
            tions. *American Psychological Society*, 3(1):23–27, Jan. 1992.

[LGea93]   P. J. Lang, M. K. Greenwald, and M. M. Bradley et al. Looking at
            pictures: Affective, facial, viseral and behavioral reactions. *Psychophys-
            iology*, 30:261–273, 1993.

[LLH+95]   T. Leino, J. Leppaluoto, P. Huttunen, A. Ruokonen, and P. Kuronen. Neuroendicrine responses to real and simulated ba hawk mk 51 flight. *Aviation, Space, and Environmental Medicine*, pages 108–113, February 1995.

[LRLM66]   D. T. Lyyken, R. Rose, B. Luther, and M. Maley. Correcting psychophysiological measures for individual differences in range. *Psychophysiological Bulletin*, 66:481–484, 1966.

[LV71]   David T. Lyyken and Peter H. Venables. Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, 8(5):656–672, 1971.

[MAea95]   Rollin McCraty, Mike Atkinsom, and William Tiller et al. The effects of emotions on short-term power spectrum analysis of heart rate variability. *American Journal of Cardiology*, 76:1089–1093, 1995.

[Mat99]   The Mathworks, Inc., 3 Apple Hill Drive,Natick, MA 01760-2098. *Matlab Version 5.3.0.10183 (R11) On-line User's Manual*, 1999.

[MP98]   Teresa Marrin and Rosalind Picard. A methodology for mapping gestures to music using physiological signals. *Submitted to: International Computer Music Conference*, 1998.

[OCC88]   Andrew Ortony, Gerald Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[Rus80]   James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[Sch64]   Stanley Schachter. The interaction of cognitive and physiological determinants of emotional state. In Leonard Berkowitz, editor, *Advances in Experimental Psychology*, volume 1, pages 49–80. Academic Press, New York, 1964.

[Sel56]   Hans Selye. *The Stress of Life*, chapter 1-7. McGraw-Hill, 1956.

[Sel80]   Hans Selye. *Selye's Guide to Stress Research*. Van Nostrand Reinhold Company, 1980.

[SF90]   Michael E. Dawson Anne M. Schell and Diane L. Filion. The electrodermal system. In Cacioppo and Tassinary, editors, *Principles of Psychophysiology: Physical, social and inferential elements*, chapter 10, pages 295-324. Cambridge University Press, Cambridge, first edition, 1990.

[SMR+97]   T. Starner, S. Mann, B. Rhodes, J. Levine, J. Healey, D. Kirsch, R. Picard, and A. Pentland. Augmented reality through wearable computing. *Presence*, 6(4):386-398, Winter 1997.

[SSea93]   J. Paul Spiers, Bernard Silke, and Ultan McDermott et al. Time and frequency domain assessment of heart rate variability: A theoretical and clinical appreciation. *Clinical and Autonomic Research*, 3:145-158, 1993.

[Sta95]   T. E. Starner. Wearable computing. Perceptual Computing Group, Media Lab 318, MIT, Cambridge, MA, 1995.

[The89]   C. W. Therrien. *Decision Estimation and Classification*. John Wiley and Sons, Inc., New York, 1989.

[The92]   C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1992.

[Tho94]   Thought Technology Ltd., 2180 Belgrave Ave. Montreal Quebec Canada H4A 2L8. *ProComp Software Version 1.41 User's Manual*, 1994.

[VP99]   Elias Vyzas and Rosalind W. Picard. Offline and online recognition of emotion expression from physiological data. In *Workshop on Emotion-Based Archetectures, at the Third International Conference on Autonomous Agents*, Seattle, WA, May 1999.

[vRAKea93] Conny van Ravenswaaij-Arts, Louis A. A. Kollee, and et al. Heart rate variability. *Annals of Internal Medicine*, 118:436–447, 1993.

[Vyz99] Elias Vyzas. *Recognition of Emotional and Cognitive States Using Physiological Data*. PhD thesis, MIT, Cambridge, MA, June 1999.

[WPK84] W. M. Winton, L. Putnam, and R. Krauss. Facial and autonomic manifestations of the dimensional structure of emotion. *Journal of Experimental Social Psychology*, 20:195–216, 1984.

[YLL+97] Hannamaija Ylonen, Heikki Lyytinen, Tuomo Leino, Juhani Leppaluoto, and Pentti Kuronen. Heart rate responses to real and simulated ba hawk mk 51 flight. *Aviation, Space, and Environmental Medicine*, 68(7):601–605, July 1997.

[Zim96] Thomas Zimmerman. Personal area networks (PAN): Near-field intrabody communication. *IBM Systems Journal*, 35:609–618, 1996.