Short-term emotion assessment in a recall paradigm

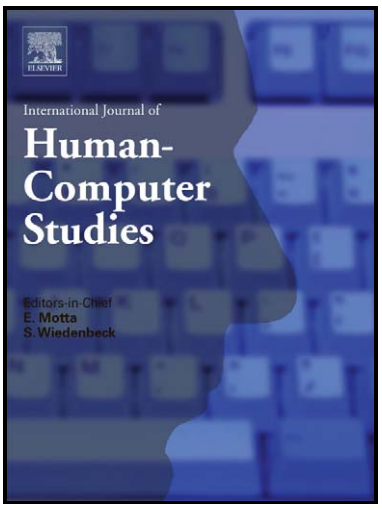
Chanel, Guillaume; Kierkels, Joep Johannes Maria; Soleymani, Mohammad; Pun, Thierry

How to cite

CHANEL, Guillaume et al. Short-term emotion assessment in a recall paradigm. In: International journal of human-computer studies, 2009, vol. 67, n° 8, p. 607–627. doi: 10.1016/j.ijhcs.2009.03.005

This publication URL: https://archive-ouverte.unige.ch/unige:47415
Publication DOI: 10.1016/j.ijhcs.2009.03.005

Author's Accepted Manuscript



www.elsevier.com/locate/ijhcs

Short-term emotion assessment in a recall paradigm

Guillaume Chanel, Joep Kierkels, Mohammad Soleymani, Thierry Pun

PII: S1071-5819(09)00043-3
DOI: doi:10.1016/j.ijhcs.2009.03.005
Reference: YIJHC 1519

To appear in: *Int. J. Human–Computer Studies*

Received date: 4 July 2008
Revised date: 19 March 2009
Accepted date: 27 March 2009

Cite this article as: Guillaume Chanel, Joep Kierkels, Mohammad Soleymani and Thierry Pun, Short-term emotion assessment in a recall paradigm, *Int. J. Human–Computer Studies* (2009), doi:10.1016/j.ijhcs.2009.03.005

Short-term emotion assessment in a recall paradigm

Guillaume Chanel*, Joep Kierkels, Mohammad Soleymani, Thierry Pun
Computer Science Departement – University of Geneva
Route de Drize 7
CH - 1227 Carouge, Switzerland

Abstract

The work presented in this paper aims at assessing human emotions using peripheral as well as electroencephalographic (EEG) physiological signals on short-time periods. Three specific areas of the valence-arousal emotional space are defined, corresponding to negatively excited, positively excited, and calm-neutral states. An acquisition protocol based on the recall of past emotional life episodes has been designed to acquire data from both peripheral and EEG signals. Pattern classification is used to distinguish between the three areas of the valence-arousal space. The performance of several classifiers has been evaluated on ten participants and different feature sets: peripheral features, EEG time-frequency features, EEG pairwise mutual information features. Comparison of results obtained using either peripheral or EEG signals confirms the interest of using EEG's to assess valence and arousal in emotion recall conditions. The obtained accuracy for the three emotional classes is 63% using EEG time-frequency features which is better than the results obtained from previous studies using EEG and similar classes. Fusion of the different feature sets at the decision level using a summation rule also showed to improve accuracy to 70%. Furthermore, the rejection of non confident samples finally led to a classification accuracy of 80% for the three classes.

Keywords: emotion assessment and classification, affective computing, signal processing.

1 Introduction

Emotions are part of any natural communication between humans, generally as non-verbal cues. Until recently affective communication was not implemented in human computer interfaces. Nowadays researchers in human-computer interaction (HCI) have recognized the importance of emotional aspects and started to include them in the design of new interfaces using one of two possible approaches, either as evaluation indicators or as components to be inserted in the human-computer loop. The first approach consist in using emotion assessment as a tool for evaluating attractiveness, appreciation and user experience of software (Hazlett and Benedek, 2007). Such an assessment can be done by using different self-report methods (Isomursu et al., 2007) or by inferring emotional states from others measures such as physiological signals (Mandryk et al., 2006; Picard and Daily, 2005).

The second approach aims at bringing the machine closer to the human by including emotional content in the communication and is known as affective computing (Picard, 1997). According to Picard, affective computing "proposes to give computers the ability to recognize [and] express […] emotions". Synthetic expression of emotions

* Corresponding author. Tel.: (+41) 22 379 01 83; fax: (+41) 22 379 0250.
E-mail address: guillaume.chanel@unige.ch.

can be achieved by enabling avatars or simpler agents to have facial expressions, different tones of voice, and empathic behaviours (Brave et al., 2005; Xu et al., 2006). Detection of human emotions can be realized by monitoring facial expressions (Cohen et al., 2003; Cowie et al., 2001; Fasel and Luettin, 2003), speech (Cowie, 2000; Ververidis and Kotropoulos, 2006) , postures (Coulson, 2004; Kapoor et al., 2007) and physiological signals (see Section 2). Fusion of these different modalities to improve the recognition accuracy has also been studied (Kapoor et al., 2007; Kim et al., 2005; Pantic and Rothkrantz, 2003; Zeng et al., 2008).The present work focuses on the emotion assessment aspect, especially from different physiological signals.

Fig. 1 presents a framework describing how emotion assessment could be integrated in human-computer interfaces. As proposed by Norman (1990) the interaction with a machine, from the point of view of the user, can be decomposed in execution / evaluation cycles. After identifying his/her goals, the user starts an execution stage. It consists in formulating his intentions, specifying the necessary sequence of actions and executing those actions. Next, the computer executes the given commands and output results through the available modalities. The second stage is the evaluation which is realized by: perceiving computer outputs, interpreting them and evaluating the outcome (*i.e.* are the goals satisfied?).

<Figure 1>

According to the cognitive theory of emotions, emotions are issued from a cognitive process called appraisal that evaluates a stimulus according to several criteria such as goal relevance and consequences of the event (Cornelius, 1996; Sander et al., 2005; Scherer, 2001). For this reason, an emotional evaluation step, corresponding to the appraisal process, was added in Fig. 1 at the evaluation stage. Elicitation of emotions is known to be related to changes in several components of the organism such as physiological, motor and behavioral components (Scherer, 2001). It is thus possible to consider those changes as emotional cues that can be used to automatically detect the elicited emotion after being recorded by the adequate sensors. For this purpose it is necessary to extract features of interest from the recorded signals and perform classification using previously trained models from data including the associated emotion. The detected emotion can then be used to adapt the interaction by modifying command execution. The information presented on the output modalities can also directly be influenced by the emotional adaptation, for instance by synthesizing an emotional response on screens and speakers.

Several applications can be derived from this framework, some of them going beyond human-computer interfaces to reach human-machine interfaces in general and even human-human interfaces. In gaming, emotion assessment can be used for better understanding of the playing conditions that lead to emotional activation (Mandryk and Atkins, 2007) and for maintaining involvement of a player by adapting game difficulty or content (Chanel et al., 2008; Chen, 2007; Rani et al., 2005). Learning in a computer mediated environment can elicit various types of feeling that are currently not handled. Detection of frustration for instance, if followed by proper adaptation of the learning strategy (Choi et al., 2007; Kapoor et al., 2007), will certainly help to maintain learner's interest. Identification of critical states, such as stress, panic, and boredom, can be very useful in situations such as driving or when performing dangerous operations (Benoit et al., 2006; Healey, 2000). Another possible

application is the use of emotion recognition to help severely disabled persons express their feeling.

In addition to the cognitive theory, several theories of emotions have developed over the past century (Cornelius, 1996). These different views gave rise to different models of emotions. The most famous models are the basic emotions (Ekman et al., 1987) and the valence-arousal space (Russell, 1980). Basic emotions are defined as the emotions that are common across cultures and selected by nature because of their high survival functions. The valence-arousal space allows for a continuous representation of emotions on two axes: valence, ranging from unpleasant to pleasant, and arousal, ranging from calm to excited.

As stated before, there is extensive literature regarding emotion assessment from speech and facial expressions. However these two modalities suffer from several disadvantages. Firstly, they are not always available: the user may not look at the camera or speak all the time. Secondly, they do not always reflect the true emotional state of the user, since facial and voice expressions are often faked because of social rules or by deliberate choice. Finally, they should be regarded as the output of the emotional process, meaning that the emotion could have been elicited prior to the speech or expression.

An alternative is to use physiological signals both from the central and peripheral system. The Jamesian theory (Cornelius, 1996) emphasizes the importance of peripheral signals as it suggests there is some specific pattern of physiology for each particular emotion. Even though this statement is often disputed in the psychology literature (Stemmler et al., 2001) and the cognitive theory suggest that physiological signals are also an output of the emotional process, several studies from human computer interaction have shown the usefulness of peripheral activity for emotion assessment in diverse conditions (see Section 2.1). Those studies mainly use signals duration of the order of the minute to obtain accurate emotion assessment, however shorter durations are necessary for real time applications. The cognitive theory stresses the importance of the central nervous system i.e. the brain. It has been shown that correlates of emotions can be observed from brain activity, especially in the pre-frontal cortex and the amygdala (Adolphs et al., 2003; Damasio et al., 2000; Davidson, 2003; Davidson et al., 2000; Rolls, 2000). Brain activity can be measured using several techniques such as Electroencephalography (EEG) and others. Since it is the least intrusive and comparatively cheap, EEG is certainly one of the most usable modality for real and everyday applications. Brain activity can also be useful for short term emotion assessment since activity related to an emotional stimulus should occur shortly after the stimulus onset. Despite of this, the few studies trying to use EEG signals for emotion assessment did not obtain convincing results.

