Recognition of Emotional and Cognitive States Using Physiological Data

by

Elias Vyzas

Bachelor of Engineering in Mechanical Engineering,
Imperial College, London (1994),
Master of Science in Mechanical Engineering,
Massachusetts Institute of Technology (1997),
Master of Science in Electrical Engineering and Computer Science,
Massachusetts Institute of Technology (1997)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Mechanical Engineer

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1999

Author .....
Department of Mechanical Engineering
May 24, 1999

Certified by .....
Rosalind W. Picard
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Certified by .....
Thomas B. Sheridan
Professor of Engineering & Applied Psychology
Thesis Reader

Accepted by .....
Ain A. Sonin
Graduate Officer
Department of Mechanical Engineering

Recognition of Emotional and Cognitive States Using Physiological Data

by

Elias Vyzas

Submitted to the Department of Mechanical Engineering
on May 24, 1999, in partial fulfillment of the
requirements for the degree of
Mechanical Engineer

Abstract

This thesis presents the application of several pattern recognition techniques on physiological data as a means to provide useful information about human emotional or cognitive states. As these states may be correlated with the well-being and performance of subjects, knowledge of these states could improve the human-computer interaction, increase productivity, and reduce accidents.

We first focus on a method for recognizing the emotional state of a person who is deliberately expressing one of eight emotions. Four physiological signals were measured and six features of each of these signals were extracted. We investigated three methods for the recognition: (1) Sequential floating forward search (SFFS) feature selection with K-nearest neighbors classification, (2) Fisher Projection (FP) on structured subsets of features with MAP classification, and (3) A hybrid SFFS-FP method. Each method was evaluated on the full set of eight emotions as well as on several subsets. The day-to-day variations within the same class often exceeded between-class variations on the same day. We present a way to take account of the day information, resulting in an improvement to the Fisher-based methods. The SFFS attained a rate of 88% for a trio of emotions, while the Fisher Projection attained the best performance on the full set of emotions, 81.25%. We extend the previous study by building an online classifier so that it can be used for real-time applications. The performance is comparable to that of the offline version. These findings demonstrate that there is significant information in physiological signals for classifying the affective state of a person who is deliberately expressing a small set of emotions.

We then look into cognitive load under different driving conditions. Subjects are asked to drive in a driving simulator around several curves. Messages appear on the screen prompting the driver to either brake immediately to a standstill or to continue driving. In parts of the experiment the driver is asked to perform a simple mathematical task on the phone. Several measures of the subjects' behavior are recorded, including driving parameters such as lane deviation, distance and time to lane crossing, steering entropy, and braking delay, mistakes in addition, and physiological data (EMG, BVP, GSR, HR, Respiration). Results show that although the majority of braking delays (irrespective of the phone task) lay between -0.5 and +0.5 seconds of the average no-phone delay, there were a few cases in which subjects pressed the brakes significantly later (0.5-2.5 seconds after the average no-phone delay). Out of

315 messages prompting subjects to brake while they were not engaged on a phone task, only twice did their breaking delay exceed the average; out of 642 messages prompting subjects to brake while they were engaged on a phone task, the delay exceeded the average 41 times. The effect of the mathematical task can also be seen in a 10% higher mean reaction time and a four times larger variance when subjects were on the phone compared to when they were not on the phone. Furthermore, people were on the phone in 9 out of the 10 cases that subjects mistakenly pressed the brake pedal while the message prompted them to continue driving, as well as in 6 out of the 7 cases that subjects did not show any reaction while the message prompted them to brake. We separated the responses into 2 classes, a normal and a slow one. Using the physiological data and similar pattern recognition techniques as mentioned above we predicted the class of the next delay with 65% success for an individual subject. These results indicate that the existence of specific secondary tasks while driving may adversely affect the reaction time of the driver, while use of physiological data may help in predicting such potentially dangerous situations.

Acknowledgments

I would like to thank my supervisor, Professor Rosalind Picard for all her help, support and inspiration, and for being such a nice person to work with, as well as the reader of my thesis, Professor Thomas Sheridan. I would also want to thank Professor Elias Gyftopoulos, for his advice and his belief in me.

I would also like to thank Jennifer Healey for overseeing the data collection and her help with the physiological sensors; Anil Jain and Doug Zongker for providing us with the SFFS code, and Tom Minka for helping with its use; Teresa Marrin for helping with the EMG sampling rate analysis; Paris Smaragdis for helping with the Speech Generation software; Tom Minka for helpful discussions and suggestions about pattern recognition algorithms.

Drs. Andrew Liu and Erwin Boer have been very helpful during the experiments in the driving simulator, sharing their work and insights with me. I would also like to thank Cambridge Basic Research for allowing me to use their facilities.

Above all, I feel the need to thank my parents and several friends (they know who they are), and hope that wherever these friends end up, we will still be as close.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The application of several pattern recognition techniques on physiological data can provide useful information about human emotional or cognitive states. Physiological data can help in the recognition of the level of cognitive load, of frustration involved in the performance of a task or of the presence of stress. These states can be highly correlated with performance. If we could know when users are in an unfomfortable or even dangerous emotional/cognitive situation, we could inform them, or intervene and postpone or take over some low-priority tasks. This could help improve the human-computer interaction and reduce accidents, thus making workplaces friendlier and safer.

1.1 Use of physiological data for recognition of emotional states

Part of this thesis addresses emotion recognition, specifically the recognition by computer of affective information expressed by people, through use of physiological and other data. This is part of a larger effort in "affective computing," computing that "relates to, arises from, or deliberately influences emotions" [14]. Affective computing has numerous applications and motivations, one of which is giving computers the skills involved in so-called "emotional intelligence," such as the ability to recognize a

person's emotions. Such skills have been argued to be more important in general than mathematical and verbal abilities in determining a person's success in life [7]. Recognition of emotional information is a key step toward giving computers the ability to interact more naturally and intelligently with people.

The research in this section focuses on recognition of emotional states during deliberate emotional expression by an actress [19]. The process included the following eight states: Neutral (no emotion), Anger, Hate, Grief, Platonic Love, Romantic Love, Joy, and Reverence. Four physiological signals of the actress were recorded during the deliberate emotional expression. The signals measured were electromyogram (EMG) from the jaw representing muscular tension or jaw clenching, blood volume pressure (BVP) and skin conductivity (GSR) from the fingers, and respiration from chest expansion. Data was gathered for approximately 3 minutes for each of the eight emotional states, and the process was repeated for several sessions, over the course of weeks.

Very little work has been done on pattern recognition of emotion from physiological signals, and there is controversy among emotion theorists whether or not emotions do occur with unique patterns of physiological signals. Some psychologists have argued that emotions might be recognizable from physiological signals given suitable pattern recognition techniques [2], but nobody has yet to demonstrate which physiological signals, or which features of those signals, or which methods of classification, give reliable indications of an underlying emotion, if any. The thesis suggests signals, features, and pattern recognition techniques for offline recognition of all 8 emotions examined and presents results suggesting that emotions can be recognized from physiological signals at significantly higher than chance probabilities.

Emotion recognition can be very useful if it occurs in real time. That is, we would like the computer to be able to sense the emotional state of the user the moment he actually is in this state (online recognition), rather than analyzing the data later, when the user is already in another state (offline recognition). This could be considered in combination with the model of an underlying mood, which may change over longer periods of time. In that respect, the classification rate of a time window given a

previous time window may yield useful information. The question is how frequently should the estimates of the baseline be updated to accommodate for the changes in the underlying mood. In addition, it appears that although the underlying mood changes the features' values for all emotions, it affects much less the relative positions with respect to each other. We are investigating ways of exploring this, and expect it to yield much higher recognition results.

Most of the data manipulation in this thesis is done using MATLAB which is relatively slow compared to C/C++ and other compiled programming languages but has very good vector/matrix manipulation abilities. Any real-life real-time application will probably not be using MATLAB, so manipulating large vectors at every time step will probably make the whole process too slow. Therefore, in the online version of the algorithm we only use features whose values can be updated at every time step with minimal computational cost. The same features proposed in the offline version can be iteratively updated using simple algorithms, thus minimizing the size of data to be manipulated in real time.

1.2 Use of physiological data for recognition of cognitive states

The second part of the thesis involves the study of cognitive load and performance under different driving conditions. According to the Mental Workload model [20], performance drops sharply when a person enters the Mental Overload regime. Therefore, talking on a cellular phone or performing some other secondary task may only be dangerous under some very demanding driving conditions and harmless in most other routine driving cases, and the onset of the dangerous regime may be person-specific. We would like to be able to recognize such potentially dangerous situations so that they can be avoided. If we could somehow predict from a variety of measures that the driver is close to the onset of overload, we might be able to prevent it by, for example, temporarily preventing the cellular from ringing.

Subjects were asked to drive a driving simulator past several curves while keep-

ing their speed close to a predetermined constant value. In some cases they were simultaneously asked to listen to random numbers from a speech-synthesis software and perform simple mathematical tasks. Several measures drawn from the subjects' driving behavior were examined as possible indicators of either the subjects' performance or their mental workload. These included lane deviation, distance and time to lane crossing, and steering entropy [1, 13]. Cases with a sharp drop in performance were identified. The study was used as a guideline for a more thorough experiment where subjects' physiological data were recorded and then used to predict these cases of inadequate performance.

1.3 Thesis summary

In the chapters that follow, we elaborate on the specific experiments conducted, the pattern recognition methods used and the results obtained. In Chapter 2 we analyze the emotion expression experiment, the features extracted from the physiological data, the pattern recognition methods applied and the best results for several offline classifiers. In Chapter 3 we use longer data, including transitions between emotions. We also use several extra features, and mention the best results for the same classifiers used previously. In Chapter 4 we extend the previous experiment by building an online classifier based on the previous data, features and classifiers. We mention the best results and some potential problems with a real-life application. In Chapter 5 we look into cognitive load, through a pilot study conducted with subjects "driving" through curves under various speeds in a simulator, while performing simple mathematical tasks. The chapter includes the experimental setup, results and observations. These observations are then used in Chapter 6 to set up another experiment where subjects' physiological data are recorded as they drive and perform simple mathematical tasks, similar to the previous ones.

Chapter 2

Offline emotion expression recognition from physiological data

2.1 Introduction

This chapter addresses emotion recognition, specifically the recognition by computer of affective information expressed by people, through use of physiological data. This is part of a larger effort in "affective computing," computing that "relates to, arises from, or deliberately influences emotions" [14]. Affective computing has numerous applications and motivations, one of which is giving computers the skills involved in so-called "emotional intelligence," such as the ability to recognize a person's emotions. Such skills have been argued to be more important in general than mathematical and verbal abilities in determining a person's success in life [7]. Recognition of emotional information is a key step toward giving computers the ability to interact more naturally and intelligently with people.

The research described here focuses on recognition of emotional states during deliberate emotional expression by an actress. The actress, trained in guided imagery, used the Clynes method of sentic cycles to assist in eliciting the emotional states [3]. For example, to elicit the state of "Neutral," (no emotion) she focused on a blank

piece of paper or a typewriter. To elicit the state of "Anger" she focused on people who aroused anger in her. This process was adapted for the eight states: Neutral (no emotion) (N), Anger (A), Hate (H), Grief (G), Platonic Love (P), Romantic Love (L), Joy (J), and Reverence (R). Simultaneous with the visualization she pushed with her hand on a firm surface in a way that was intended to express each state. This effort at physical expression is supposed to help increase the intensity of the emotions felt.

The specific states one would want a computer to recognize will depend on the particular application. The eight emotions used in this research are intended to be representative of a broad range, which can be described in terms of the "arousal-valence" space commonly used by psychologists [12]. The arousal axis ranges from calm and peaceful to active and excited, while the valence axis ranges from negative to positive. For example, anger was considered high in arousal, while reverence was considered low. Love was considered positive, while hate was considered negative.

There has been prior work on emotional expression recognition from speech and from image and video; this work, like ours, has focused on deliberately expressed emotions. The problem is a hard one when you look at the few benchmarks which exist. In general, people can recognize affect in neutral-content speech with about 60% accuracy, choosing from among about six different affective states [16]. Computer algorithms can match this accuracy but only under more restrictive assumptions, such as when the sentence content is known. Facial expression recognition is easier, and the rates computers obtain are higher: from 80-98% accuracy when recognizing 5-7 classes of emotional expression on groups of 8-32 people [22, 6]. Facial expressions are easily controlled by people, and easily exaggerated, facilitating their discrimination.

Emotion recognition can also involve other modalities such as analyzing posture, gait, gesture, and a variety of physiological features in addition to the ones described in this paper. Additionally, emotion recognition can involve prediction based on cognitive reasoning about a situation, such as "That goal is important to her, and he just prevented her from obtaining it; therefore, she might be angry at him." Such a framework for analysis of affective dynamics has been developed under Affect Control Theory [9, 17]. The best emotion recognition is likely to come from pattern recognition

and reasoning applied to a combination of all of these modalities, including both low-level signal recognition, and higher-level reasoning about the situation [14].

For the research described here, four physiological signals of an actress were recorded during deliberate emotional expression. The signals measured were electromyogram (EMG) from the jaw, representing muscular tension or jaw clenching, blood volume pressure (BVP) and skin conductivity (GSR) from the fingers, and respiration from chest expansion. Data was gathered for each of the eight emotional states for approximately 3 minutes each. This process was repeated for several weeks. The four physiological waveforms were each sampled at 20 samples a second. The experiments use 2000 samples per signal, for each of the eight emotions, gathered over 20 days (Fig. 2-1). Hence there are a total of 32 signals a day, and 80 signals per emotion.

Very little work has been done on pattern recognition of emotion from physiological signals, and there is controversy among emotion theorists whether or not emotions do occur with unique patterns of physiological signals. Some psychologists have argued that emotions might be recognizable from physiological signals given suitable pattern recognition techniques [2], but nobody has yet to demonstrate which physiological signals, or which features of those signals, or which methods of classification, give reliable indications of an underlying emotion, if any. This paper suggests signals, features, and pattern recognition techniques for offline recognition of all 8 emotions examined, and presents results suggesting that emotions can be recognized from physiological signals at significantly higher than chance probabilities.

2.2 Choice of features

A very important part in recognizing emotional states, as with any pattern recognition procedure, is to determine which features are most relevant and helpful. This helps both in reducing the amount of data stored and in improving the performance of the recognizer.

Let the four raw signals, the digitized EMG, BVP, GSR, and Respiration wave-

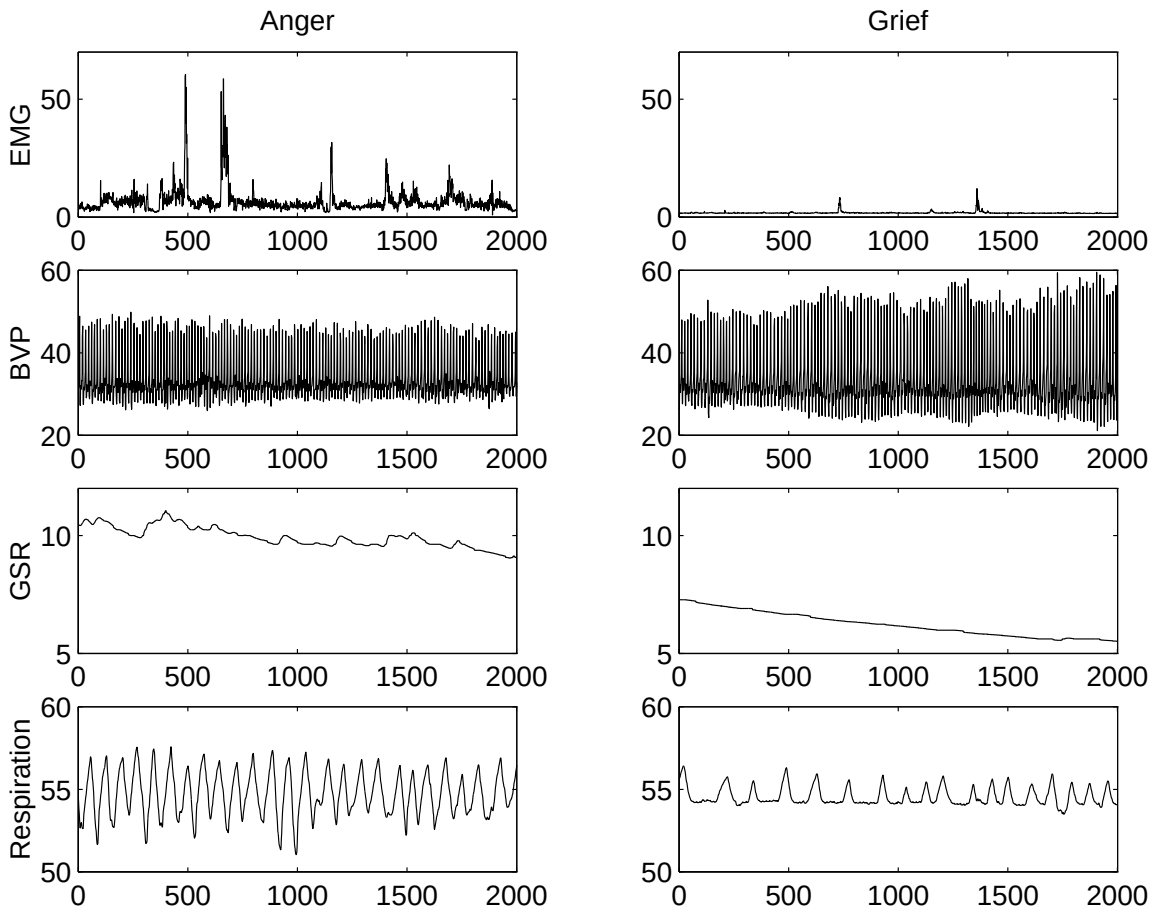

Figure 2-1: Examples of four physiological signals measured from an actress while she intentionally expressed anger (left) and grief (right). From top to bottom: electromyogram (microvolts), blood volume pressure (percent reflectance), galvanic skin conductivity (microSiemens), and respiration (percent maximum expansion). The signals were sampled at 20 samples a second. Each box shows 100 seconds of response. The segments shown here are visibly different for the two emotions, which was not true in general.

forms, be designated by $(S^i), i = 1, 2, 3, 4$. Each signal is gathered for 8 different emotions each session, for 20 sessions. Let S_n^i represent the value of the n^{th} sample of the i^{th} raw signal, where $n = 1...N$ and $N = 2000$ samples. Let $\tilde{S}_n^i$ refer to the normalized signal (zero mean, unit variance), formed as:

$$\tilde{S}_n^i = \frac{S_n^i - \mu^i}{\sigma^i} \qquad i = 1, ..., 4$$

where μ^i and σ^i are the means and standard deviations explained below. We extract 6 types of features for each emotion, each session:

1. the means of the raw signals (4 values)

$$(\mu^i) = \frac{1}{N} \sum_{n=1}^{N} S_n^i \qquad i = 1, ..., 4 \tag{2.1}$$

2. the standard deviations of the raw signals (4 values)

$$(\sigma^i) = \left(\frac{1}{N-1} \sum_{n=1}^{N} \left(S_n^i - (\mu^i) \right)^2 \right)^{1/2} \qquad i = 1, ..., 4 \tag{2.2}$$

3. the means of the absolute values of the first differences of the raw signals (4 values)

$$(\delta_1^i) = \frac{1}{N-1} \sum_{n=1}^{N-1} |S_{n+1}^i - S_n^i| \qquad i = 1, ..., 4 \tag{2.3}$$

4. the means of the absolute values of the first differences of the normalized signals (4 values)

$$(\tilde{\delta}_1^i) = \frac{1}{N-1} \sum_{n=1}^{N-1} \left| \tilde{S}_{n+1}^i - \tilde{S}_n^i \right| = \frac{(\delta_1^i)}{(\sigma^i)} \qquad i = 1, ..., 4. \tag{2.4}$$

5. the means of the absolute values of the second differences of the raw signals (4 values)

$$(\delta_2^i) = \frac{1}{N-2}\sum_{n=1}^{N-2}\left|S_{n+2}^i - S_n^i\right| \qquad i = 1, ..., 4 \tag{2.5}$$

6. the means of the absolute values of the second differences of the normalized signals (4 values)

$$(\tilde{\delta}_2^i) = \frac{1}{N-2}\sum_{n=1}^{N-2}\left|\tilde{S}_{n+2}^i - \tilde{S}_n^i\right| = \frac{(\delta_2^i)}{(\sigma^i)} \qquad i = 1, ..., 4 \tag{2.6}$$

Therefore, each emotion is characterized by 24 features, corresponding to a point in a 24-dimensional space. The classification can take place in this space, in an arbitrary subspace of it, or in a space otherwise constructed from these features. The total number of data in all cases is 20 points per class for each of the 8 classes, 160 data points in total.

Note that not all the features are independent; in particular, two of the features are nonlinear combinations of the other features. We expect that dimensionality reduction techniques will be useful in selecting which of the proposed features contain the most significant discriminatory information.

2.3 Dimensionality reduction

There is no guarantee that the features chosen above are all appropriate for emotion recognition. Nor is it guaranteed that emotion recognition from physiological signals is possible. Furthermore, a very limited number of data points—20 per class—is available. Hence, we expect that the classification error may be high, and may further increase when too many features are used. Therefore, reductions in the dimensionality of the feature space need to be explored, among with other options. Here focus on three methods for reducing the dimensionality, and evaluate the performance of these methods.

2.3.1 Sequential Floating Forward Search

The Sequential Floating Forward Search (SFFS) method [15] is chosen due to its consistent success in previous evaluations of feature selection algorithms, where it has recently been shown to outperform methods such as Sequential Forward and Sequential Backward Search (SFS, SBS), Generalized SFS and SBS, and Max-Min, [10] in several benchmarks. Of course the performance of SFFS is data dependent and the data here is new and difficult; hence, the SFFS may not be the best method to use. Nonetheless, because of its well documented success in other pattern recognition problems, it will help establish a benchmark for the new field of emotion recognition and assess the quality of other methods.

The SFFS method takes as input the values of n features. It then does a non-exhaustive search on the feature space by iteratively adding and subtracting features. It outputs one subset of m features for each m, $2 \leq m \leq n$, together with its classification rate. The algorithm is described in detail in [15]. For each subset size it maintains the criterion value $J(X_m)$ of the best feature subset X_m of that size found so far, as well as the subset which gave this value. The first two features are selected using the SFS method described in [11]. The rest of the algorithm can summarized in the following 3 steps, quoted from [15]:

- *Step 1:* **Inclusion.** *Select the most significant feature with respect to X and add it to X. Continue to step 2.*

- *Step 2:* **Conditional exclusion.** *Find the least significant feature k in X. If it is the feature just added, then keep it and return to step 1. Otherwise, exclude the features k. Note that X is now better than it was before step 1. Continue to step 3.*

- *Step 3:* **Continuation of conditional exclusion.** *Again, find the least significant feature in X. If its removal will (a) leave X with at least 2 features, and (b) the value of $J(X)$ is greater than the criterion value of the best feature subset of that size found so far, then remove it and repeat step 3. When these two conditions cease to be satisfied, return to step 1.*

2.3.2 Fisher Projection

Fisher projection [5] is a well-known method of reducing the dimensionality of the problem in hand, which involves less computation than SFFS. The goal is to find a projection W of the data to a space of fewer dimensions than the original where the classes are well separated. The algorithm is summarized below. First we define the *within-class scatter matrix*:

$$S_W = \sum_{j=1}^{c} \sum_{x \in \chi_j} (x - m_j)(x - m_j)^T \tag{2.7}$$

where c is the number of classes, x are the data points in the original space, χ_j is the subset of data in class j, m_j is the sample mean for class j in the original space and m is the overall sample mean. We also define the *between-class scatter matrix*:

$$S_B = \sum_{j=1}^{c} n_j (m_j - m)(m_j - m)^T \tag{2.8}$$

where n_j is the number of data points in class j. Then the original features, x, are projected through a linear transformation matrix W (to be determined) to a lower dimensional space. Therefore, the new features y are given by:

$$y = W^T x \tag{2.9}$$

and the scatter matrices in the new space are given by:

$$\tilde{S}_W = \sum_{j=1}^{c} \sum_{y \in \psi_j} (y - \tilde{m}_j)(y - \tilde{m}_j)^T = W^T S_W W \tag{2.10}$$

$$\tilde{S}_B = \sum_{j=1}^{c} n_j (\tilde{m}_j - \tilde{m})(\tilde{m}_j - \tilde{m})^T = W^T S_B W \tag{2.11}$$

where $\tilde{m}_j$ is the sample mean for class j in the reduced space and $\tilde{m}$ is the overall sample mean in the reduced space.

The criterion $J(W)$ to be maximized is defined as the ratio of the determinant of the between-class scatter matrix over the determinant of the within-class scatter

matrix in the projected data:

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{2.12}$$

It turns out [21] that the columns of a W maximizing J are the generalized eigenvectors that correspond to the largest eigenvalues in:

$$S_B w_i = \lambda_j S_W w_i \tag{2.13}$$

Due to the nature of the Fisher projection method, the data can only be projected down to $c-1$ (or fewer if one wants) dimensions, assuming that originally there are more than $c-1$ dimensions and c is the number of classes.

It is important to keep in mind that if the amount of training data is inadequate, or the quality of some of the features is questionable, then some of the dimensions of the Fisher projection may be a result of noise rather than a result of differences among the classes. In this case, Fisher might find a meaningless projection which reduces the error in the training data but performs poorly in the testing data. For this reason, projections down to fewer than $c-1$ dimensions are also evaluated in the paper.

Furthermore, since 24 features is high for the amount of training data here, and since the nature of the data is so little understood that these features may contain superfluous measures, we decided to try an additional approach: applying the Fisher projection not only to the original 24 features, but also to several "structured subsets" of the 24 features, which are described further below. Although in theory the Fisher method finds its own most relevant projections, the evaluation conducted below indicates that better results are obtained with the structured subsets approach.

Note that if the number of features n is smaller than the number of classes c, the Fisher projection is meaningful only up to at most $n-1$ dimensions. Therefore in general the number of Fisher projection dimensions d is $1 \leq d \leq \min(n, c) - 1$. For example, when 24 features are used on all 8 classes, all $d = [1, 7]$ are tried. When 4 features are used on 8 classes, all $d = [1, 3]$ are tried.

2.3.3 Hybrid SFFS with Fisher Projection (SFFS-FP)

As mentioned above, the SFFS algorithm proposes one subset of m features for each m, $2 \leq m \leq n$. Therefore, instead of feeding the Fisher algorithm with all 24 features or with structured subsets, we can use the subsets that the SFFS algorithm proposes as our input to the Fisher Algorithm. Note that the SFFS method is used here as a simple preprocessor for reducing the number of features fed into the Fisher algorithm, and not as a classification method. We call this hybrid method SFFS-FP.

2.4 Evaluation

We now describe how we obtained the results shown in Table 2.3. A discussion of these results follows below.

2.4.1 Methodology

The Maximum a Posteriori (MAP) classification is used for all Fisher Projection methods. The leave-one-out method is chosen for cross validation because of the small amount of data available. More specifically, here is the algorithm that is applied to every data point:

1. The data point to be classified (the testing set only includes one point) is excluded from the data set. The remaining data set will be used as the training set.

2. In the case where a Fisher projection is to be used, the projection matrix is calculated from only the training set. Then both the training and testing set are projected down to the d dimensions found by Fisher.

3. Given the feature space, original or reduced, the data in that space is assumed to be Gaussian. The respective means and covariance matrices of the classes are estimated from the training data.

4. The posterior probability of the testing set is calculated: the probability the test point belongs to a specific class, depending on the specific probability distribution of the class and the priors.

5. The data point is then classified as coming from the class with the highest posterior probability.

The above algorithm is first applied on the original 24 features (**Fisher-24**). Because this feature set was expected to contain a lot of redundancy and noise, we also chose to apply the above algorithm on various "structured subsets" of 4, 6 and 18 features defined as follows:

Fisher-4 All combinations of 4 features are tried, with the constraint that each feature is from a different signal (EMG, BVP, GSR, Respiration). This gives a total of $6^4 = 1296$ combinations, which substantially reduces the $\binom{24}{4}$ =10626 that would result if all combinations were to be tried. The results of this evaluation may give us an indication of which type of feature is most useful for each physiological signal.

Fisher-6 All combinations of 6 features are tried, with the constraint that each feature has to be of a different type: (1)-(6). This gives a total of $4^6 = 4096$ combinations instead of $\binom{24}{6}$=134596 if all combinations were to be tried. The results of this evaluation may give us an indication which physiological signal is most useful for each type of feature.

Fisher-18 All possible combinations of 18 features are tried, with the constraint that exactly 3 features are chosen from each of the types (1)-(6). That again gives a total of $4^6 = 4096$ combinations, instead of $\binom{24}{18}$=134596 if all combinations were to be tried. The results of this evaluation may give us an indication which physiological signal is least useful for each feature.

The SFFS software we used included its own evaluation method, K-nearest neighbor (kNN) [4], in choosing which features were best. For the SFFS-FP method, the procedure below was followed: The SFFS algorithm outputs one set of m features for each $2 \leq m \leq n$, and for each $1 \leq k \leq 20$. All possible Fisher projections are then calculated for each such set.

Another case, not shown in Table 2.3, was investigated. Instead of using a Fisher projection, we tried all possible **2-feature subsets**, and evaluated their class according to the maximum a posteriori probability, using cross-validation. The best classification in this case was consistently obtained when using the mean of the EMG signal (feature μ^1 above) and the mean of the absolute value of the first difference of the normalized Respiration signal (feature $\tilde{\delta}_1^4$ above) as the two features. The only result almost comparable to other methods was obtained when discriminating among Anger, Joy and Reverence where a linear classifier scores 71.66% (43/60). When trying to discriminate among more than 3 emotions, the results were not significantly better than random guessing, while the algorithm consumed too much time in an exhaustive search.

Attempting to discriminate among 8 different emotional states is unnecessary for many applications, where 3 or 4 emotions may be all that is needed. We therefore evaluated the three methods here not only for the full set of eight emotion classes, but also for sets of three, four, and five classes that seemed the most promising in preliminary tests.