The objective of this study is to investigate the use of EEG modality, peripheral modality and fusion of these two for emotion assessment on short time periods. This work will thus focus on the red parts of Fig. 1. Previous studies related to emotion assessment from physiological signals and novelties of the present study are detailed in Section 2. A protocol based on the recall of past emotional episodes was used to reliably induce emotions from three areas of the valence-arousal space. This protocol was used instead of a real HCI framework because it allows for better control over when an emotional episode starts (timing) and over the strength and valence of the recalled emotion (content). The protocol, the signal acquisition system and the algorithms developed to extract different feature sets from the recorded signals are

presented in Section 3. To estimate underlying emotions from the computed feature sets, various pattern recognition techniques were applied (Section 4). Finally results are discussed in Section 5.

2 Emotion assessment from physiological signals

2.1 State of the art

Over the last years, emotion recognition from physiological signals has received much interest. Table 1 provides a (non exhaustive) list of relevant studies. Unfortunately, it is difficult to make comparisons between these studies because they differ on several criterions. Six criteria are introduced below to help the reader gain insight into the current state of the art as well as to discuss important aspects of emotion assessment from physiological signals.

<Table 1>

I. **Number of participants:** in a study that includes a high number of participants the results can be regarded as more significant. Another point of importance is to know if the model obtained for emotion identification is user-specific or not. In the first case, a new model will have to be generated for each new user.

II. **Emotion elicitation:** Picard et al. (2001) divided emotion elicitation approaches in two categories: subject-elicited and event-elicited. In the first category, emotions can be generated by asking the participant to act as if he/she was feeling a particular emotion (for instance by mimicking the facial expression of anger) or to remember past emotional episodes of his/her life. This method has been often used in facial expression recognition and it has been shown by Ekman et al. (1983) that this method is effective to induce specific peripheral activity. In the second category, it is possible to use images, sounds, video clips or any emotionally evocative stimuli. There already exist stimuli databases that have been designed for the purpose of emotion elicitation such as the International Affective Picture or Digitized Sound System (IAPS, IADS)(Lang et al., 2005). These databases are generally accompanied by affective evaluations from experts or average judgments of several people. However, Chanel et al. (2006) found that for the purpose of emotion assessment, even in the presence of predefined evaluation labels, it is important to use label based on self-assessments from the participant since the actually felt emotion can strongly differ from the expected one, depending on a subject's past experience. Emotion elicitation is also influenced by the number and the complexity of the targeted emotions.

III. **Time:** the duration of an affective phenomenon can be used to define time categories that range from the "full blown emotions", lasting for some seconds or minutes, to the traits, lasting for years if not all the lifetime (Cowie et al., 2001). In between are categories such as moods or emotional disorders. In human computer interaction, most of the applications under consideration deal with what Cowie et al. (2001) define as "full blown emotions" thus managing phenomena that last from seconds to minutes. In an ideal affective interface, the emotion of the user should be detected as soon as possible (let's say in few seconds) in order to take the proper decision that directly matches the user expectation and not one that was expected minutes before. Synchronization of the different modalities is also an issue since the activation of physiological outputs can occur at different times after the stimulus onset. For instance, a change in temperature of the skin is much

slower than a change in brain activity that occurs few milliseconds after emotion elicitation.

IV. **Sensors / modalities:** various sensors can be used to measure physiological activity. Most of the devices used in modern research observe the peripheral nervous system activity by measuring sudation, blood pressure, heart rate, respiration and intensity of muscles contraction. Sudation can be inferred using GSR (Galvanic Skin Response), SCR (Skin Conductance Response) or EDR (ElectroDermal Response) sensors that measure the resistance (or the conductance) of the skin. There are two main ways to measure blood pressure: by using a plethysmograph (continuous relative value) or an inflatable cuff (discrete scaled value). Heart rate can also be computed from the continuous monitoring of blood pressure using plethysmography or by using an electrocardiogram (ECG). Respiration is generally measured using a respiration belt that measure thoracic expansion. It is also possible to use more complex apparatus to infer inspired and expired volume of air and CO_2. Finally, electromyograms provide a way to measure muscle contractions.
For usage in HCI applications, the activity of the central nervous system can be recorded by two main devices. EEGs are measured by electrodes that register the electrical potentials at the surface of the head skin due to neuronal activity. Functional Near-Infrared Spectroscopy (fNIRS) detects the light that travels through the cortex tissues and is used to monitor levels of oxygenated and deoxygenated blood that follow changes in brain activity.
Sensors used for emotion assessment should be chosen carefully so that they do not disturb the user. Firstly, sensors should not be uncomfortable for the user in order to avoid undesired emotions such as pain and frustration. Secondly, they should not prevent the use of usual modalities, such as keyboards or mice, for instance by monopolizing the hands of the user.
In our view, a modality can be defined as a form of communication, it thus involve both information transfer and interpretation of the information. When physiological sensors are used for affective computing they switch from the standard status of sensors, which only monitor and record a physiological activity, to the concept of modalities that interpret the physiological signals as an emotional information usable by the system for interaction. It is then necessary to merge these modalities to perform emotions assessment in a reliable way, taking into account redundant and complementary information such as the relation that exists between heart rate variability and respiration (Bailón et al., 2007; Pelzer et al., 1994).

V. **Emotion models / classes:** frequently used models are sets of discrete emotions like Ekman's six basic emotions (Ekman et al., 1987), and the valence-arousal space (Russell, 1980). Basic or discrete emotions have been widely used in literature, mainly because they are intuitive and are known to have a significant correlation with facial muscle activity. However, one can question the usability of some emotion labels depending on the task one wants to perform. For instance, the disgust state can be relevant to infer reactions from movies while it is generally useless in more classical human-machine interaction. Moreover, the emotion felt is often a mixture of several basic emotions yielding a more complex representation using probabilistic, fuzzy or multi-labels models (Devillers et al., 2005). As an alternative to discrete emotions, the valence-arousal model has the advantages of being continuous, which better represents the strength of an emotion, as well as more general in the sense that many applications can use the valence-arousal

space to keep the user away from negative emotional states and favor the inducement of positive feelings. It is also possible to project the more intuitive labels of emotion such as fear or joy to points or areas in this space. One drawback of this continuous model is that different labels of emotions are sometimes very close to each other when projected into the valence and arousal space. For instance, fear and anger are both negative-excited emotions. In this case it is possible to add a third dimension, generally called dominance (control over the situation) (Russell, 1980), that helps to distinguish between these emotional states.

VI. **Methods:** a wide range of methods has been used to infer affective states. Most of them are part of machine learning and pattern recognition fields. Classifiers like k-Nearest Neighbors (KNN), Functional Discriminant Analysis (FDA), neural networks, Support Vector Machines (SVM's), Relevance Vector Machines (RVM's) and others (Bishop, 2006) are useful to detect emotional classes of interest (Table 1). Regression techniques can also be used to obtain continuous estimation of emotions (Soleymani et al., 2008). Prior to inferring emotional states it is important to define some physiological features of interest. It is very challenging to find with certainty some features in physiological signals that always correlate with affective status of users. Those variables frequently differ from one user to another and they are also very sensitive to day to day variations as well as to the context of the emotion induction. To perform this selection researchers generally apply feature selection or projection algorithms like Sequential Floating Forward Search (SFFS) (Pudil et al., 1994) or Fisher projection (Duda et al., 2001).

As can be seen from Table 1 there are large differences in classification accuracy even for studies that employ the same number of classes. Although this can be partly explained by the factors detailed above, we believe that such variations mostly result from: the differences in emotion elicitation strategy, the type of physiological signals (modalities) used, and the chosen model of emotions (or emotional classes).

For emotions elicited by an event the best results obtained on more than three classes are those presented in (Lisetti and Nasoz, 2004) and (Wagner et al., 2005). One interesting point is that even though they use different stimuli (film clip versus music) they both use a method to control for the validity of the elicited emotions. In (Lisetti and Nasoz, 2004), the film clips presented were chosen according to the results obtained in a pilot study. In this pilot study, several participants were asked to evaluate clips by stating the emotion they felt as well as its intensity. In (Wagner et al., 2005), who worked with the Augsburger Database of Biosignals (AuDB), the participants where asked to freely choose music that matches the desired emotions. Both of these methods ensure that emotions felt during the experiment are intense and correspond to the expected ones. The good results obtained in (Picard et al., 2001) using a self induction protocol also tend to confirm the importance of reliable elicitation. In this study, the participant was the experimenter herself so that emotions to be induced were perfectly clear to her. Moreover, the participant remembered past emotional episodes for emotion elicitation; this implies a strong intensity since remembered episodes generally are those that have induced intense feelings.