2.4.2 Results

The results of all the emotion subsets and classification algorithms are shown in Table 2.3. All methods performed significantly better than random guessing, indicating that there is emotional discriminatory information in the physiological signals.

When Fisher was applied to structured subsets of features, the results were always better than when Fisher was applied to the original 24 features.

3 emotions In runs using the Fisher-24 algorithm, the two best 3-emotion subsets turned out to be the *Anger-Grief-Reverence (AGR)* and the *Anger-Joy-Reverence (AJR)*. All the other methods are applied on just these two triplets for comparison.

4 emotions In order to avoid trying all the possible quadruplets with all the possible methods, we use the following arguments for our choices:

Anger-Grief-Joy-Reverence (AGJR): These are the emotions included in the best-classified triplets. Furthermore, the features used in obtaining the best results above

Anger (High Arousal Negative Valence)	Joy (High Arousal Positive Valence)
Grief (Low Arousal Negative Valence)	Reverence (Low Arousal Positive Valence)

Table 2.1: Anger, Grief, Joy, and Reverence can be seen as placed in the four corners of a valence-arousal plot, a common taxonomy used by psychologists in categorizing the space of emotions, and may therefore be highly discriminable.

were not the same for the two cases. Therefore a combination of these features may be discriminative for all 4 emotions. Finally, these emotions can be seen as placed in the four corners of a valence-arousal plot, a common taxonomy used by psychologists in categorizing the space of emotions (See Table 2.1).

Neutral-Anger-Grief-Reverence (NAGR) In results from the 8-emotion classification using the Fisher-24 algorithm, the resulting confusion matrix (Table 2.2) shows that Neutral, Anger, Grief, and Reverence are the four emotions best classified and least confused with each other.

5-emotions The 5-emotion subset examined is the one including the emotions in the 2 quadruplets chosen above, namely the *Neutral-Anger-Grief-Joy-Reverence (NAGJR)* set.

The best classification rates obtained by SFFS and SFFS-FP are reported in Table 2.3, while the number of features used in producing these rates can be seen in Table 2.4. We can see that in SFFS a small number m_{SFFS} of the 24 original features gave the best results. For SFFS-FP a slightly larger number $m_{SFFS-FP}$ of features tended to give the best results, but still smaller than 24. These extra features found useful in SFFS-FP, could be interpreted as containing some useful information, but together with a lot of noise. That is because feature selection methods like SFFS can only accept/reject features, while the Fisher algorithm can also scale them appropriately, performing a kind of "soft" feature selection and thus making use of such noisy features.

	N	A	H	G	P	L	J	R	Total
N	**10**	2	3	0	0	1	0	4	20
A	2	**11**	2	3	0	1	1	0	20
H	5	0	**3**	3	2	4	1	2	20
G	1	3	1	**10**	2	0	2	1	20
P	0	0	5	0	**6**	2	5	2	20
L	0	2	3	2	2	**8**	3	0	20
J	1	0	2	4	5	3	**5**	0	20
R	1	0	0	0	6	2	0	**11**	20
Total	20	18	19	22	23	21	17	20	160

Table 2.2: Confusion matrix in the classification of 8 emotions, when using Fisher-24. An entry's row is the true class, the column is what it was classified as. The diagonal shows all the correctly classified data, 64 out of a total of 160, or 40%. Neutral, Anger, Grief, and Reverence are the most discriminated.

Number of Emotions	Random Guess (%)	SFFS (%)	Fisher-24 (%)	Structured subsets (%)			SFFS-FP (%)
				4-feat	6-feat	18-feat	
8	12.50	40.62	40.00	34.38	41.25	48.75	46.25
5 (NAGJR)	20.00	64.00	60.00	53.00	63.00	71.00	65.00
4 (NAGR)	25.00	70.00	61.25	61.25	70.00	72.50	68.75
4 (AGJR)	25.00	72.50	60.00	58.75	70.00	68.75	67.50
3 (AGR)	33.33	83.33	71.67	75.00	83.33	81.67	80.00
3 (AJR)	33.33	88.33	66.67	73.33	83.33	81.67	83.33

Table 2.3: Classification rates for several algorithms and emotion subsets.

In Table 2.5 one can see that for greater numbers of emotions and greater numbers of features, the best-performing number of Fisher dimensions tends to be less than the maximum number of dimensions Fisher can calculate, confirming our earlier expectations (Section 2.3.2).

2.5 Day dependence

As mentioned previously, the data were gathered in 20 different sessions, one session each day. During their classification procedure, we noticed high correlation between the values of the features of different emotions in the same session. In this section we

Number of Emotions	m_{SFFS}	$m_{SFFS-FP}$
8	13	17
5 (NAGJR)	12	15
4 (NAGR)	9	19
4 (AGJR)	7	12
3 (AGR)	2	12
3 (AJR)	6	7

Table 2.4: Minimum number of features m proposed by the SFFS algorithms which gave the best results. When a range of SFFS algorithms performed equally well, only the one proposing the fewest features is listed.

Number of Emotions	Structured subsets			Fisher-24	SFFS-FP	Ratio
	4-feature	6-feature	18-feature			
8	3/3	3/5	5/7	6/7	4,5/7	4:1
5 (NAGJR)	3/3	4/4	3/4	3/4	3/4	3:2
4 (NAGR)	3/3	3/3	3/3	3/3	3/3	0:5
4 (AGJR)	3/3	2/3	2,3/3	3/3	2/3	3:2
3 (AGR)	2/2	2/2	2/2	2/2	2/2	0:5
3 (AJR)	2/2	2/2	2/2	1/2	2/2	1:4
Ratio	0:6	2:4	3:3	3:3	3:3	11:19

Table 2.5: Number of dimensions used in the Fisher Projections which gave the best results, over the maximum number of dimensions that could be used. The last row and column give the ratio of cases where these two values were not equal, over the cases that they were.

first quantify this phenomenon by building a day (session) classifier and then use it to improve the emotion classification results by including the day information in the features.

2.5.1 Day classifier

We use the same set of 24 features, the Fisher algorithm, and the leave-one-out method as before, only now there are $c = 20$ classes instead of 8. Therefore the Fisher projection is meaningful from 1 to 19 dimensions. The resulting "day classifier" using the Fisher projection and the leave-one-out method with MAP classification, yields

a classification accuracy of 133/160 (83%), when projecting down to 6,9,10 and 11 Fisher dimensions. This is better than all but one of the results reported above, and far better than random guessing (5%). We note the following on this result:

- The signals, as well as the features extracted from them, are highly dependent on the day the experiment is held.
- This can be because, even if the actress is intentionally expressing a specific emotion, there is still an underlying emotional and physiological state which affects the overall results of the day.
- This may also be related to technical issues, like the amount of gel used in the sensing equipment (for the BVP and GSR signals), or external issues like the temperature in a given day, affecting the perspiration and possibly the blood pressure of the actress.
- It should be expected that a more sophisticated algorithm would give even better results. For example we only tried using all 24 features, rather than a (structured, SFFS, or SFFS-FP) subset of them.

A possible model for the emotions explaining part of the day dependence could be thought of as follows: At any point in time the physiological signals are a combination of a long-term slow-changing mood (for example a day-long frustration) or physiological situation (for example lack of sleep) and of a short-term emotion caused by more sudden changes in the environment (for example the arrival of some bad news). It is not clear how the different emotions that coexist at any given time affect the behavior or the physiology of a subject. Nevertheless, it seems that in the current context knowledge of the day (as part of the features) may help in establishing a baseline which could in turn help in recognizing the different short-term emotions within a day. This baseline may be as simple as subtracting a different value depending on the day, or something more complicated.

It is also relevant to consider conditioning the recognition tests on only the day's data, as there are many applications where the computer wants to know the person's

emotional response *right now* so that it can change its behavior accordingly. In such applications, not only are interactive-time recognition algorithms needed, but they need to be able to work based on only present and past information, i.e., causally. In particular, they will probably need to know what range of responses is typical for this person, and base recognition upon deviations from this typical behavior. The ability to estimate a good "baseline" response, and to compare the present state to this baseline is important.

2.5.2 Establishing a day-dependent baseline

According to the results of the previous section, the features extracted from the signals are highly dependent on the day the experiment was held. Therefore, we would like to augment the set of features to include both the **Original** set of 24 features and a second set incorporating information on the day the signals were extracted.

The Day Matrix

Let us think of a case where the data come from only 2 different days and only 1 feature is extracted from the data (This is the only way the following manipulations can be visualized, but the manipulation trivially extends to more features). Although the feature values of one class are always related to the values of the other classes in the same way (for example the mean EMG for anger may always be higher than the mean EMG for Joy), the actual values may be highly day-dependent (Fig. 2-2a). To alleviate this problem an extra dimension can be added before the features are fed to the Fisher Algorithm (Fig. 2-2b). If the data came from 3 different days, 2 extra dimensions would have to be added rather than one (Fig. 2-2c), etc. Therefore, in the general case $D-1$ extra dimensions are needed for data coming from D different days, and 19 extra dimensions are needed in our case. The above can be also seen as using the minimum number of dimensions so that each of D points can be at equal distance from all others. Therefore the $D-1$ dimensional vector will contain the coordinates of one such point for each day. This vector is the same for all emotions recorded in the same day.

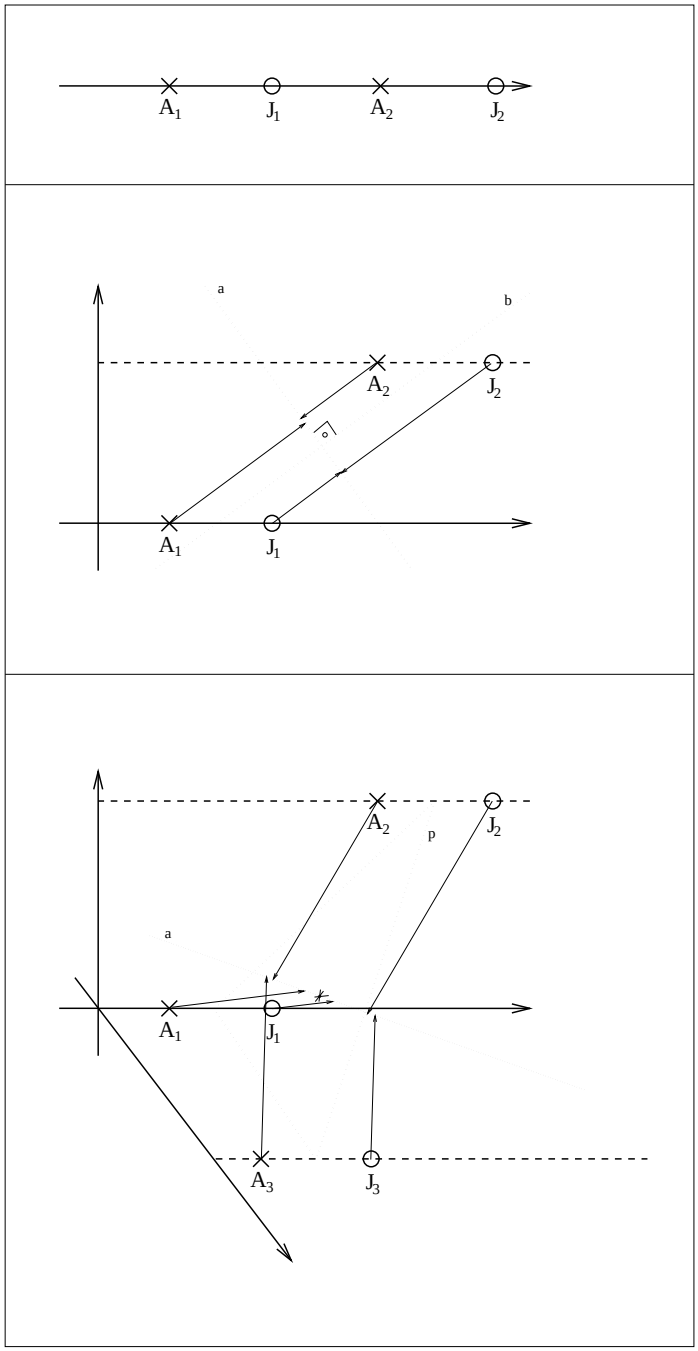

Figure 2-2: Fictitious example of a highly day-dependent feature for 2 emotions from 2 different days. (a) The feature values for (A)nger and (J)oy from 2 different days. (b) Addition of an extra dimension allows for a line b to separate Anger from Joy. The data can be projected down to line a, so the addition of the new dimension did not increase the final number of features. (c) In the case of data from 3 different days, addition of 2 extra dimensions allows for a plane p to separate Anger from Joy. The data can again be projected down to line a, not increasing the final number of features.

Feature Space	SFFS (%)	Fisher (%)	SFFS-FP (%)
Original (24)	40.62	40.00	46.25
Original+Day (43)	N/A	49.38	50.62

Table 2.6: Classification Rates for the 8-emotion case using several algorithms and methods for incorporating the day information. "N/A" denotes that SFFS feature selection is meaningless if applied to the Day Matrix.

Feature Space	SFFS (%)	Fisher (%)	SFFS-FP (%)
Original (24)	42.86	39.29	45.00
Orig.+Day (43)	N/A	39.29	45.71
Orig.+Base. (48)	49.29	40.71	54.29
Orig.+Base.+Day (67)	N/A	35.00	49.29

Table 2.7: Classification Rates for the 7-emotion case using several algorithms and methods for incorporating the day information. "N/A" denotes that SFFS feature selection is meaningless if applied to the Day Matrix.

Another approach that we investigated involves constructing a **Baseline Matrix** where the Neutral (no emotion) features of each day are used as a baseline for (subtracted from) the respective features of the remaining 7 emotions of the same day. This gives an additional 24x20 matrix for each emotion.

The complete 8-emotion classification results can be seen in Table 2.6, while the 7-emotion classification results can be seen in Table 2.7. Random guessing would be 12.50% and 14.29% respectively. The results are several times that for random guessing, indicating that significant emotion classification information has been found in this data.

2.6 Conclusions

The results suggest that there is significant information in physiological signals for classifying the affective state of a person who is deliberately expressing a small set of emotions. They also reveal a very high day dependence which can be seen both as

something to be aware of, as well as something worth exploiting to better understand and better recognize human emotions.

Success rates above 80% when recognizing 3 emotions and 50% when recognizing 8 emotions are encouraging and signify that physiological data contain information about the human emotional state. Nevertheless it is very important to keep in mind that these were intentionally expressed emotions, of only one subject, expressed in the same sequence every time (with unknown interactions between emotions) and all had similar duration (something not necessarily true with "real" emotions). In the next chapter we will look into some slightly different data available and try some new features. These data are longer than the ones used in this chapter and include the transition periods between emotions.

Chapter 3

Offline Recognition using improved data and features

3.1 Introduction

3.1.1 Data

In the previous sections, we used data consisting of 2000 samples-per-signal, for each of the eight emotions, gathered over 20 days (Fig. 2-1).