Diverse types of physiological activity measurements from both the peripheral and the central nervous system have been used to assess emotions. Up to now, most of the studies using EEG's have shown unconvincing results (Chanel et al., 2006; Sakata et al., 2007) to infer emotional states from brain activity. One could conclude that EEG signals in the present state of the art are not effective for emotion

assessment; the present work however will argue against this. Describing the state of the art for emotion recognition based on peripheral signals is a real challenge because most studies performed classification on feature sets that include features from many types of signals, thus preventing analysis for single modalities. However, there are signals that are employed in nearly all studies, like GSR and heart rate (extracted from ECG or plethysmography). These two signals are known to correlate well with affective states (Ekman et al., 1983; Lang et al., 1993; Sinha et al., 1992). In many results it can also be seen that EMG signals are usable for emotion assessment (Rainville et al., 2006; Rani et al., 2006; Sinha and Parsons, 1996; van den Broek et al., 2006; Wagner et al., 2005). Generally, EMG electrodes are positioned to measure facial muscle activity like the venter frontalis (raising of eye-brows), zygomatic (smiling) and the corrugator supercili (frowning of eyebrows). As for facial expression analysis, measuring facial activity is strongly relevant; however placing those sensors on the face is strongly invasive which could hamper their usage in a concrete application.

One of the most evident observations that can be made from Table 1 is that different models of emotions lead to different classification accuracies. This is especially clear when comparing the basic-emotions model, which generally includes more than three categories, to emotions in the valence-arousal space model, including two or three categories. Thanks to the works of Wagner et al. (2005) and Picard et al. (2001), it is possible to compare results on valence-arousal classes to those obtained on basic emotions classes in an intra-study framework. As can be observed from (Table 1) the accuracies reported with the valence-arousal representation are similar to those reported with basic emotions. However since the number of classes is higher for basic emotion (4 to 8) than for valence-arousal classes (2), basic emotions can be considered as being better classified. Moreover, identification of valence classes is generally harder than identification of arousal classes (Table 1), which supports the idea that peripheral activity is better correlated with arousal than valence (Lang et al., 1993). No clear differences can be observed in the number of classes or labels name for basic emotions.

2.2 Toward emotional assessment using EEG

Fairly recent psychological studies regarding the relations between emotions and the brain are uncovering the strong implication of cognitive processes in emotions (Adolphs et al., 2003; Damasio et al., 2000; Davidson, 2003; Davidson et al., 2000; Rolls, 2000). From the state of the art, it can however be observed that up to now few studies have investigated the usefulness of EEG for emotion assessment. From those studies only one (Chanel et al., 2007) proposed a self-induction method for emotion elicitation: recall of past emotional episodes. Apart from the advantages of self-induction paradigms detailed in Section 2.1, this method has the advantage of activating many brain areas because cognitive processes related to memory retrieval are located throughout the brain (Damasio et al., 2000; Smith et al., 2006). All those factors make this elicitation method a good candidate for emotion assessment from EEG and peripheral signals. The current paper is a significant extension of (Chanel et al., 2007) which was a preliminary study promising enough to warrant further work. In the present article, the methodological approach is extended to investigate the fusion of peripheral and central signals. Also, the number of participants in the experimental study has increased from 1 to 10.

EEG was mostly used for BCI's (Brain Computer Interfaces) (Vaughan et al., 2003) where the goal is to provide an interface for disabled persons who cannot use standard muscular paths to communicate their intentions. Notice that because of the sensitivity of EEG sensors to noise and the fact that they often require gel to be applied on the surface of the skin, some researchers have avoided using them for others HCI applications. However, the success of BCI's led to impressive progress in the development of new EEG sensors. For instance, the use of active electrodes (MettingVanRijn et al., 1996) combined with properly designed algorithm (Kierkels et al., 2006; Romero et al., 2008) can greatly reduce noise effects. New easier to use caps and dry electrodes are also being developed. Currently some lightweight and potentially mass market EEG systems start appearing at a modest price (http://www.emotiv.com/, http://www.neurosky.com/). Although such devices only have few sensors, they demonstrate that the use of EEG is no longer the domain of physicians and scientists. For those reasons, it is now legitimate to address the problem of emotion assessment based on EEG signals analysis.

Another aspect that should be emphasized relates to the temporal dimension. Some peripheral variables require quite a long time to stabilize. For instance, heart rate variability should not be computed on epochs of less than a minute (Berntson et al., 1997). In (Salahuddin et al., 2007) the authors analyzed the usability of heart rate variability on different time periods and concluded that 50 s of signals are necessary to accurately monitor mental stress in real settings. Apart from (Haag et al., 2004) and (Leon et al., 2007) there are no studies that try to use short term monitoring of emotions using only peripheral signals (studies where the time of a trial is not specified can be assumed as lasting for minutes considering that stimuli are music or film clips). However being able to perform emotion identification in only a few seconds is certainly critical to allow for real-time applications.

Having in mind the previous considerations, the present study aims at investigating the usefulness of EEG and peripheral signals in a self-induction paradigm on short time segments of 8 s. In order to be as much application independent as possible, we used the valence arousal space as a prior model to define three emotional classes of interest that are calm-neutral, positive-excited and negative-excited.

3 Data collection

3.1 Acquisition protocol

As can be seen from the state of the art, designing an adequate protocol for eliciting emotion and recording physiological signals is not an easy task. In (Picard et al., 2001) five factors that can influence recordings were defined: subject-elicited vs. event-elicited, laboratory setting vs. real world, focus on expression vs. feeling of the emotion, openly-recorded vs. hidden recording and emotion-purpose vs. other-purpose. The following protocol description addresses those five criterions as well as those emphasized in the precedent Section.

In the present study a subject-elicited method, using recall of strong emotional episodes, is employed to elicit reliable and short time emotions. An episode is defined as a situation that lasted for a while and potentially containing several events and actions with the same emotional orientation. An example is the funeral of a relative including events such as moments of the ceremony and the burial. The elicited emotions are considered reliable because *(i)* thinking of the same episodes ought to produce similar reactions from one trial to another, *(ii)* emotional episodes

are often stored in memory because the emotions felt were quite intense,*(iii)* emotional recall is a cognitive task that induces EEG activity (Damasio et al., 2000; Smith et al., 2006) as well as modify peripheral activity (Rainville et al., 2006; Sinha et al., 1992).

Compared to other studies (Lisetti and Nasoz, 2004; Picard et al., 2001; Wagner et al., 2005), where emotions are elicited and assessed over several minutes, the duration of an emotion epoch is merely 8 s. This epoch duration was chosen because it is the maximum duration that allows maintaining the total length of the protocol below one hour to avoid participant fatigue. Within the requirement of the one hour duration this epoch is maximized for three reasons. Firstly, some peripheral features need to be determined over a sufficiently long period of time in order to be reliably computed; this is for instance the case for statistical features extracted from heart rate. In general, an epoch of 8 s should suffice for this purpose if we exclude the very low frequency features such as low frequency heart rate variability (Berntson et al., 1997). Secondly, recalling past episodes and eliciting the corresponding emotions are difficult tasks and participants might need a few seconds to accomplish them. Thirdly, the reaction time of peripheral activity from the moment where the emotion is elicited is of several seconds, with the GSR being the slowest response with a lag around 3-4 seconds.

The 11 participants (7 males, 4 females) who took part in the study were aged from 26 to 40. One week before the recording, participants were told to retrieve from their memory one excited-positive and one excited-negative episode that had occurred in their life and that they consider as being most powerful. On the day of the experiment, each participant was given a consent form where the context, the goal and a short explanation of the experiment were provided. Participants had to sign this consent form to continue further and could stop the experiment whenever they wanted. After signing the consent form, sensors were attached to the participant who was seated in front of a computer screen. A precise description of the protocol was provided with a support demonstration. This corresponds to Picard's criteria for an open-recording (participants knew they were recorded), emotion-purpose (participants new the objective of the study), and laboratory settings (participant are recorded in a controlled environment).

The complete recording session was divided into trials. During each trial participants had to accomplish a particular task according to the visual cue displayed on the monitor after a random duration display of a dark screen (Fig. 2). This task could be to self-generate one of the two excited emotions by using the past emotional episodes of their life as a support, or to stay calm and relax in order to define a third emotional state called calm-neutral. A total of T = 300 trials (100 trials per emotional state) were performed in a random order. Since facial muscle artifacts can contaminate EEG signals, participants were encouraged not to express their feelings through facial expressions, not to blink, and not to close their eyes during the 8 s of recordings (despite this some involuntary facial expressions artifacts can still remain in the signals). Emphasis was thus put on the feeling of emotions rather than on the cognitive task of remembering and on the motor expressions of emotions. A resting period of unlimited duration to relax and stretch muscles was proposed to participants after each block of 30 trials. As can be seen from Fig. 2, the chosen emotional states do not cover all areas of the valence-arousal space, especially in the bottom half of the space. This choice was made because there are actually few

emotions that are calm-negative or calm-positive (Hanjalic and Xu, 2005; Lang et al., 2005).