The data were originally gathered in 34 sessions where the 8 different emotions were expressed one after the other. Each full session lasted around 25 minutes, resulting in around 28 to 33 thousand samples per signal, with each emotion being around 2 to 5 thousand samples long, due to the randomness of the Clynes method of eliciting the emotional states [3]. In several occasions one or more sensors failed during parts of the experiment. The first 20 sessions were the ones used in the previous sections, choosing the last 2000 samples from each emotional state while trying to avoid parts where the sensors had failed. The question which remained was if any information could be extracted from the uninterrupted data, like transition characteristics, or if an online classifier could be built. Therefore, we revisited the data from the full sessions and chose 20 days in which the sensors did not fail during any part of the experiment. 16 of the original days and another 4 which had not been used before were included.

We call this new set of data "Set B", with "Set A" being the original data analyzed in the previous sections. Some comparative results between the common days of the two slightly different sets of data can be seen in Table 3.2.

3.1.2 Features

Using peak detection on the Blood Volume Pressure signal, the Heart Rate can be calculated. The same 6 features proposed in Section 2.2 can be extracted from the Heart Rate as well. Additionally, a set of 11 other features have been proposed for use with these physiological data. Quoting from [8], these features are: *the mean EMG activity, the mean and mean slope of the skin conductivity, average heart rate and heart rate change, and the normalized mean, variance, and four power spectral density characteristics of the respiration signal.* We would like to see if the inclusion of any of the above features can improve classification.

3.1.3 Results

The results can be seen in Tables 3.3, 3.4 and 3.5. The confusion matrix for the case with the highest recognition rate, 81.25% can be seen in Table 3.1. Note that the total number of different features is 40 (rather than 41) because the mean EMG that was proposed in [8] was already included in the original 24 features.

We can see that in most cases, a small number m_{SFFS} of the original features gave the best results in SFFS. For SFFS-FP a slightly larger number $m_{SFFS-FP}$ of features tended to give the best results. These extra features found useful in SFFS-FP but not in pure SFFS, could be interpreted as containing some useful information, but together with a lot of noise. That is because feature selection methods like SFFS can only accept/reject features, while the Fisher algorithm can also scale them appropriately, performing a kind of "soft" feature selection and thus making use of such noisy features.

From this point on, all results mentioned are based on these new longer data.

	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 3.1: Confusion matrix in the classification of 8 emotions, when using SFFS-FP, starting with all 40 features and without using the Day Matrix. An entry's row is the true class, the column is what it was classified as. The diagonal shows all the correctly classified data, 130 out of a total of 160, or 81.25%.

Data	Without Day Matrix (%)	With Day Matrix (%)
Set A	42.97	46.09
Set B	54.69	54.69

Table 3.2: Comparative classification rates for the 16 common days (128 data points in total) between Data Sets A and B, using 24 features fed to the Fisher Algorithm. The results suggest that using the longer data (Set B) improves classification performance.

3.2 Conclusions

The results here confirm and expand upon our earlier results, which suggested that there is significant information in physiological signals for classifying the affective state of a person who is deliberately expressing a small set of emotions.

Success rates above 80% when recognizing 8 emotions are extremely high, even compared to the other existing methods of emotion recognition. Nevertheless it is very important to keep in mind that these were intentionally expressed emotions, of only one subject, expressed in the same sequence every time (with unknown interactions between emotions) and all had similar duration (something not necessarily true with "real" emotions). Therefore, plenty of work has to be done until a robust and easy-to-

Number of Features	Without Day Matrix			With Day Matrix	
	SFFS (%)	Fisher (%)	SFFS-FP (%)	Fisher (%)	SFFS-FP (%)
24	49.38	51.25	56.87	54.37	63.75
30 (incl. HR)	52.50	56.87	60.00	58.75	63.75
11 (other)	60.62	70.00	70.63	61.25	63.12
40 (incl. HR, other)	65.00	77.50	81.25	77.50	78.75

Table 3.3: Comparative classification rates for all 20 days (160 data points in total) of Data Set B and different features and methods used. The Day Matrix adds 19 features to the data fed to the Fisher Algorithm.

Number of Features	Without Day Matrix		With Day Matrix
	m_{SFFS}	$m_{SFFS-FP}$	$m_{SFFS-FP}$
24	14	16	19
30 (incl. HR)	5	7	22
11 (other)	11	7	7
40 (incl. HR, other)	8	25	32

Table 3.4: Number of features m proposed by the SFFS algorithms that gave the best results in Data Set B. When a range of SFFS algorithms performed equally well, only the one proposing the fewest features is listed.

use emotion recognizer is built. A first step could be looking into real time emotion recognition. That is, if the computer can sense the emotional state of the user the moment he/she actually is in this state, (online recognition), rather than analyzing the data later, when the user is already in another state (offline recognition). This could be considered in combination with the model of an underlying mood, which may change over longer periods of time. In that respect, the classification rate of a time window given a previous time window could yield useful information. In addition, it appears that although the underlying mood changes the features' values for all emotions, it affects much less the relative positions with respect to each other. The question is how frequently should the estimates of the baseline be updated to accommodate for the changes in the underlying mood. Also, how often can new data be incorporated and at any moment in time how have back do we have to look in

Number of Features	Without Day Matrix		With Day Matrix	
	Fisher	SFFS-FP	Fisher	SFFS-FP
24	7	7	4	4
30 (incl. HR)	4	5	3	4
11 (other)	5	6	5	3
40 (incl. HR, other)	7	5	7	6

Table 3.5: Number of dimensions used in the Fisher Projections which gave the best results, out of a maximum of 7 dimensions. When a range of Fisher Projections performed equally well, only the one using the fewest dimensions is listed.

order to establish the currect emotion? In the next chapter we look more closely at some of these questions and build a classifier capable of recognizing emotional states from physiological data in real time.

Chapter 4

Online Recognition

4.1 Introduction

Each day of Data Set B contains a continuous stream of data running through 8 different emotions. This data set is then appropriate for training and testing an online algorithm.

4.2 The iterative algorithm

Most of the data manipulation in this thesis has been done using MATLAB which is relatively slow compared to C/C++ and other compiled programming languages but has very good vector/matrix manipulation abilities. Any real-life real-time application will probably not be using MATLAB, so manipulating large vectors at every time step will probably make the whole process too slow. Therefore, in the online version of the algorithm we will only use features whose values can be updated at every time step with minimal computational cost. The 6 features per signal proposed previously can be iteratively updated using the following algorithm (where S_{N+1} is the value of the signal at the time step newly incorporated in the data and W is the width of the moving window in number of time steps):

For $3 \leq N < W$

$$\mu_{N+1} = \frac{N}{N+1}\mu_N + \frac{1}{N+1}S_{N+1} \tag{4.1}$$

$$\sigma_{N+1} = \left(\frac{N-1}{N}\sigma_N^2 + \frac{1}{N}S_{N+1}^2 - \frac{N+1}{N}\mu_{N+1}^2 + \mu_N^2\right)^{1/2} \tag{4.2}$$

$$\delta_{1_{N+1}} = \frac{N-1}{N}\delta_{1_N} + \frac{1}{N}|S_{N+1} - S_N| \tag{4.3}$$

$$\tilde{\delta}_{1_{N+1}} = \frac{\delta_{1_{N+1}}}{\sigma_{N+1}} \tag{4.4}$$

$$\delta_{2_{N+1}} = \frac{N-2}{N-1}\delta_{2_N} + \frac{1}{N-1}|S_{N+1} - S_{N-1}| \tag{4.5}$$

$$\tilde{\delta}_{2_{N+1}} = \frac{\delta_{2_{N+1}}}{\sigma_{N+1}} \tag{4.6}$$

And for $N \geq W$

$$\mu_{N+1} = \mu_N + \frac{1}{W}\left(S_{N+1} - S_{N+1-W}\right) \tag{4.7}$$

$$\sigma_{N+1} = \left(\sigma_N^2 + \frac{1}{W-1}(S_{N+1}^2 - S_{N+1-W}^2) - \frac{W}{W-1}\left(\mu_{N+1}^2 - \mu_N^2\right)\right)^{1/2} \tag{4.8}$$

$$\delta_{1_{N+1}} = \delta_{1_N} + \frac{1}{W-1}\left(\left|S_{N+1} - S_N\right| - \left|S_{N+2-W} - S_{N+1-W}\right|\right) \tag{4.9}$$

$$\tilde{\delta}_{1_{N+1}} = \frac{\delta_{1_{N+1}}}{\sigma_{N+1}} \tag{4.10}$$

$$\begin{aligned}\delta_{2_{N+1}} = \; & \delta_{2_N} + \tfrac{1}{W-2}\left(\left|S_{N+1} - S_{N-1}\right|\right. \\ & \left. - \left|S_{N+3-W} - S_{N+1-W}\right|\right)\end{aligned} \tag{4.11}$$

$$\tilde{\delta}_{2_{N+1}} = \frac{\delta_{2_{N+1}}}{\sigma_{N+1}} \tag{4.12}$$

The estimates for the first few steps can be calculated using the offline formulae (Eqns. 2.1-2.6)

The above iterations assume a continuous feed of data, therefore we will be using

the long continuous data of Set B, as mentioned earlier. Using all 5 signals (EMG, BVP, GSR, Respiration, and HR), gives a total of 30 features that can be calculated for every position of the moving window, for each one of the days.

4.3 Training data

Given that this is an online algorithm, it is not clear if we should use data from emotions of one day in the training of the classifier for other emotions of the same day. Therefore, *assuming that a person does not re-train the algorithm during the day*, we only use features from other days to train the classifier. Because of the small amount of days available, we use the leave-one-out method (2.4.1). This means that a new classifier is trained using 19 days and tested on the one left out, with the process repeated for all 20 days. Each day has around 30 thousand time steps, so a moving window can produce around that many sets of 30 features. But using all these sets for training would make the problem computationally very hard, requiring extreme amounts of disk space, memory and time, and would be almost useless, as consecutive time steps have very highly correlated features (for example in a 1000-time step moving window, 999 data points would be the same between consecutive time steps). Therefore, we arbitrarily choose to use a subset of 200 sets of features per emotion, updating around every 15 time steps. This produces 30400 training sets of features (200 sets of features per emotion times 8 emotions per day times 19 days) which are then fed into the Fisher Algorithm to produce a reduced dimensionality Fisher Projection.

4.4 Testing data

Using the Fisher Projection matrix, we calculate the posterior probabilities for all the sets of features (around 30 thousand data points) of the day we are testing and classify each one as coming from the emotion with the highest posterior probability.

4.5 Data labeling and moving window size

In the offline version, features were calculated from segments of data known to fully belong to only one emotion. In the online version, features are calculated based on data from a moving window. When the window includes the transition from one emotion to the next, features are calculated from data coming from 2 different emotions. It is not clear if these features should be included in the training of the classifier, and to which emotion. Similarly it is not clear if the classifier should be expected to classify these features to the previous or the next emotion during the testing phase. We expect our decisions on the training phase to influence the performance of the classifier in the testing phase.

The objective of an online emotion classifier is to first recognize as *correctly* as possible the emotional state of the user (high classification rate), and second to recognize it as *soon* as possible (high sensitivity). The former suggests a large window size, to minimize variance in the features within a class. It would also require that the features be considered as belonging to the previous emotion if most of the window is still in the previous emotion. On the contrary, the latter suggests a small window size, and the features of a window including the smallest part of a new emotion to be considered as belonging to the new emotion. Taking into account the above tradeoffs, we built and compared several classifiers, varying the following parameters:

W: We compare 5 different window sizes W (100, 200, 500, 1000, and 2000 time steps long). We also try combinations of 2, 3, 4 and all 5 window sizes. This is done by feeding to the Fisher Projection Algorithm a multiple of the 30 features calculated from each different window size for each data point (60 features when using 2 windows, 90 features when using 3 windows, etc.) Besides the 5 single-window cases, there are 10 pairs, 10 triplets, 5 quadruplets and 1 case of all 5 window sizes used, therefore a total of 31 different window size combinations.

W_{train_1}: A data point's features are used in the training of the new emotion when it is at least W_{train_1} time steps into the new emotion. We compare classifiers with $0 < W_{train_1} \leq W$. Normalizing provides $w_{train_1} = \frac{W_{train_1}}{W}$, $0 < w_{train_1} \leq 1$.

When using multiple moving windows, they are all positioned so that their individual w_{train_1}'s are equal. Therefore when a window of 100 data points is 20 data points into a new emotion (W_{train_1}=20, w_{train_1}=0.2), a window of 1000 data points is 200 data points into the new emotion (W_{train_1}=200, w_{train_1}=0.2).

W_{train_2}: A data point's features are used in the training of the previous emotion when it is at most W_{train_2} time steps into the new emotion. We compare classifiers with $-\frac{W}{2} \leq W_{train_2} \leq \frac{W}{2}$. Normalizing provides $w_{train_2} = \frac{W_{train_2}}{W}$, $-0.5 \leq w_{train_2} \leq 0.5$. When using multiple moving windows, they are all positioned so that their individual w_{train_2}'s are equal, as mentioned above.

W_{test_1}: A data point is expected to be classified as belonging to the new emotion when it is at least W_{test_1} time steps into the new emotion. We compare classifiers with $0 < W_{test_1} \leq W$. Testing has to occur online, so multiple moving windows have to incorporate the same new data point simultaneously, (their individual W_{train_1}'s are equal). Therefore when a window of 100 data points is 20 data points into a new emotion, a window of 1000 data points is also 20 data points into the new emotion (W_{test_1}=20 in both cases), not 200 data points as in the training.

W_{test_2}: A data point is expected to be classified as belonging to the previous emotion when it is at most W_{test_2} time steps into the new emotion. We compare classifiers with $-\frac{W}{2} \leq W_{test_2} \leq \frac{W}{2}$.

4.6 Definition of performance

In the case of an online algorithm, there are several options for how to define performance. We could try to combine the posterior probabilities of all data points in one emotion and end up with an overall posterior probability from which we could classify the whole segment. Alternatively we could use simple voting among the classification results of all data points within one emotion to come up with an overall classification of the whole segment. None of these methods are natural, because in real life we will not know the emotion boundaries of the data we are trying to classify. (Although such pre-segmented classification is what was used in the facial and vocal expression

recognition results alluded to in 2.1.)

Another measure of performance is the *data point* classification success rate. This is the ratio of the total number of data points correctly classified over the total number of data points in the day for which a classification was attempted. The results analyzed later use this definition of performance, but overall segment-classification performance will also be mentioned.

4.7 Results

In all 31 window-size combinations, the best results were obtained when the data were projected down to 7 Fisher dimensions ($c - 1$). This is probably because the increase in training data helps in reducing the effect of noise in the features, making all 7 dimensions contain useful information, unlike in the offline version.

In all single-window cases, the larger the window size, the better the results. In all other cases, the larger the maximum window size used, the better the results.

In all cases, the results when using a combination of window sizes were at least as good, and in most cases significantly improved, over using any subsets of these window sizes.

When using multiple moving windows in the training, they all have equal normalized overlaps w_{train_1}'s, as mentioned in Section 4.5. They also have equal w_{train_2}'s. Therefore Figures 4-1 and 4-3 plot performance against w_{train_1} and w_{train_2} respectively. Contrary to the training, when using multiple moving windows in the testing, they all have equal unnormalized overlaps W_{train_1}'s, as mentioned in Section 4.5. They also have equal W_{train_2}'s. Therefore Figures 4-2 and 4-4 plot performance against W_{test_1} and W_{test_2} respectively.