<Figure 2>

Data were recorded using the Biosemi Active II system (http//www.biosemi.com). EEG signals were recorded using 64 surface electrodes positioned according to the 10-10 system. Plugged on the same system to simplify synchronization, other sensors were used to record peripheral activity: a GSR (Galvanic Skin Response) sensor to evaluate sudation, a respiration belt to record abdominal expansion and a plethysmograph to measure blood pressure. Both EEG and peripheral signals were sampled at 1024 Hz. GSR electrodes were positioned on the tops of the middle finger and index finger of the same hand, the respiration belt was tied around the participant abdomen and the plethysmograph was clipped on the thumb. Even if currently the placement of those sensors is quite invasive because they prohibit the use of one hand, the wireless and wearable sensors of the future will help increase user comfort. Moreover, since heart rate can be computed from the continuous monitoring of blood pressure, the use of a plethysmograph sensor avoids having to use an ECG sensor.

After data acquisition, participants were asked to report on their experiences in an informal interview. Participants were not asked to provide a detailed description of the chosen episodes because we believe that for personal and ethical reasons a participant may hesitate to refer to his/her strongest emotional experiences. For this reason the differences in the cognitive tasks between different trials could not be fully controlled, however, as argued in (Damasio et al., 2000) the known effectiveness of mental imagery as an elicitator of powerful emotions can compensate this problem.

The present protocol for off-line acquisition of physiological signals is very close to those encountered in the BCI community so that the conclusions drawn from this paper may also have some impact in this direction. An emotion elicitation task can then be regarded as a mental task that the user tries to perform in order to communicate his or her feelings. This can be useful for severely disabled people that cannot directly express their emotions. Current BCI paradigms (Kronegg et al., 2007; Lotte et al., 2007; Vaughan et al., 2003) aim to detect brain activity that corresponds to complex tasks (mental calculus, imagination of finger taping, etc.) not related to the objective of the user (moving a mouse cursor, choosing a letter, etc.). Generally the user needs training before using such systems. In case the objective of the user is to express an emotion, classical BCI tasks (e.g., imagination of finger tapping) seem to be really far from this objective and it is more appropriate to use tasks such as the remembering of a similar emotional episode.

3.2 Features extraction

This Section describes three feature sets computed to represent physiological activity; one for the peripheral signals and two others for EEG signals. Since emotion assessment will be performed for each participant separately no baseline was computed to normalize participant physiological signals. Each feature is computed for all trials and for all participants.

3.2.1 Peripheral features

Several features extracted from physiological signals have been shown to be related to emotional activity (Lisetti and Nasoz, 2004) and their effectiveness is now fully

demonstrated as explained in Section 2.1. The current study used the following peripheral signals: GSR, blood pressure and chest cavity expansion. All signals were first filtered by a moving average filter to remove noise. For this purpose we used filters of length 512 samples for GSR, 128 for blood pressure, and 256 for respiration. Those different lengths were chosen to remove high frequencies without corrupting oscillations of interest in the different signals.

GSR provides a measure of the resistance of the skin (electrodermal activity) by positioning two electrodes on the tops of two fingers. The resistance decreases due to an increase of activity in sweat glands, which usually occurs when one is experiencing emotions such as stress or surprise. The resistance then slowly returns to its baseline level. This decrease in resistance generally occurs 3 to 4 s after stimulus onset and can be characterized by its amplitude and its duration. In (Lang et al., 1993) the authors also discovered that the mean value of the GSR is related to the level of arousal. The features extracted from electrodermal activity are presented in Table 2 and were designed to represents the characteristics of the GSR activity.

A plethysmograph is a device that uses infrared light to measure tissues blood volume, thus providing a continuous monitoring of relative blood pressure. Since heart pulses expand and contract the microvasculature, it is also possible to compute the heart rate from the plethysmograph signal. A method to determine heart rate from a blood volume pressure signal is proposed in (Aboy et al., 2005). However this method is based on a complex analysis that requires recordings of long duration. Since in this study the duration of a trial is 8 s, an alternative approach is proposed. Heart peaks were assumed to be the local maxima of the signal which were obtained by finding samples were the derivative is zero and the amplitude is switching from an increase to a decrease. If two peaks fall in the same interval of 0.5 s then only the peak with the highest amplitude is kept. This interval is chosen based on the assumption that the heart rate will not exceed 120 beats per minutes (BPM) which is somehow reasonable since the participant is sitting in front of a computer screen without performing significant physical activity. Blood pressure and heart rate variability are variables that have significant correlation with defensive reactions (Healey, 2000) and pleasantness of stimuli (Lang et al., 1993). Sinha et al. (1992) refers to the increase of blood pressure during fear and anger as one of the most consistent findings in emotion research from autonomic activity. Rainville et al. (2006) observed an increase in heart rate for many basic emotions. The features used to represent heart rate and blood pressure signals are listed in Table 2.

The respiration signal is obtained by a belt that measures the expansion of the abdomen related to the quantity of inspired and expired air. Respiration rate ranges from 0.1Hz to 0.35 Hz at rest, while it can reach 0.7Hz during exercise, but in the case of measuring respiration with a belt, irregular respiration leads to appearance of energy in higher frequencies. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear (Kim, 2004; Rainville et al., 2006). Laughing is known to affect the respiration pattern by introducing high-frequency fluctuations of the recorded signal. To capture those fluctuations, features from both the frequency and time domain are therefore used. Features of the frequency domain are obtained by computing the Fast Fourier Transform (FFT) of the original signal and of a selection of frequency bands of interest. A list of both temporal and frequency features can be found in Table 2.

<Table 2>

3.2.2 EEG features

As explained in Section 1, the cognitive theory of emotions provides a strong motivation to go toward emotion assessment using signals from the central nervous system. Several researchers have shown the implication of brain structures in emotional processes by analyzing emotional impairment after a brain traumatism and brain activity in controlled conditions. The two main areas of the brain that relate to emotional activity are the pre-frontal cortex and the amygdala (Davidson et al., 2000). From the study of pre-frontal EEG alpha waves Davidson (2003) demonstrated the lateralization of this area, with a higher brain activity in the right hemisphere for negative or withdrawal stimuli and the opposite pattern for positive or approach stimuli. Amygdala activation seems to be more related to negative emotions, such as fear. There is currently no consensus about a possible lateralization of the amygdala and its involvement in positive emotions. These brain regions are certainly not the only ones involved in emotional processes. For instance, Aftanas et al. (2004) reported differences in Event Related Desynchronization / Synchronization (ERD/ERS) during the visualization of more or less arousing images. Those differences were observed in theta bands for the parietal and occipital areas, alpha bands for the frontal areas and gamma bands for the complete scalp. Concerning the particular case of the recall paradigm, Smith et al. (2006) showed an augmentation of activity in the connections between the hippocampus and the amygdala during the recollection of negative events compared to neutral events, while Damasio et al. (2000) found activation of many cortical areas during the feeling of self-generated emotions.

For the current study, it is hard to predict which brain areas are supposed to be activated since this strongly depends on the memories the participant used to relive the emotion. For instance, memories of a rather auditory nature will activate the auditory cortex while visual memories will activate the occipital cortex. Moreover, as the structures involved in recollection of events are deep in the brain and hard to precisely capture using EEG electrodes, two widely applicable feature extractions methods were chosen. One assisted in the computation of power features for all electrodes within limited frequency bands, while the other focused on the common information contained in each pair of electrodes. Prior to extracting these features from EEG data, noise needs to be removed by pre-processing the signals. Environment noise and drifts were removed by applying a 4-45Hz bandpass filter while other noises were considered as non-significant. The second step was to re-reference all electrode signals to a Laplacian reference.

Once EEG signals were pre-processed, the first set of EEG features was extracted by computing the Short-Time Fourier Transform (STFT) for each electrode with a sliding window of 512 samples and 50% overlap between two consecutive windows. For each of the 64 spectrograms (one per electrode), we selected 9 frequency bands ranging from 4Hz to 22Hz (Δf = 2Hz); this was done according to psycho-physiological literature (Aftanas et al., 2004; Davidson, 2003). The total number of features extracted by this method is 16704 (64 electrodes x 9 frequency bands x 29 time frames).

For the second set of features, mutual information (MI) between pairs of electrodes is proposed as a measure of statistical dependencies between different areas of the brain. This set of features was motivated by studies that demonstrated

synchronization of brains areas in emotional processes (Ansari-Asl et al., 2007; Grandjean et al., 2008). With the assumption that the signal of electrode i for a given trial is a stochastic process with probability mass function $P(X_i)$ then the mutual information between electrodes i and j for this trial is expressed as:

$$I(X_i;X_j) = H(X_i) - H(X_i \mid X_j)$$

$$H(X_i) = -\sum_{x_i} P(X_i = x_i)\log(P(X_i = x_i))$$

$$H(X_i / X_j) = -\sum_{x_i,x_j} P(X_i = x_i, X_j = x_j)\log(P(X_i = x_i / X_j = x_j))$$

where $H(X_i)$ and $H(X_i \mid X_j)$ are respectively the entropy and conditional entropy of random variables X_i and X_j. Mutual information was computed using Moddemeijer's matlab toolbox (Moddemeijer, 1989)(available at http://www.cs.rug.nl/~rudy/matlab/) that estimates the different distributions based on histograms and automatically determines the appropriate bin size. The MI feature vector $\mathbf{f}^{MI}$ of this trial is then constructed by concatenation of mutual informations between each pairs of electrodes:

$$\mathbf{f}^{MI} = [I(X_1,X_2)\ldots I(X_1,X_M), I(X_i,X_{i+1})\ldots I(X_i,X_M), I(X_{M-1},X_M)]$$

The total number of features of a trial for M=64 electrodes is then: $\sum_{i=1}^{M-1} i = 2016$.