Using data points from the start of a new emotion, even though the window still includes data from the previous emotion ($w_{train_1} << 1$) in the training, seems to slightly improve the results (Fig. 4-1). On the contrary, using these data points in the testing, slightly worsens the overall results (Fig. 4-2). Therefore, they help improve the training of the classifier, but they themselves are not classified as well as

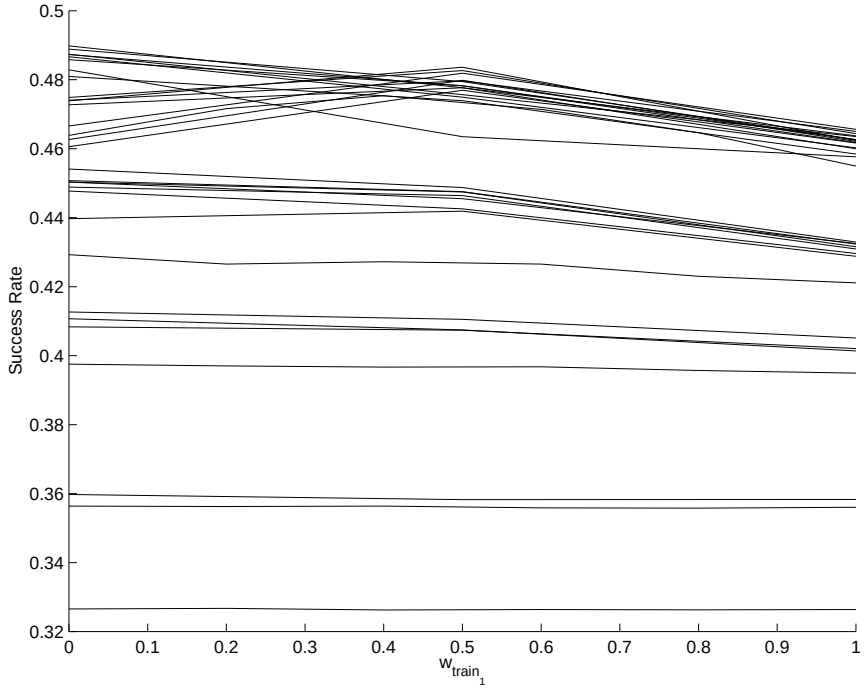

Figure 4-1: Success rate vs. normalized overlap w_{train_1} for different combinations of window sizes. Using data points from the start of a new emotion, even though the window still includes data from the previous emotion ($w_{train_1} << 1$) in the training, seems to slightly improve the results.

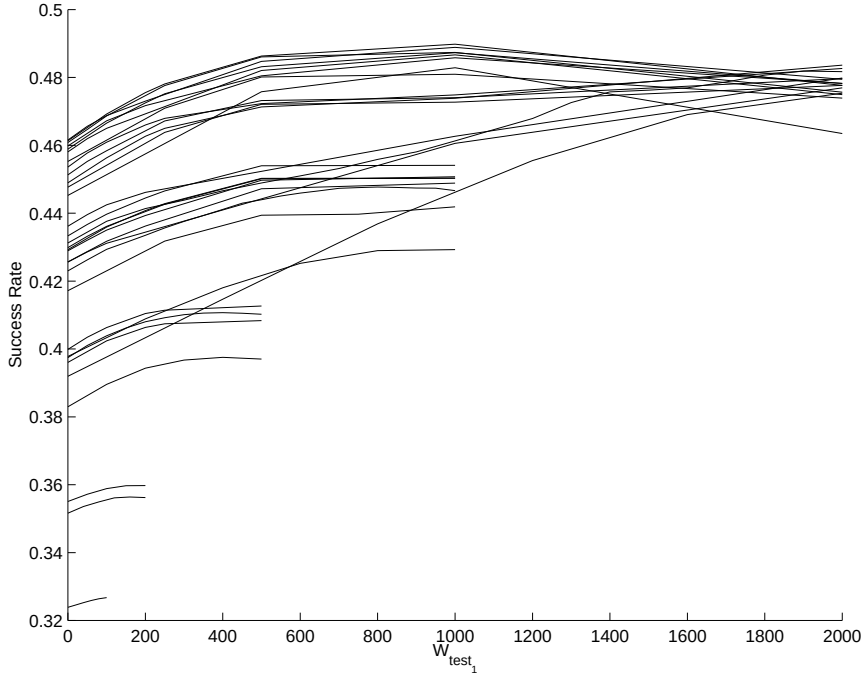

Figure 4-2: Success rate vs. unnormalized overlap W_{test_1} for different combinations of window sizes. Using data points from the start of a new emotion while the window still mostly includes data from the previous emotion (left part of the curves) in the testing, seems to slightly worsen the results.

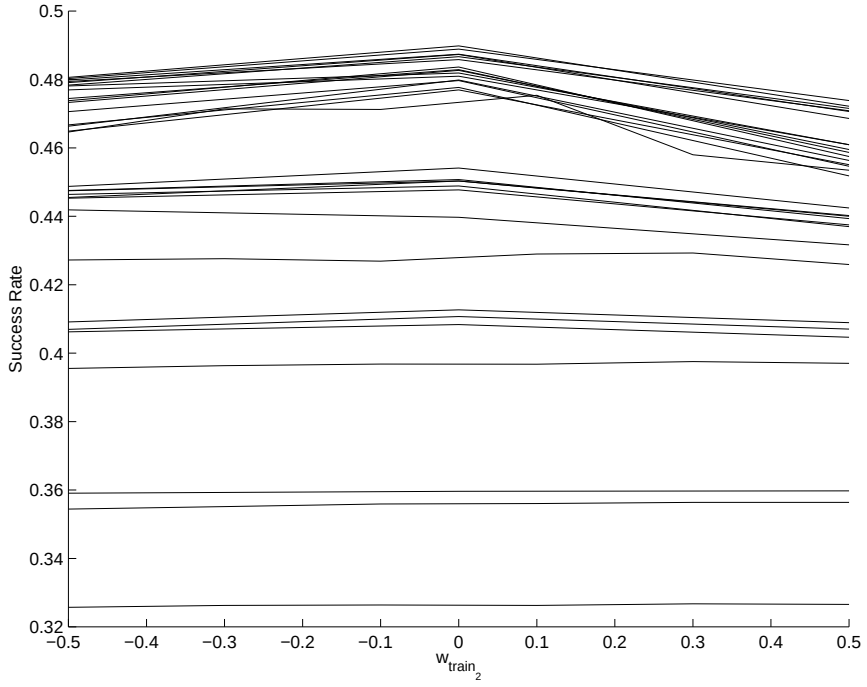

Figure 4-3: Success rate vs. normalized overlap w_{train_2} for different combinations of window sizes. Excluding data points that include the start of the next emotion segment ($w_{train_2} < 0.5$) in the training, slightly improves the results.

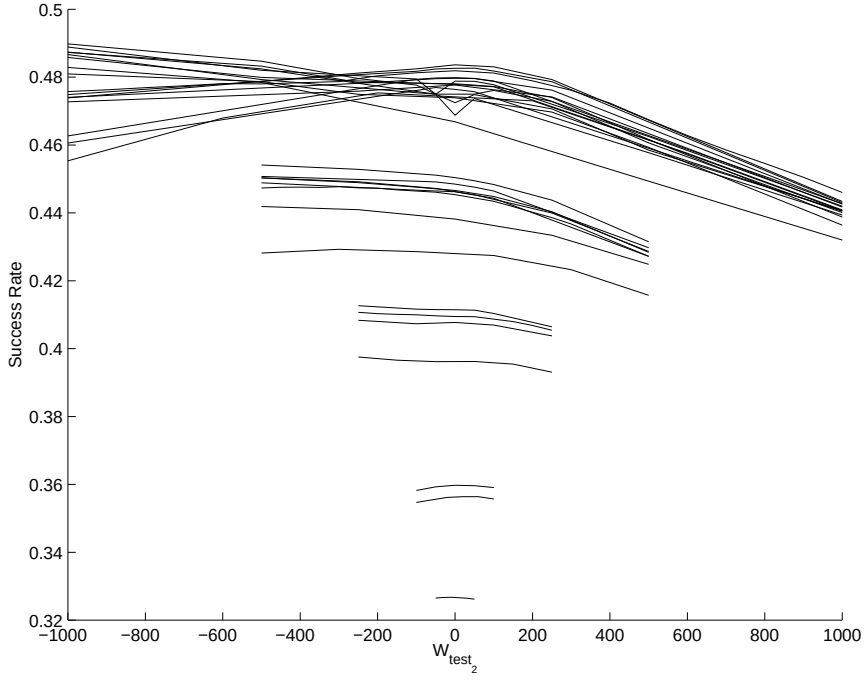

Figure 4-4: Success rate vs. unnormalized overlap W_{test_2} for different combinations of window sizes. Excluding data points from the end of an emotion segment (left part of the curves) in the testing, significantly improves the overall results.

the middle section of the emotions.

Excluding windows that include data points from the start of the next emotion segment ($w_{train_2} < 0.5$) in the training, slightly improves the results (Fig. 4-3). Similarly, excluding these data points from the testing significantly improves the overall results (Fig. 4-4). It seems that the data towards the end of each emotion segment is not helpful in the training of the classifier, and is not classified as well as the middle section of each emotion segment. We inquired with the actress who provided the data, and she indicated that trying to express a specific emotion steadily for 3 minutes often got boring; hence the data towards the end of each segment might not be as representative of the emotion as the earlier and middle portions of the segment.

The highest data point classification success rate was obtained when combining all 5 window sizes, and it was 48.98%. It should be noted that the segment classification success rate reached 60%, while the offline version using the same methods (Fisher Projection method, 30 features, without Day Matrix) gave a (segment classification) success rate of 56.87% (Table 3.3). Unfortunately, in most real-life applications, presegmented data will not be available.

4.8 Conclusions

Plenty of work needs to be done before a robust and easy-to-use emotion recognizer is built, but a first step was made, by looking into online emotion expression recognition and trying to solve some of the issues that arise. Results from the online classifier were very encouraging, comparable to the offline version's results using the same features and methods. They confirm and expand upon our earlier results, which suggested that there is significant information in physiological signals for classifying the affective state of a person who is deliberately expressing a small set of emotions.

Chapter 5

Cognitive Load: Pilot Study

5.1 Introduction

This chapter describes two sets of pilot studies on cognitive load under different driving conditions. Physiological signals were not gathered at this point. Five subjects were asked to drive a driving simulator past several curves. In parts of the experiment they were also asked to listen to random numbers from a speech-synthesis software and perform simple mathematical tasks. Several measures drawn from the subjects' driving behavior were examined as possible indicators of either the subjects' performance or their cognitive load. The studies were used as a guideline for a more thorough experiment which follows (6) and includes the recording of physiological data.

5.2 Motivation

Every day more people use their cellular phones while driving. This may have a serious impact on their driving performance, and subsequently on the safety of people on the road.

According to the Mental Workload model [20], performance drops sharply when a person enters the Mental Overload regime. Therefore, talking on a cellular phone may only be dangerous under some very demanding driving conditions and harmless

in most other routine driving cases, and the onset of the dangerous regime may be person- and situation-specific. We would like to be able to recognize such potentially dangerous situations so that they can be avoided. If we could somehow predict from a variety of measures that the driver is close to the onset of overload, we might be able to prevent it by, for example, temporarily turning the phone off or warning the driver to be more attentive to the road.

In the specific experiment, the measures used were extracted from real-time data of the subjects' driving patterns as recorded in the state of the car at each time step. They involve measures of performance and measures of workload. These may be useful in understanding that a person is close to his/her Mental Overload limit, either because of the level of difficulty of the task or alternatively because the person's effective Mental Overload limit is temporarily much lower (due to substance-abuse, fatigue, sleep deprivation, or the user's underlying emotional state). No physiological data have been recorded at this point, although such data are used in the second study, described in the following chapter.

5.3 First study

5.3.1 Experiment

The driving simulator used for the purposes of the experiment is part of Cambridge Basic Research, a laboratory of Nissan Research & Development, Inc. and it consists of a Nissan 240sx (Fig. 5-1), an SGI Octane rendering the virtual world and a PC recording the state of the car in real time. An Apple Macintosh was used for the speech synthesis. The five subjects who participated in the experiment were all male MIT students from various departments and backgrounds, and all had moderate driving experience. The path they were asked to follow was identical in all the runs and can be seen in Fig. 5-2. It consisted of 10 right turns, none of which was more than 180 degrees, and none of which had a radius of curvature less than 40 meters, as well as the equivalent 10 left turns. The turns were randomly placed along the length of the

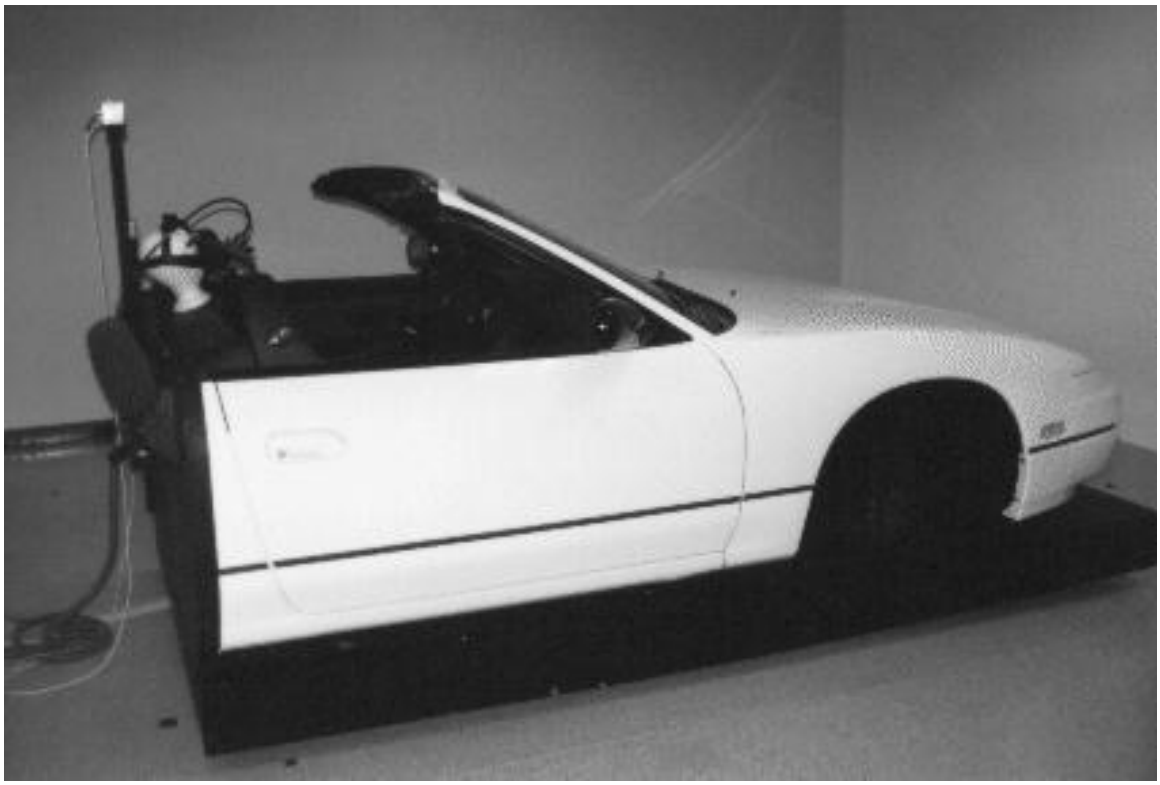

Figure 5-1: The Nissan 240sx driving simulator at Cambridge Basic Research.

path, with alternating right and left turns and small straight parts of various lengths in between. The fact that the finishing direction is parallel to the starting one is because each right turn has an equal and opposite left turn. Subjects said they did not learn the road, especially because the turns were not sudden or steep, so there was no need to anticipate them, although they all noticed the alternating right-left pattern.