4 Methods

This section presents the methods used to assess emotions from the recorded physiological signals, to fuse the different set of features and to improve on the classification accuracy by rejecting samples having low confidence value.

4.1 Classification

Since each recorded trial corresponds to a particular emotional state, it is easy to formulate a classification task (called CPN for "calm", "positive", "negative") where the three ground-truth classes ω_c, ω_p, ω_n correspond to calm-neutral, positive-excited and negative-excited patterns. A target class vector $Y^{CPN} = [y_1\ldots,y_i,\ldots,y_T]^T$ is constructed, where $y_i \in \{ \omega_c, \omega_p, \omega_n \}$ represents the class of the trial i. We also address other classification tasks by constructing different target vectors to distinguish between the following emotional states: negative excited vs. positive-excited (NP), calm-neutral vs. positive-excited (CP), calm-neutral vs. negative excited (CN), calm vs. excited (CE) by regrouping samples of the positive-excited and negative-excited states.

As summarized in Fig. 3, there are three sets of features, F^{Periph}, F^{STFT} and F^{MI} that contain respectively peripheral features, STFT EEG features and MI EEG features for all trials. Those feature sets are associated with the class vectors Y^{CPN}, Y^{NP}, Y^{CP}, Y^{CN} and Y^{CE}, depending on the classification task to address. Notice that F^{STFT} and F^{MI} are high-dimensional features spaces which is a problem for real applications where time and storage issues are of importance. However, the current study focuses on the improvement of the classification accuracy. The issue of reducing the dimensionality of the problem was addressed in (Chanel et al., 2007) where it was

shown that feature selection does not clearly improve results of the best classifier. Only linear classifiers are applied on those two feature sets since there is always a linear boundary that can completely separate training samples of the different classes and linear classifiers solutions are well regularized so that they give better generalized solutions.

<Figure 3>

In the present study, different classifiers were trained on the three feature sets to recover the ground truth classes. Then the best classifier was chosen for each feature set for later fusion (see Section 4.2). To evaluate the accuracy of each classifier, a leave-one-out strategy was chosen. This involves using each feature vector in turn as the test set and the remaining ones as the learning set. At each step, a classifier is trained from the learning set and then applied to the test sample. This leave-one-out strategy was chosen since it provides the maximum possible size of the learning set. This is preferable in this problem because the number of samples (T=300) is very low compared to the size of the EEG feature spaces.

4.1.1 Discriminant analysis

Two discriminant analysis methods, namely the linear discriminant analysis (LDA) and the Quadratic discriminant analysis (QDA) are used in this paper. Both are based on the Bayes rule to find the class with the highest posterior probability $P(\omega_i \mid \mathbf{f})$ (Duda et al., 2001). Under the assumption that the conditional distributions $P(\mathbf{f} \mid \omega_i)$ are Gaussians with different means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ this rule automatically defines a quadratic decision boundary (hence the name QDA for the associated classifier):

$$P(\omega_i \mid \mathbf{f}) = \frac{P(\mathbf{f} \mid \omega_i).P(\omega_i)}{\sum_{i=1}^{C} P(\mathbf{f} \mid \omega_i).P(\omega_i)} = \frac{\mathcal{N}(\mathbf{f} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).P(\omega_i)}{\sum_{i=1}^{C} \mathcal{N}(\mathbf{f} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).P(\omega_i)}$$

Vectors $\boldsymbol{\mu}_i$ and matrices $\boldsymbol{\Sigma}_i$ are computed from the learning set. In the case where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j, \forall i \neq j$ the boundary becomes linear, yielding an LDA classifier. With the LDA it is sufficient to compute a single covariance matrix $\boldsymbol{\Sigma}$ from the complete learning set without distinction between classes.

Here the prior probability $P(\omega_i)$ was defined as 1/K where K is the number of classes. Due to the high number of EEG features and low number of samples, discriminant analysis can fall in the singularity problem. In this case we used the diagonalized version where covariance matrices are assumed to be diagonal, containing the variances of all features. The Matlab statistics toolbox (v. 5.0.1) implementation of those algorithms was used in this study.

4.1.2 Support Vector Machines (SVM's)

A SVM is a two class maximum margin classifiers that tries to maximize the distance between the decision surface and the nearest point to this surface as well as to minimize the error on the training set. SVM's minimize an upper bound on the expected risk rather than only the error on the training data, thus enabling good generalization as well as interesting performance in high dimensional feature spaces (Chanel et al., 2007; Hua et al., 2005). Here, both linear and radial basis function (RBF) kernels are used. In the case of RBF kernels, the size of the kernel is chosen

based on the results of 5-fold cross-validation procedures with values ranging from 5.10^{-3} to 0.5 with step 5.10^{-3}. The C parameter that regulates the tradeoff between error minimization and margin maximization is empirically set to 1.

There are two drawbacks to the use of SVM's as classifiers: they are intrinsically only two-class classifiers and their output is uncalibrated so that it is not directly usable as a confidence value in the case one wants to combine outputs of different classifiers or modalities. In this study the first point was addressed by using the one versus all approach where N classifiers are trained for each class and the final choice is done by majority voting. For the second point, Platt (2000) proposes to model the probability of being in one of the two classes knowing the output value of the SVM by using a sigmoid fit, while Wu et al. (2004) proposes a solution to extend this idea to multiple classes. The libSVM (Chang and Lin, 2001) Matlab toolbox was used as an implementation of these algorithms.

4.1.3 Relevance Vector Machines (RVM's)

RVM's (Tipping, 2001) are algorithms that have the same functional form as SVM's but embedded in a Bayesian learning framework. They have shown to provide results similar to SVM's with generally sparser solutions. They have the advantage that they directly give an estimation of the posterior probability of having class ω_i.

RVM's try to maximize the likelihood function of the training set using a linear model including kernels. The main difference with more classical probabilistic discriminative models is that a different prior is applied on each weight thus leading to sparse solutions that should generalize well. In this paper, the multiclass RVM version presented in (Zhang and Malik, 2005) was used.

4.2 Fusion and ambiguity rejection

The interest of fusing peripheral and EEG features for emotion assessment was shown in (Chanel et al., 2006) through a simple concatenation of feature sets. The fusion of different EEG features is also known to improve results for EEG signals classification (Blankertz et al., 2003; Lotte et al., 2007). In a classification problem, fusion can be applied at the sensor level to raw acquired data, at the feature level, at the classifier level, and at the decision level by combining the output of classifiers (Sanderson and Paliwal, 2004).

This paper focuses on the fusion at the decision level. More specifically, accuracies of several classifiers are evaluated on the feature sets F^{Periph}, F^{STFT} and F^{MI}. For each of those feature sets, the classifier with the best accuracy is selected so that we obtain three best classifiers named q^{Periph}, q^{STFT} and q^{MI}. The final decision is then made by combining the decisions of two or all of those classifiers.

If the classifiers output some confidence measures on their decision, combining decisions of classifiers can be done using summation rules and product rules. In this work, the probabilistic outputs of classifiers are used as a measure of confidence. The sum rule is thus defined as follow for a given trial:

$$g_i = \frac{\sum_{q \in Q} P_q(\omega_i \mid \mathbf{f})}{\sum_{i=1}^{K} \sum_{q \in Q} P_q(\omega_i \mid \mathbf{f})} = \sum_{q \in Q} \frac{1}{|Q|} P_q(\omega_i \mid \mathbf{f})$$

where Q is the ensemble of the classifiers chosen for fusion, $|Q|$ the number of such classifiers and $P_q(\omega_i \mid \mathbf{f})$ is the posterior probability of having class ω_i according to classifier q. The final choice is done by selecting the class ω_i with the highest g_i. It can be observed that g_i can also be viewed as a confidence measure on the class given by the fusion of classifiers.

As a final step, rejection of trials that have a confidence value g_i below a threshold δ was performed to improve classification accuracy. When a sample is rejected because the confidence value is not sufficiently high, the sample is not classified and the classification accuracy is computed only on the samples with high confidence. The percentage of rejected samples as well as the accuracy computed on the remaining samples thus become of function of the δ threshold. A good value for δ would be one that provides a compromise between accuracy maximization and rejection rate minimization. The proper value for this parameter will be discussed in Section 5.2.2.

5 Results and discussion

This chapter reports and discusses the classification accuracies obtained for emotion assessment. Section 5.1 details participants' reports and explains why the protocol is considered to be successful in eliciting emotions. Section 5.2 reports the accuracies obtained for each combination of classifier and feature set. Those results are then used to associate the optimal classifier to each feature set for later fusion. Finally, Section 5.3 shows the positive effects of fusion and rejection of samples having a low confidence value.

5.1 Participants reports and protocol validation

Out of the 11 recorded participants 10 reported a successful elicitation of the emotions by recalling emotional episodes. As can be seen from Fig. 6, which represents the average accuracies obtained from the 10 participant cited above, the peripheral activity is useful to distinguish between different classes of emotions. This implies that different patterns of physiological activity where induced for each emotional task and thus supports the idea that emotions were successfully elicited. However, all participants reported that it was really difficult to stay concentrated throughout the entire recording. A recurring observation was also that switching from one emotion to another very quickly was sometimes confusing and hard to accomplish. The effects of such observations can be missing trials where the participants did not accomplish the requested task, the elicitation of the undesired boredom emotion which can interfere with positive and negative excited trials, and noisy EEG signals due to fatigue.

In the protocol presented in Section 3.1 brain activity can be induced by two cognitive components: the actual events of the episode (for instance thinking of someone crying) and the emotion elicitation following the event. Since our aim is to detect emotions, it is important to control that the events used to induce emotions were not always the same to ensure that what is detected from brain signals is the emotion and not the cognitive task related to the event (for instance mental imagery of the act of crying). Since participants did not report about the episodes they used to induce emotions it is difficult to control for this, however the following remarks lead us to assume the protocol is valid:

- two participants reported that they thought of different episodes within the same category (ie. positive-excited and negative-excited). The classification accuracy obtained from the signals of one of those two participants actually corresponds to the best results across the 10 participants. The other one obtained average accuracies;
- one participant reported that he thought to the episodes without concentrating on the feeling of emotions which resulted in a weak emotion elicitation. All the accuracies computed from the signals of this participant are at the random level;
- since an episode was defined as including several emotional events of the same category, it is unlikely that the participants always thought of the same event to elicit one of the emotions;
- the participants were explicitly told to focus on the feeling of emotions and emotions were successfully elicited as stated above.

Notice that the participant who did not concentrate on the feeling of emotions was removed for further analysis since he did not follow the protocol properly.

5.2 Single modality results

Figs. 4, 5, 6 respectively present the mean accuracy across participants for the STFT EEG features, the MI features and the peripheral features. The accuracies of different modalities and classifiers are compared below to answer the following questions: what is the effectiveness of EEG and peripheral features to assess emotions according to the different classification schemes and which classifiers should be used for latter fusion of feature sets.

<Figure 4>

<Figure 5>

<Figure 6>

STFT EEG features provided interesting results with a mean classification accuracy of 63% for three classes and a SVM classifier (the random level is at 33% accuracy). The best average accuracy for two classes is obtained from the CP classification task with nearly 80% of well classified trials (random level at 50%), followed by the CE and NP classification tasks with respectively 78% and 74% of accuracy. For all participants and all classification tasks, the results are higher than the random levels (33% for three classes and 50% for two classes). MI features seem to be a bit less suitable for emotion classification than STFT features with an approximate decrease of well classified trials of 2% to 4%, except for the NP classification task where a slight performance increase was noted. It is hard to compare those results to the state of the art because there are only few studies using EEG. In (Chanel et al., 2006) the best accuracy on two and three arousal classes was respectively of 72% and 58%. In this study the highest accuracies for the CE and CPN classification tasks are respectively of 88% and 86.3%. The best result for a two class task is obtained on the NP task with 96% of accuracy. In (Sakata et al., 2007) an accuracy of 29% was obtained for 6 different emotional classes while the accuracy was of 42% for identification of 5 emotional states in (Takahashi, 2004). Our results are thus superior to the previous studies using the EEG modality for detecting emotions expressed in the valence-arousal space and in alignment with results obtained on emotional labels.

To check for the usability of this emotional protocol as a new BCI paradigm our results were compared to BCI accuracies. In (Guger et al., 2003) the authors showed that around 75% of the 99 untrained participants that took part in a two class BCI paradigm without feedback obtained accuracies between 60% and 79%. The distributions of the accuracies for our recall paradigm are similar; however more participants should be recorded to validate this statement. Our results are also far from those of more recent BCI studies where the accuracy can reach more than 90% for two classes for many untrained participants (Kronegg et al., 2007). This can be due to the definition of mental task that are chosen to activate well separated areas of the brain, contrary to the task definition used in this study.

If the three Figs. 4, 5 and 6 are compared, it is obvious that the EEG features lead to better accuracy than peripheral features for all classification schemes. For peripheral features the LDA classifier is the best with an average accuracy of 51% for three classes and around 66% for two classes (except for the NP classification task with accuracy around 61%). Results ranged from nearly the random level up to around 80% for two class formulations and from 37% up to 75% for three classes, showing the importance of this modality for at least one participant. However there is an exception, the CE classification task, where the LDA does not have the best accuracy. In this task the sparse kernel machines have better accuracies but they were sensitive to the unbalanced nature of this configuration with 200 samples belonging to the excited class and 100 samples belonging to the calm class. As can be seen from the confusion matrices of Table 3, sparse kernel machines tend to always assign the excited class to test samples. Those results where thus considered as irrelevant and the LDA classifier chosen as the most relevant classifier for fusion.

<Table 3>

Compared to the state of the art of emotion assessment from peripheral signals and time segments of similar duration our results are under those reported. In (Haag et al., 2004) 90% and 97% of accuracy was obtained using time windows of 2 s for valence and arousal assessment respectively. However the accuracy they report represents the number of samples from which the output of a neural network regressor falls in a 20% interval of the target value. Thus this accuracy cannot be directly compared to classification tasks. In (Leon et al., 2007) the classification strategy discriminated three emotional states (neutral, positive and negative) with an accuracy of 71% from 6 s signals. However this accuracy was obtained on only one participant after training the algorithm on 8 participants. To give an example of the variability of results that can be obtained from a participant to another, in our study results ranged from 40% for the worst participant to 81% for the best considering only the best classifier. However, the classification strategy used in (Leon et al., 2007) included a detection of signals corruption, which demonstrates the importance of such a procedure for correct emotion assessment.

The large differences in accuracy between the EEG and peripheral features can be explained by two factors. Firstly, the protocol is based on a cognitive elicitation of emotions where participants are asked to remember past emotional episodes which ensures strong brain activities. Moreover, the emphasis was put on the internal feeling of emotions rather than on the expression of emotion that can help to induce peripheral reactions (Ekman et al., 1983). Secondly, the 8 s length of trials may be

too short for a complete activation of peripheral signals while it may be sufficient for EEG signals.

For both EEG and peripheral features there is always high variability of results across the participants. For instance the accuracies ranged from 48% to 92% to classify emotions in three classes. This variability can be explained by the fact that the participants had more or less difficulty in accomplishing the requested tasks as reported during the interview. Another remark that holds for all feature sets is that the detection of arousal states is more accurate than the detection of valence sates. This is not surprising for peripheral activity since it is known to better correlate with the arousal axis than with the valence axis (Lang et al., 1993), and sheds some new light on the usability of EEG for the detection of arousal and valence. Notice that the standard deviation is lower for arousal identification then for all other combination of emotional states showing that arousal states are detected with more stability across participants.

This study also allows comparing the performances of the different classifiers in the three feature spaces. For the peripheral feature set (Fig. 6), the classifiers have relatively similar accuracies except for the QDA which performs poorly compared to the others. Since this algorithm needs to compute a covariance matrix for each class, the low number of samples that are available for learning (around 100 per class) explains this result. The RBF SVM does not perform as well as the other classifiers for the two classes formulations, suggesting that those problems are linear by nature. For the high dimensional spaces of EEG features the LDA accuracy is always about 10% below the results obtained by SVM classification. This confirms the effectiveness of SVM's in high dimensional spaces (Hua et al., 2005). One of the goals of the present work was also to determine which of the RVM and probabilistic SVM would have the best accuracies in order to use the best algorithm for the purpose of fusion. As can be seen from Figs. 4 and 5, the probabilistic SVM performs as well as the standard SVM demonstrating the interest of such a classifier to perform fusion on the basis of standardized scores. The RVM classifier outperforms the LDA, showing its adequacy for high dimensional spaces but does not outperform the SVM. An explanation could be that RVM's generally used less support vectors than SVM's which is not desirable in those undersampled classification tasks where good generalization is hard to obtain.

5.3 Fusion and rejection results

5.3.1 Fusion

Fusion of classifier decisions is done according to the explanation given in Section 4.2. According to the obtained results, fusion was performed choosing probabilistic SVM as the classifiers for EEG features sets (q^{STFT} and q^{MI}), and the LDA as the classifier for the peripheral feature set (q^{Periph}).