Each subject first did a small number of trial runs, usually 2, during which he was also allowed to listen to a sample of the speech synthesizer saying some numbers. He then proceeded with the testing runs.

There were 9 different runs, a matrix of 3x3 different cases (Table 5.1), each 200 seconds long, in all of which the driver had to try to preserve a nominal speed as close to a preset constant value as possible. These values were 30, 40, and 50 miles per hour (MPH), although the perceived speeds reported were around 25%-50% slower. In 3

	No Addition	Easy Every 3	Hard Every 8
30	1	6	2
40	7	3	8
50	4	9	5

Table 5.1: The sequence of runs was constructed to minimize the effects of learning.

cases there was no associated phone task. In another 3 cases the subject had to answer simple mathematical questions: add "1" to a random single digit number ([0-9]), read by the speech-synthesizer every 3 seconds. In the remaining 3 cases the subject had to answer slightly harder mathematical questions: add two random non-negative numbers read by the speech-synthesizer every 8 seconds, whose sum was between 0 and 99. The interface for the sound/speech was a head-mounted speaker/microphone, resembling some new integrated car phones. The sequence by which these 9 runs were executed was identical for all subjects; this sequence is numbered in Table 5.1.

5.3.2 Analysis and results

The data which were available for analysis are the following:

- Car data: Every 50 ms a line is appended to the data file containing:
  - Elapsed time
  - 2-D car position
  - 2-D car velocity
  - Accelerator pedal depression
  - Brake pedal depression
- Speech data: Audio file of mathematical questions and subjects' responses

From the above, the following values can be calculated for each run and for each subject:

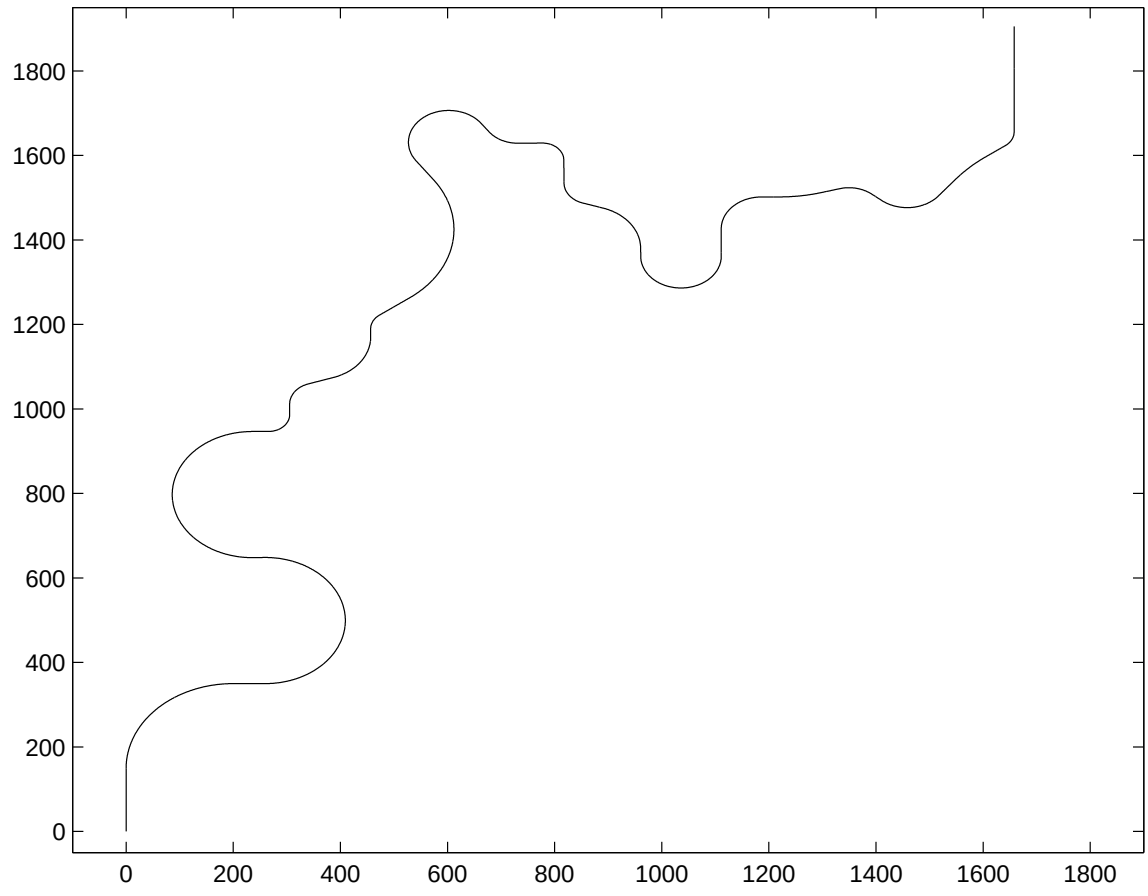

Figure 5-2: The path to be followed by the subjects had 10 right turns, 10 symmetric left turns, and a total length of 4.5 kilometers. The full length was only used in the high speed runs.

- Steering angle prediction error variance and entropy H_s [1, 13] : For each time step $t(i)$, a quadratic function is fitted to the steering angles

$$\theta(i-1), \theta(i-2), \theta(i-3)$$

of the last 3 time steps

$$t(i-1), t(i-2), t(i-3)$$

respectively. Then, using extrapolation, the angle

$$\tilde{\theta}(i)$$

is calculated as an estimate of $\theta(i)$ and the steering angle prediction error:

$$e(i) = \theta(i) - \tilde{\theta}(i)$$

is calculated. The error probability distribution is estimated by using a histogram, and then the entropy is calculated by:

$$H_s = \sum_{j=1}^{N} (p(j) * \log_N(p(j)))$$

where N is the number of equally spaced bins used in the histogram and p is the number of samples in bin j normalized by the total number of samples.

- Accelerator pedal variance and entropy: The accelerator pedal entropy is calculated using the same technique as above but with the prediction being a constant, the mean accelerator pedal depression of the whole session.

- Speed variance and entropy: The speed entropy is calculated using the same technique followed for the accelerator pedal depression.

- Mean lane deviation: Mean distance from the center of the street.

	Easy Addition	Hard Addition
30	2/1000	4/375
40	1/1000	4/375
50	0/1000	6/375

Table 5.2: Total number of errors over the total number of trials for each different test case.

- Normalized mean lane deviation: Same as above, except for the subtraction of the subject-specific mean from each subject's results, to account for any trend of driving on the right side of the street.
- Lane deviation variance and entropy
- Mean distance and time to lane crossing: The distance and time the car can travel before exiting the current lane, if it keeps the current heading (i.e. straight ahead) and speed.
- Speech task errors

All of these measures drawn from the subjects' driving behavior were examined as possible indicators of the subjects' cognitive load and performance. No monotonic behavior was observed with increasing difficulty of either task, and no behavior was consistent between different subjects.

5.3.3 Remarks

We can think of human resources as a *bucket* of limited capacity, with the higher priority tasks using the lower available resources. Then, if too many or too demanding tasks need to be performed, the person will enter the so-called Mental Overload regime (the bucket is going to overflow) with the tasks of perceived lowest priority being ignored first.

The apparent lack of correlation between the two tasks in the specific study may be an indication that the tasks were too easy. Figure 5-3 shows the model that

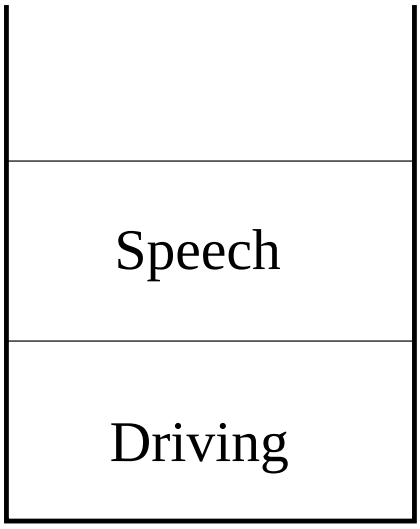

Figure 5-3: Mental Workload model for first pilot study. Driving is considered to be a higher priority task than the speech. No overload occurs, as the tasks are very easy.

probably describes this situation. Several subjects said that after they stabilized the car to a speed close to the desired, they would lay their foot against the side and thus keep the speed constant without any further effort (Fig. 5-4). The sound of the engine gave an extra cue. Also, the path was made out of constant curvature turns, so after subjects found the correct steering angle they just had to keep the wheel almost steady, rather than continuously introduce corrections. These are very unrealistic situations, because in real life the presence of vibrations from the road, wind, and inclines, among other things, make keeping a constant speed and driving through curves much harder of a task. As mentioned above, a common remark was that the car "felt" like it was going slower than the nominal speed, with the perceived speed reported usually being around 50%-75% of the nominal one.

The analysis was further complicated by the presence of multiple relatively uncorrelated measures of workload such as the ones related to the steering wheel (e.g. steering entropy) and the ones related to the accelerator pedal (e.g. accelerator entropy). Similarly, there were multiple relatively uncorrelated measures of performance, such as the ones related to the position of the car (e.g. lane deviation, distance to lane crossing) and the car speed (e.g. speed variance). It cannot be known exactly how

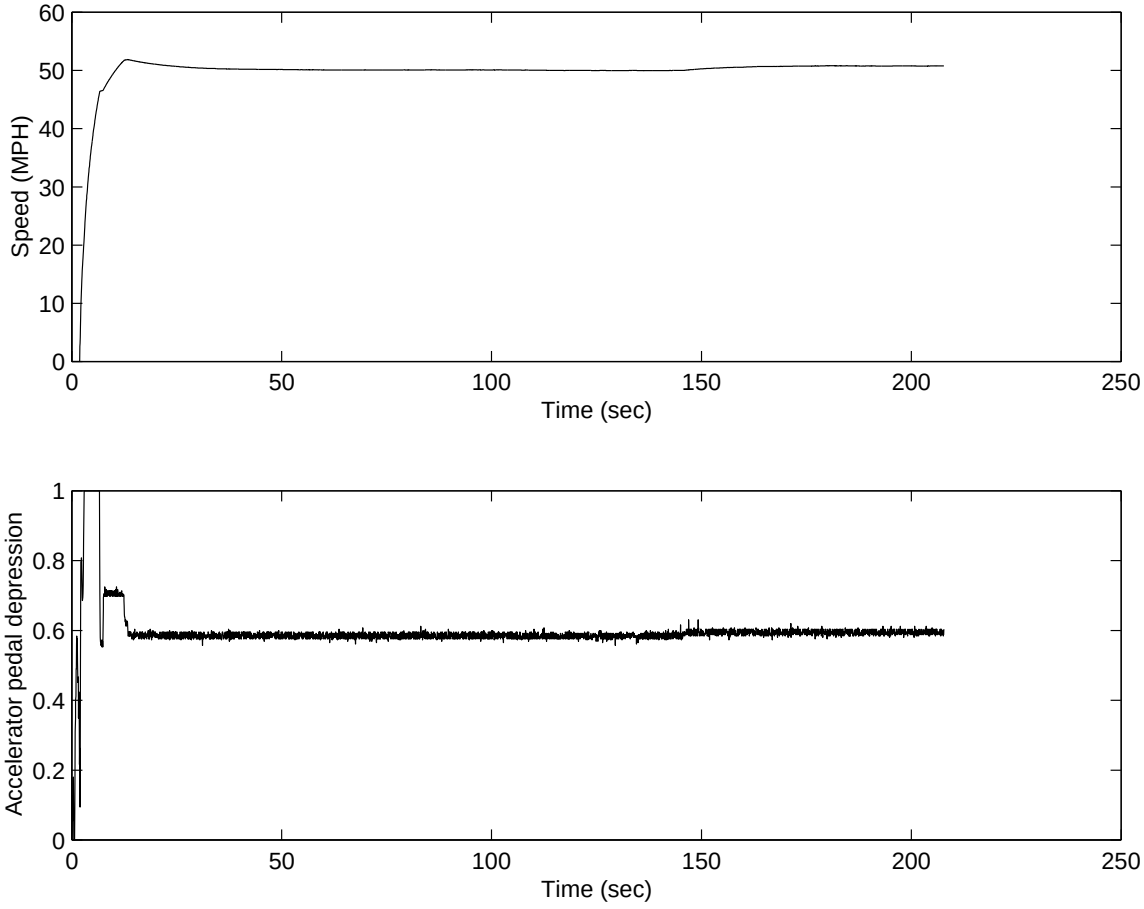

Figure 5-4: Example car speed and accelerator pedal depression vs. time.

these unrelated measures interact, making it harder to estimate "global" measures of workload and performance. It was noticed, for example, that at least one person clearly favored keeping a constant speed around steep curves, at the expense of going slightly off-course, while another favored staying in the middle of the lane, by decelerating in advance.

Finally, with respect to the speech tasks, although adding two numbers whose sum is less than 100 is for most people more difficult than adding 1 to a given digit, it is not clear whether doing the former every 8 seconds is much harder than doing the latter every 3 seconds.

We try to make use of these conclusions in building the next set of experiments.

5.4 Second study

5.4.1 Experiment

In the second pilot study, several modifications were introduced. First, one of the phone tasks was eliminated. The subjects were now asked only to add two numbers whose sum was between 0 and 99, in intervals of 3, 5, 7, or 9 seconds. Instead of asking the driver to keep a desired speed, nominal speeds of 30, 60, 90, and 120 MPH were imposed. To minimize road predictability and constant curvature turns, a new path was created as a sum of 4 sinusoids of different amplitudes and frequencies, which can be seen in Fig. 5-5. The specific sequence of runs is displayed in Table 5.3.

Unfortunately the first and only subject to participate in this study had to stop after the first 9 runs, after feeling sick. Nevertheless, results from these runs contributed to further modifications.