Results from the fusion of MI and STFT EEG features as well as fusion of all EEG and peripheral features are presented in Fig. 7. As can be seen, combining EEG feature sets increased the best average accuracy by 2% to 4% while combining the three feature sets increased it by 3% to 7%. In all the present cases combining feature sets leads to an increase in average accuracy, even when fusing modalities with low accuracies such as the peripheral signals. This demonstrates the importance of combining multiple sources of information from both the central en peripheral nervous system in emotion detection from physiological signals. There are

two studies that tried to fuse peripheral and EEG information, both at the feature level (Chanel et al., 2006; Takahashi, 2004). In (Takahashi, 2004) the authors found that the fusion did not improve accuracy compared to EEG classification while in (Chanel et al., 2006) an increase was reported only for some classifiers and sets of classes. Fusion of the different modalities at the feature level was performed in the present work but the results are not reported because no increase of accuracy was found. This emphasizes the importance of fusion at the decision level for emotion assessment.

<Figure 7>

5.3.2 Rejection

Finally, samples with low confidence values are rejected using the method described in Section 4.2 and the corresponding increase in accuracy is analyzed in Fig. 8. This was done for the CPN, NP and CE classification task because they are the most relevant for HCI applications. In Fig. 8, only the results of the CPN configuration are presented for the trials of all 10 participants (3000 trials) and different values of the δ threshold. Since the label of each trial is already determined after fusion, it is possible to compare the number of badly classified trials that are rejected to the correctly classified ones. As can be seen from Fig. 8, no samples are rejected until δ reach the value of 33%, which is normal since $\max_i g_i$ cannot be inferior to 33% (there is the constraint $\sum_{i=1}^{K} g_i = 1$). The number of rejected samples that are badly classified is higher than the number of correctly classified samples until δ becomes higher than 47%. We choose this value to stop rejecting samples since most of the badly classified samples are rejected at this point.

This value corresponds to a mean accuracy across participants of 80%, thus increasing it by about 10%. This is to be compared with the 70% accuracy when performing fusion without rejection, but at the cost of rejecting 40% of the samples. Such high rejection rate could seem problematic for a real application, but is however compensated by the short recording period needed to perform classification and give a decision. For instance if two consecutive trials are rejected, and the third one correctly classified the whole process would still be completed within 25 s. Using the same value of 40%, the percentage of rejected samples for the NP and the CE classification task, the increase of accuracy was respectively of 11% and 10%, resulting in an accuracy of 89% and 92%. This shows the interest of rejecting samples to improve classification accuracy for other classification tasks.

<Figure 8>

6 Conclusions and future work

This paper proposes an approach to classify emotions in the three main areas of the valence-arousal space by using physiological signals from both the peripheral nervous system and the central nervous system. A protocol based on the recall of past emotional episodes was designed to acquire short-term emotional data from 11 participants. From the data of 10 participants we extracted three feature sets, two for EEG signals and one for peripheral signals. Using the different feature sets, the accuracy of several classifiers was compared on the discrimination of the different combinations of three emotional states. The fusion of the three feature sets at the

classifiers decision level, by combining the probabilistic outputs of classifiers, was analyzed. Finally, rejection of trials where the confidence of the resulting classification is low was performed. In the case the trials with low confidence are those that are misclassified such rejection should lead to an increase of accuracy.

Results showed the importance of EEG signals for emotion assessment by classification as they had better accuracy than peripheral signals on the 8 s of recorded signal. Classification of time-frequency features derived from the EEG signals provided an average accuracy of 63% for three emotional classes and between 73% and 80% for two classes. However, peripheral features were shown to increase accuracy when fused with EEG features. Fusion of different EEG feature sets also increased the performance of the emotion assessment to obtain 70% of accuracy on three classes by fusing the three physiological feature sets. Finally, the rejection of 40% of samples having a low confidence value increased the accuracy to up to 80%.

Since following the stimulus onset emotional processes in brain and peripheral signals are expected to be observable at different times, the exploration of different time resolutions is needed to determine the time scales favorable to emotional assessment from EEG and peripheral activity. For this purpose a protocol where the exact time of the emotion elicitation is known should be designed. The high number of electrodes used in this study is also an issue since it leads to a high dimensional space where classification is difficult and it forbids the use of this system for real applications. Our study (Ansari-Asl et al., 2007), based on the data from the same protocol, is a first step in this direction.

Analysis of EEG in other elicitation contexts should also be performed to confirm the efficiency of EEG features for emotional assessment in less cognitive tasks, as well as when interacting with computer interfaces. For HCI, the described work can also be used as a guideline to decide which classification strategy to use. Finally, while the rejection of non-reliable trials has been shown to improve accuracy, the percentage of rejected samples is high and further analysis should be conducted to confirm that this rejection can improve the information transfer rate.

7 Acknowledgements

This work has been supported by the European Networks of excellence Petamedia (http://www.petamedia.eu) and Similar (http://www.similar.cc), as well as by the Swiss National Science Foundation. The authors gratefully acknowledge Prof. Klaus Scherer, Dr. David Sander, Dr. Didier Grandjean and Dr. Sylvain Delplanque from the Swiss Center for Affective Sciences (http://www.affective-sciences.org) for a number of helpful discussions.

References

Aboy, M., McNames, J., Thong, T., Tsunami, D., Ellenby, M.S., Goldstein, B., 2005. An automatic beat detection algorithm for pressure signals. IEEE Transactions on Biomedical Engineering 52(10), 1662-1670.

Adolphs, R., Tranel, D., Damasio, A.R., 2003. Dissociable neural systems for recognizing emotions. Brain and Cognition 52(1), 61-69.

Aftanas, L.I., Reva, N.V., Varlamov, A.A., Pavlov, S.V., Makhnev, V.P., 2004. Analysis of Evoked EEG Synchronization and Desynchronization in Conditions

of Emotional Activation in Humans: Temporal and Topographic Characteristics. Neuroscience and Behavioral Physiology 34(8), 859-867.

Ansari-Asl, K., Chanel, G., Pun, T., 2007. A channel selection method for EEG classification in emotion assessment based on synchronization likelihood. to be published in 15th Eur. Signal Proc. Conf. (Eusipco 2007), Poznan, Poland.

Bailón, R., Laguna, P., Mainardi, L., Sörnmo, L., 2007. Analysis of Heart Rate Variability Using Time-Varying Frequency Bands Based on Respiratory Frequency. IEEE 29th Int. Conf. EMBS, Lyon, France.

Benoit, A., Bonnaud, L., Caplier, A., Ngo, P., Lawson, L., Trevisan, D., Levacik, V., Mancas, C., Chanel, G., 2006. Multimodal Focus Attention and Stress Detection and Feedback in an Augmented Driver Simulator. 3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI), Athens, Greece.

Berntson, G.G., Bigger, J.T., Eckberg, D.L., Grossman, P., Kaufmann, P.G., Malik, M., Nagaraja, H.N., Porges, S.W., Saul, J.P., Stone, P.H., VanderMolen, M.W., 1997. Heart rate variability: Origins, methods, and interpretive caveats. Psychophysiology 34(6), 623-648.

Bishop, C.M., 2006. Pattern recognition and machine learning, Springer.

Blankertz, B., Dornhege, G., Schafer, C., Krepki, R., Kohlmorgen, J., Muller, K.R., Kunzmann, V., Losch, F., Curio, G., 2003. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. IEEE Transactions on Neural Systems and Rehabilitation Engineering 11(2), 127-131.

Brave, S., Nass, C., Hutchinson, K., 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. International Journal of Human Computer Studies 62(2), 161-178.

Chanel, G., Ansari-Asl, K., Pun, T., 2007. Valence-arousal evaluation using physiological signals in an emotion recall paradigm. IEEE SMC and International Conference on Systems, Man and Cybernetics, Smart cooperative systems and cybernetics: advancing knowledge and security for humanity, Montreal, Canada.

Chanel, G., Kronegg, J., Grandjean, D., Pun, T., 2006. Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals, in: B. Gunsel, A. K. J., A. M. Tekalp, B. Sankur (Eds.), Multimedia Content Representation, Classification and Security, Springer LNCS, Istanbul, Turkey, 4105, pp. 530-537.

Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T., 2008. Boredom, Engagement and Anxiety as Indicators for Adaptation to Difficulty in Games. 12th International MindTrek Conference: Entertainment and Media in the Ubiquitous Era, ACM, Tampere, Finland.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM : a library for support vector machines.

Chen, J., 2007. Flow in games (and everything else) - A well-designed game transports its players to their personal Flow Zones, delivering genuine feelings of pleasure and happiness. Communications of the ACM 50(4), 31-34.

Choi, D.H., Kim, J., Kim, S.H., 2007. ERP training with a web-based electronic learning system: The flow theory perspective. International Journal of Human-Computer Studies 65(3), 223-243.

Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S., 2003. Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Understanding 91(1-2), 160-187.

Cornelius, R.R., 1996. The Science of Emotion. Upper Saddle River, NJ, Prentice-Hall.

Coulson, M., 2004. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. Journal of Nonverbal Behavior 28(2), 117-139.

Cowie, R., 2000. Describing the emotional states expressed in speech. ISCA Workshop on Speech and Emotion, Northern Ireland.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32-80.

Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J., Hichwa, R.D., 2000. Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat Neurosci 3(10), 1049-1056.