5.4.2 Analysis and results

The analysis which followed was similar to the one in the first study. This time several driving performance measures were found to be deteriorating with higher speeds. Also the speech performance deteriorated as intervals between questions got

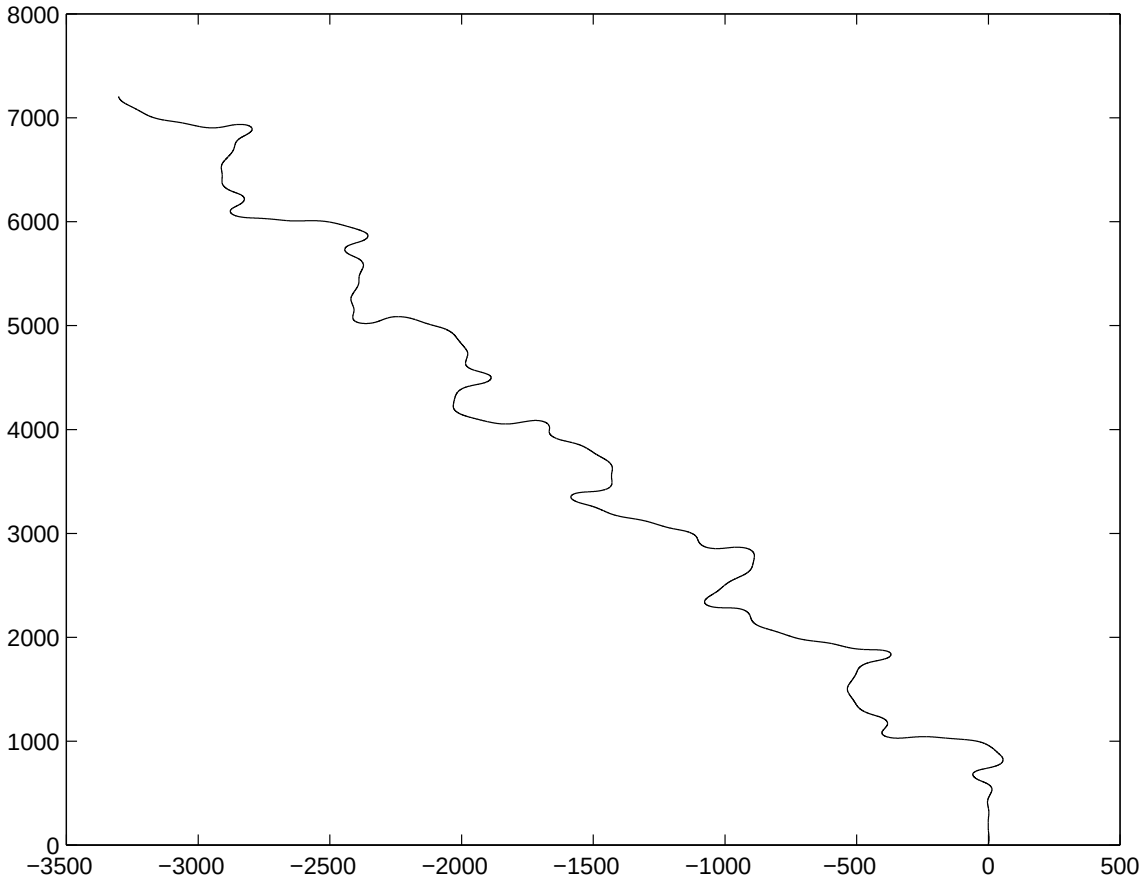

Figure 5-5: The path to be followed by the subjects was a sum of 4 sinusoids of different amplitudes and frequencies, and a total length of 15 kilometers. The full length was only used in the high speed runs.

	No Addition	Addition Every 9	Addition Every 7	Addition Every 5	Addition Every 3
30	1	9	17	5	13
60	8	16	4	12	20
90	15	3	11	19	7
120	2	10	18	6	14

Table 5.3: The sequence of runs was constructed to minimize the effects of learning.

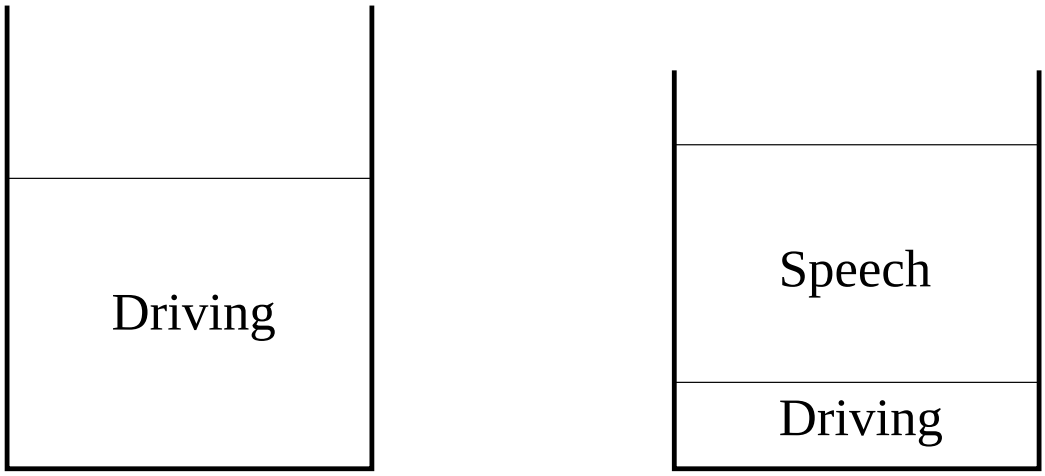

Figure 5-6: Mental Workload model for second pilot study. The assumption is that driving needs a lot of low level resources (on the left), and only a few high level congnitive resources (on the right). The speech task needs a lot of high level cognitive resources. Sometimes overload occurs, almost independently for the two tasks, when the needed resources of one sort exceed the height of the respective bucket.

smaller. Nevertheless, still there appeared to be no correlation between the speech task and driving performance, or speed and the speech performance.

5.4.3 Remarks

This time each task individually was hard enough to produce a deterioration in this task's performance. Nevertheless, it seems the two tasks share very few common resources, as if people use different parts of their brains. This may be because the continuous driving task can be mostly performed "in the back of our minds", while our highly cognitive resources get dedicated to the speech task. Figure 5-6 shows the model that probably describes this situation. Our model for the previous study should also be updated with the existence of multiple kinds of resources, rather than just one. This raises the following issue: if another event requiring cognitive resources suddenly appears, will talking on the phone make it harder to deal with? In the next chapter we set up such an experiment where events suddenly demand the driver's cognitive resources. We then use the driver's physiological data to predict the driver's ability (or inability) to cope with such a challenge.

Chapter 6

Cognitive Load and Physiological Data

6.1 Experiment

The study presented here is very similar to the second pilot study described in the previous chapter: the path was again generated as a sum of sinusoids, the speed was imposed, the math questions were of the same type. The most significant change was that messages appeared on the screen which prompted the subject to either *Continue driving at same speed* or to *Brake immediately to 0 mph.* In order for the subject to be able to use the brakes, the gas pedal was turned into a switch: when it was pressed, the predetermined speed was instantly imposed, while when it was released the car was allowed to slow down because of friction, wind resistance, or braking. There were 4 different runs, all combinations of 2 different speeds and 2 different intervals between math questions. In some occasions, when subjects had more time, another 4 runs followed. Each run was 6 minutes long, with no math questions during the first and last minutes. Messages appeared every 10 seconds in a random sequence which guaranteed 3 *Brake* and 3 *Continue* messages in the first minute, 12 from each during the 4 minutes the subject was on the phone, and 3 each again in the last minute. The specific sequence of runs and the setup of each run are displayed in Tables 6.1 and 6.2. Finally, the sum of the math questions now was constrained to be between

	Addition Every 3	Addition Every 9
60	1,8	3,6
120	4,5	2,7

	Addition Every 3	Addition Every 9
60	2,7	4,5
120	3,6	1,8

Table 6.1: The sequences of runs were constructed to minimize the effects of learning. Some subjects were given the first sequence, others were given the second one.

Duration (minutes)	1	4	1
Speech Task	No	Yes	No
Brake events	3	12	3
Continue events	3	12	3

Table 6.2: Each run consisted of 3 parts.

11 and 99, excluding the really easy single digit summations, and the number of summations with carry-over was controlled to be exactly half of the total number of questions. The only information withheld from the subjects was (1) that the path was always the same, (2) the path itself, (3) the timing of the messages on the screen, (4) that there was an equal number of instances of the two messages, and (5) that we were controlling for the number of cases with a carry over (See Appendix). These would increase the effect of learning and create further secondary tasks, with people trying to memorize the road or to predict when the next message would appear. In some cases it was necessary to decrease the rate of questions from every 3 to every 4 seconds, and/or the speed from 120 MPH to 100 MPH, when subjects said they found it impossible to perform the tasks. More curtains were used, to block a larger amount of light from the driving simulator room than in the previous studies, in case this reduced the chances of subjects getting sick, but another two subjects still had to interrupt the experiment.

The messages prompting the driver to brake or to continue driving were used to examine the driving performance on discrete tasks of a person talking on the phone. The messages were almost exactly the same length, to make the subject devote some cognitive resources before deciding on some action. This is because in real life one has more than one possible action (press brakes, steer left etc.) when such discrete events occur, so one has to assess the situation before acting. The setup with a first and last minute without math questions was included to give a baseline performance and baseline physiological data. Finally, the control for carry-overs in the math was included after several subjects mentioned that these are the hardest and it was clear from preliminary results that they were responsible for the vast majority of the summation mistakes.

6.2 Driving results

The most important new performance measure is the driver's reaction time. This is the time from the moment a *Brake* message appears on the screen until the moment the subject presses the brake pedal. A typical plot of response delay with and without speech task can be seen in Figure 6-1. It turns out that the same pattern appeared consistently across all subjects: although the majority of delays lay between 0.7 and 1.4 seconds (irrespective of the phone task), there were a few cases in which subjects pressed the brakes significantly later (1.5-3.5 seconds). These latter cases occurred almost entirely while subjects were performing the speech task. The complete set of delays for all subjects can be seen in Figure 6-2. Each subject's delays are normalized by the mean delay in the absence of the speech task.

Further analysis revealed that out of a total of 972 times a *Brake* message appeared, there were 7 cases where the driver never pressed the brakes. In 6 of these 7 cases the driver was talking on the phone. Finally, out of a total of 972 times a *Continue* message appeared, there were 10 cases where the driver mistakenly pressed the brakes. In 9 of these 10 cases the driver was talking on the phone.

If reaction times consistently increased by a small amount in the presence of sec-

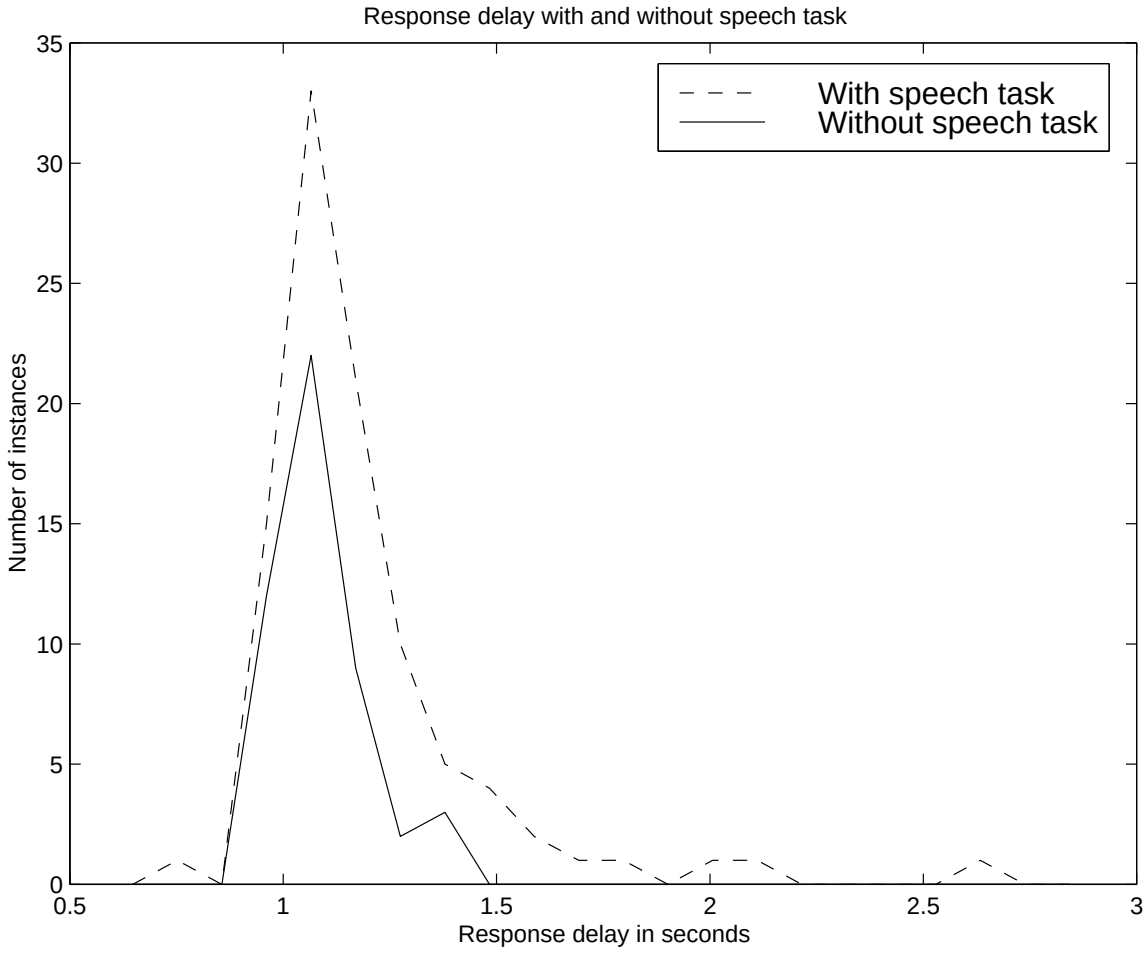

Figure 6-1: A typical plot of response delay with and without speech task.

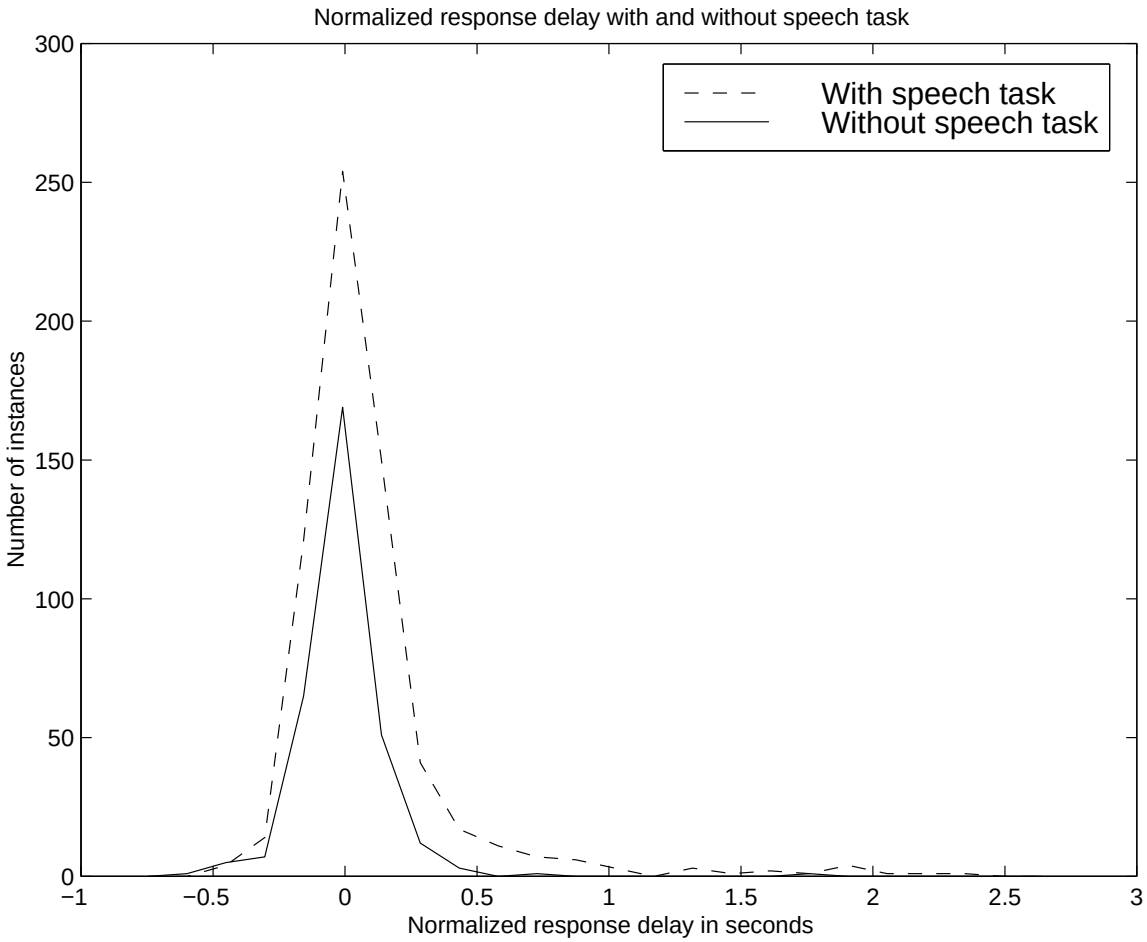

Figure 6-2: Response delay for 10 subjects. Each subject's delays are normalized by the mean delay in the absence of the speech task.

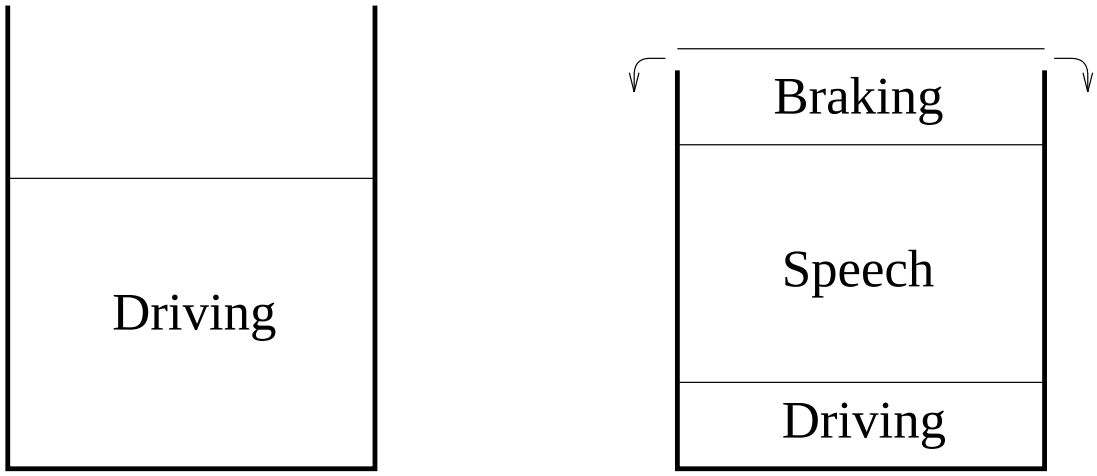

Figure 6-3: Mental Workload model for study. Overload occurs often when subjects are asked to brake while talking on the phone. As only the lowest priority resources are available then, braking is the one to overflow when demand exceeds resources.

ondary tasks, people could compensate by taking greater precautions. Unfortunately, the specific study shows that drivers' braking delays generally tend to be comparable (0.7-1.5 seconds) with and without the speech task, except for a few times when the delays are *significantly* longer. Furthermore, while on the phone, drivers have a higher tendency to mistakenly press the brakes after a *Continue* message or completely ignore a *Brake* message while they are on the phone. Because the need to take immediate action is infrequent and most of the times drivers seem to react appropriately, a false sense of safety is created, further aggravating the problem. Figure 6-3 shows the model that probably describes this situation.

6.3 Physiological data

Unfortunately physiological data are available for only 1 out of the 10 subjects whose driving performance was reported above. The first subjects were used only as pilots, until the pattern in the response delay was established (so no physiology was recorded). During some experiments, there were severe problems with a computer undersampling or failing to save the data, or a sensor failing during the experiment.

For two subjects the presence of only 5 cases of high delay (out of 72) would have made it all but impossible to train any algorithm into recognizing these cases.

6.3.1 Analysis

A classifier based on physiology could ideally help identify when a subject is likely to have a significantly delayed response to a low-probability discrete event. The procedure followed in building the classifier is almost identical to the one used in the earlier chapters for emotion recognition:

- Signals: EMG, BVP, GSR, Respiration, HR, each sampled at 20 samples a second.
- Data: A window of size 1 to 5 seconds immediately preceding the appearance of a *Brake* message on the screen.
- Features: Same 6 features/signal as in Section 2.2.
- Classes: All *normal* responses (delays of 0.7-1.5 seconds) define Class 1 (55 data points); all *slow* responses (delays of more than 1.5 seconds) define Class 2 (17 data points);
- Recognition: Fisher Projection (2.3.2) with leave-one-out (2.4.1).

6.3.2 Results

There are two types of error in the classification process. The first is the mistaken classification of a *normal* response as *slow* (a false positive, also known as false alarm). The second is the mistaken classification of a *slow* response as *normal* (a false negative). In most cases these two are not equally important. We may be willing to take a lot of false positives (also called *false alarms*) in order to avoid false negatives, because of potentially dangerous situations arising from the latter. Following the convention in [18], the Receiver-Operator Curve (ROC) for the system can be seen in Fig. 6-4. We see that the point where the two errors are equal is at about 65% recognition rate.

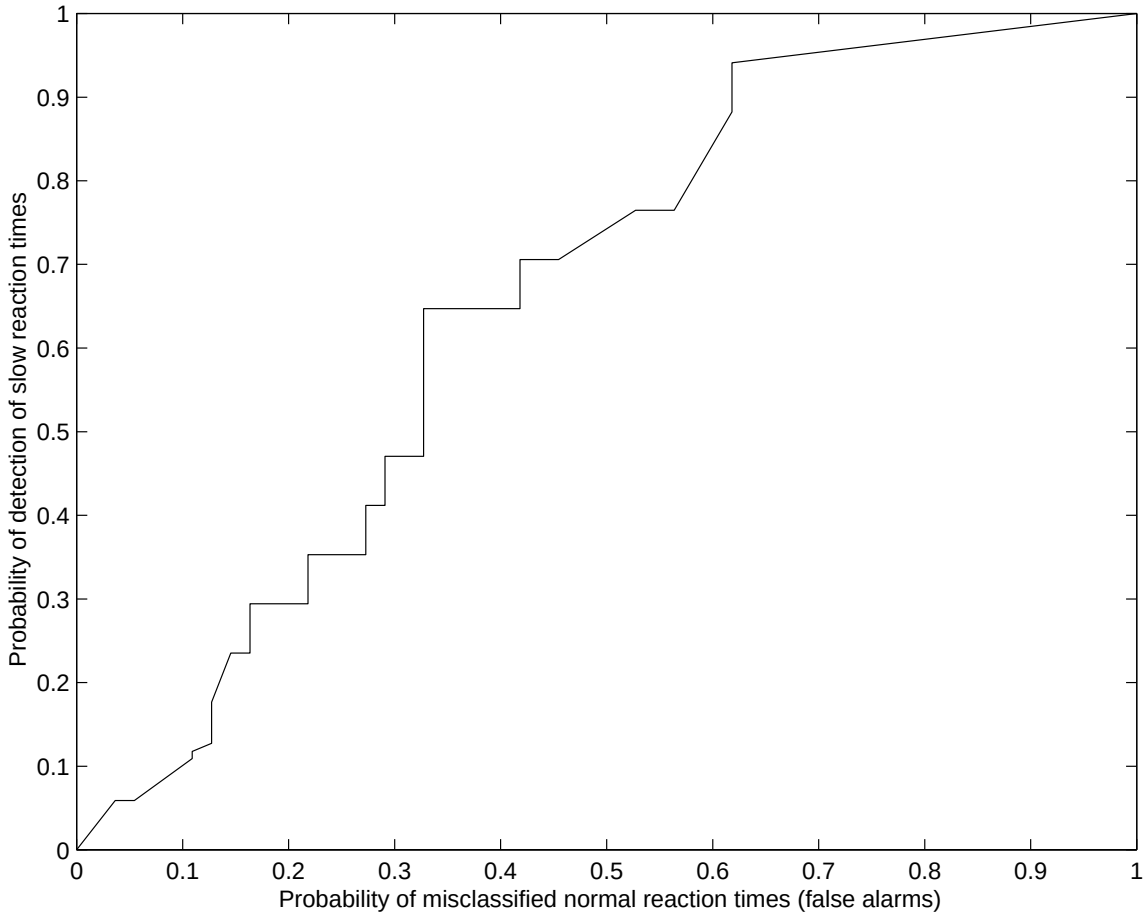

Figure 6-4: Receiver-Operator Curve for the delay classification.

6.4 Conclusions

In this chapter we showed that while performing some very demanding secondary tasks, driving performance with respect to discrete events drops with potentially very dangerous consequences. Furthermore, we showed that it may be possible to use physiological data from the drivers to predict if their reaction to an event in the near future would be "normal" or "slow". This can help in the prevention of accidents where attentiveness is crucial, like in driving, or in operating heavy machinery.

Chapter 7

Conclusions

The application of several pattern recognition techniques on physiological data was shown to provide useful information about some cases of human emotional or cognitive states.

Part of this thesis addressed the recognition, through physiology, of emotional states during deliberate emotional expression by an actress. The thesis suggested signals, features, and pattern recognition techniques for offline recognition of all 8 emotions examined and presented results suggesting that emotions can be recognized from physiological signals at significantly higher than chance probabilities.

Emotion recognition can be very useful if it occurs in real time. That is, we would like the computer to be able to sense the emotional state of the user the moment he actually is in this state (online recognition), rather than analyzing the data later, when the user is already in another state (offline recognition). The thesis showed an online algorithm is not only possible but it can reach classification rates comparable to the ones of the offline version. It suggested the use of iterative updates for several features and made it run much faster than the current sampling rate of the physiological sensors that were available for the experiments. Finally it pointed out problems and limitations: if we have no knowledge of the time scale of the duration of an emotion, it is very hard to choose a window size; if the data is not presegmented the performance drops significantly.

The second part of the thesis involved the study of cognitive load and performance

under different driving conditions. Verifying the Mental Workload model [20], it was displayed that different tasks can act cumulatively on a subject's cognitive load, with possibly detrimental effects. Furthermore, again verifying previous findings, we showed that in some cases there are multiple kinds of workload, which may or may not interact significantly. The physiology of subjects driving in a driving simulator was recorded, and used for one subject to successfully predict slow response times.

The above findings suggest that by using physiology among other things we may be able to better predict the emotional state or the cognitive load of a car driver, a heavy machinery operator, a computer user, and improve safety, performance and well-being of subjects.

Appendix

These are the instructions given to the subjects in the final driving study (Chapter 6):

Welcome,

Please make sure to read and sign our consent form. The experiment you will be asked to perform has the following format: There are 8 runs of 6 minutes each. In some of the runs you will be driving at 60 miles per hour, and in some at 120 miles per hour. These speeds are perceived as much slower than that, around half their nominal values. Your accelerator pedal operates as a switch, so as long as you press it, it imposes the predetermined speed.

At random points during these runs, a message will come up on the screen, asking you to brake immediately to 0 miles per hour, in which case you are expected to press the brake pedal until the car stops and then press the accellerator pedal again. In other occasions, a message will come up asking you to continue driving at the same speed, in which case you are expected to continue pressing the accelerator pedal.

In addition, during the middle 4 minutes of each run, you will listen to pairs of numbers on the headphone and you will be expected to say their sum to the microphone. These will either come at a rate of one every 9 seconds, or at a rate of one every 3 seconds.

Before we start, we need you to perform 2 trial runs in order to get used to the

setup, one at 60 and one at 120 miles per hour.

If you have any questions, do not hesitate to ask the experimenter. We can take a break half way through the experiment. Finally, if at any time during the experiment you feel sick, inform the experimenter immediately.

Thank you for participating in this experiment.

Bibliography

[1] E. R. Boer and A. Liu (Eds.). Cambridge basic research 1997 annual report. Technical Report CBR TR 97-7, Cambridge Basic Research, Cambridge, MA, 1997.

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

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

[4] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE T. Info. Theory*, IT-13(1):21–27, Jan. 1967.

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

[6] Irfan Essa and Alex Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.

[7] D. Goleman. *Emotional Intelligence.* Bantam Books, New York, 1995.

[8] J. Healey and R. W. Picard. Digital processing of affective signals. In *IEEE Int. Conf. on Acoust., Sp., and Sig. Proc.*, Seattle, 1998.

[9] D. R. Heise. Affect control theory: Concepts and model. *Journal of Mathematical Sociology*, 13(1-2):1–33, January-February 1987.

[10] A. K. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.

[11] J. Kittler. Feature set search algorithms. In C. H. Chen, editor, *Pattern Recognition and Signal Processing*, pages 41–60. Alphen aan den Rijn: Sijthoff & Noordhoff, Netherlands, 1978.

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

[13] O. Nakayama, T. Futami, T. Nakamura, and E. R. Boer. Development of a steering entropy method for evaluating driver workload. In *Society of Automotive Engineers*, 1998.

[14] R. W. Picard. *Affective Computing.* The MIT Press, Cambridge, MA, 1997.

[15] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, November 1994.

[16] K. R. Scherer. Ch. 10: Speech and emotional states. In J. K. Darby, editor, *Speech Evaluation in Psychiatry*, pages 189–220. Grune and Stratton, Inc., 1981.

[17] L. Smith-Lovin. Affect control theory: An assessment. *Journal of Mathematical Sociology*, 13(1-2):171–192, January-February 1987.

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

[19] Elias Vyzas and Rosalind W. Picard. Affective pattern classification, AAAI 1998 fall symposium, emotional and intelligent: The tangled knot of cognition. In *AAAI*, Orlando, FL, Oct. 1998.

[20] C. D. Wickens, S. E. Gordon, and Y. Liu. *An Introduction to Human Factors Engineering*, chapter Stress and Workload, pages 377–408. Longman, 1998.

[21] S. Wilks. *Mathematical Statistics*, pages 577–578. John Wiley, New York, 1962.

[22] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from log image sequences using optical flow. *IEEE T. Patt. Analy. and Mach. Intell.*, 18(6):636–642, June 1996.