Davidson, R.J., 2003. Affective neuroscience and psychophysiology: Toward a synthesis. Psychophysiology 40(5), 655-665.

Davidson, R.J., Jackson, D.C., Kalin, N.H., 2000. Emotion, Plasticity, Context, and Regulation: Perspectives From Affective Neuroscience. Psychological Bulletin 126(6), 890-909.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407-422.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, Wiley Interscience.

Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., 1987. Universals and cultural differences in the judgments of facial expressions of emotion. Journal of Personality and Social Psychology 53(4), 712-717.

Ekman, P., Levenson, R.W., Friesen, W.V., 1983. Autonomic Nervous-System Activity Distinguishes among Emotions. Science 221(4616), 1208-1210.

Fasel, B., Luettin, J., 2003. Automatic facial expression analysis: a survey. Pattern Recognition 36(1), 259-275.

Grandjean, D., Sander, D., Scherer, K.R., 2008. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Consciousness and Cognition 17(2), 484-495.

Guger, C., Edlinger, G., Harkam, W., Niedermayer, I., Pfurtscheller, G., 2003. How many people are able to operate an EEG-based brain-computer interface (BCI)? IEEE Transactions on Neural Systems and Rehabilitation Engineering 11(2), 145-147.

Haag, A., Goronzy, S., Schaich, P., Williams, J., 2004. Emotion Recognition Using Bio-Sensors: First Step Toward an Automatic System. Affective Dialog Systems: Tutorial And Research Workshop, Kloster Irsee, Germany.

Hanjalic, A., Xu, L.-Q., 2005. Affective video content representation and modeling. IEEE Trans. on multimedia 7(1), 143-154.

Hazlett, R.L., Benedek, J., 2007. Measuring emotional valence to understand the user's experience of software. International Journal of Human-Computer Studies 65(4), 306-314.

Healey, J.A., 2000. Wearable and Automotive Systems for Affect Recognition from Physiology. Electrical Engineering and Computer Science Dept., MIT. Doctor of Philosophy.

Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21(8), 1509-1515.

Isomursu, M., Tahti, M., Vainamo, S., Kuutti, K., 2007. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. International Journal of Human-Computer Studies 65(4), 404-418.

Kapoor, A., Burleson, W., Picard, R.W., 2007. Automatic prediction of frustration. International Journal of Human-Computer Studies 65(8), 724-736.

Kierkels, J.J.M., G.J.M, v.B., L.L.M, V., 2006. A model-based objective evaluation of eye movement correction in EEG recordings. IEEE Transactions on Biomedical Enginneering 53(2), 246-253.

Kim, J., 2004. Emotion Recognition from Physiological Measurement. Humaine European Network of Excellence Workshop, Santorini, Greece.

Kim, J., Andre, E., Rehm, M., Vogt, T., Wagner, J., 2005. Integrating Information from Speech and Physiological Signals to Achieve Emotional Sensitivity. Proc. of the 9th European Conference on Speech Communication and Technology, Lisboa, Potugal.

Kronegg, J., Chanel, G., Voloshynovskiy, S., Pun, T., 2007. EEG-based synchronized brain-computer interfaces: A model for optimizing the number of mental tasks. IEEE Transactions on Neural Systems and Rehabilitation Engineering 15(1), 50-58.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 2005. International affective picture system (IAPS): Digitized photographs, instruction manual and affective ratings. Technical Report A-6, University of Florida, Gainesville, FL.

Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O., 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology 30(3), 261-273.

Leon, E., Clarke, G., Callaghan, V., Sepulveda, F., 2007. A user-independent real-time emotion recognition system for software agents in domestic environments. Engineering Applications of Artificial Intelligence 20(3), 337-345.

Lisetti, C.L., Nasoz, F., 2004. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. Journal on applied Signal Processing 11, 1672-1687.

Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain-computer interfaces. Journal of Neural Engineering 4(2), R1-R13.

Mandryk, R.L., Atkins, M.S., 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int. Journal of Human-Computer Studies 65(4), 329-347.

Mandryk, R.L., Inkpen, K.M., Calvert, T.W., 2006. Using psychophysiological techniques to measure user experience with entertainment technologies. Behaviour & Information Technology 25(2), 141-158.

MettingVanRijn, A.C., Kuiper, A.P., Dankers, T.E., Grimbergen, C.A., 1996. Low-cost active electrode improves the resolution in biopotential recordings. 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Amsterdam, The Netherlands.

Moddemeijer, R., 1989. On Estimation of Entropy and Mutual Information of Continuous Distributions. Signal Processing 16(3), 233-248.

Norman, D.A., 1990. The Psychology of Everyday Things. New York, Doubleday / Currency.

Pantic, M., Rothkrantz, L.J.M., 2003. Toward an Affect-Sensitive Multimodal Human-Computer Interaction. Proc. of the IEEE 91(9), 1370-1390.

Pelzer, M., Schipke, J.D., Horstkotte, D., Arnold, G., 1994. Effect of Respiration on Short-Term Heart-Rate-Variability. Faseb Journal 8(5), A846-A846.

Picard, R.W., 1997. Affective computing, The MIT press.

Picard, R.W., Daily, S.B., 2005. Evaluating Affective Interactions: Alternatives to Asking What Users Feel. CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches, Portland.

Picard, R.W., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175-1191.

Platt, J., 2000. Probabilities for SV Machines, in: Smola, A., Bartlett, P., Scholkopf, B.,Schuurmans, D. (Eds.), Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, pp. 61-64.

Pudil, P., Ferri, F., Novovicová, J., Kittler, J., 1994. Floating search methods for feature selection with nonmonotonic criterion functions. Proc. of the IEEE Intl. Conf. on Pattern Recognition 2, 279-283.

Rainville, P., Bechara, A., Naqvi, N., Damasio, A.R., 2006. Basic emotions are associated with distinct patterns of cardiorespiratory activity. International Journal of Psychophysiology 61(1), 5-18.

Rani, P., Liu, C., Sarkar, N., 2006. An Empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Analysis & Applications 9, 58-69.

Rani, P., Sarkar, N., Liu, C., 2005. Maintaining Optimal Challenge in Computer Games through Real-Time Physiological Feedback. 11th HCI International, Las Vegas, USA, Lawrence Erlbaum Associates, Inc.

Rolls, E.T., 2000. Précis of The brain and emotion. Behavioral and Brain Sciences 23(2), 177-233.

Romero, S., Mananas, M.A., Barbanoj, M.J., 2008. A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case. Computers in Biology and Medecine 38(3), 348-360.

Russell, J.A., 1980. A Circumplex Model of Affect. Journal of Personality and Social Psychology 39(6), 1161-1178.

Sakata, T., Watanuki, S., Sakamoto, H., Sumi, T., Kim, Y.-K., 2007. Objective evaluation of Kansei by a complementary use of physiological indexes, brain wave and facial expressions for user oriented designs. Proc. of the10th Qmod conference, Quality Management and Organisational Development: Our Dreams of Excellence, Helsingborg, Sweden.

Salahuddin, L., Cho, J., Jeong, M.G., Kim, D., 2007. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. IEEE 29th International conference of the EMBS, Lyon, France.

Sander, D., Grandjean, D., Scherer, K.R., 2005. A systems approach to appraisal mechanisms in emotion. Neural Networks 18(4), 317-352.

Sanderson, C., Paliwal, K.K., 2004. Identity verification using speech and face information. Digital Signal Processing 14(5), 449-480.

Scherer, K.R., 2001. Appraisal considered as a process of multi-level sequential checking. New York and Oxford, Oxford University Press.

Sinha, R., Lovallo, W.R., Parsons, O.A., 1992. Cardiovascular differentiation of emotions. Psychosomatic Medicine 54(4), 422-435.

Sinha, R., Parsons, O.A., 1996. Multivariate response patterning of fear and anger. Cognition & Emotion 10(2), 173-198.

Smith, A.P.R., Stephhan, K.E., Rugg, M.D., Dolan, R.J., 2006. Task and Content Modulate Amygdala-Hippocampal Connectivity in Emotional Retrieval. Neuron 49(4), 631-638.

Soleymani, M., Chanel, G., Kierkels, J., Pun, T., 2008. Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses. IEEE International Symposium on Multimedia, Berkeley, US.

Stemmler, G., Heldmann, M., Pauls, C.A., Scherer, T., 2001. Constraints for emotion specificity in fear and anger: The context counts. Psychophysiology 38(2), 275-291.

Takahashi, K., 2004. Remarks on Emotion Recognition from Bio-Potential Signals. Proc. 2nd International Conference on Autonomous Robots and Agents, Palmerston North, New Zealand.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1(3), 211-244.

van den Broek, E.L., Schut, M.H., Westerink, J.H.D.M., van Herk, J., Tuinenbreijer, K., 2006. Computing emotion awareness through facial electromyography. Computer Vision in Human-Computer Interaction 3979, 52-63.

Vaughan, T.M., Heetderks, W.J., Trejo, L.J., Rymer, W.Z., Weinrich, M., Moore, M.M., Kubler, A., Dobkin, B.H., Birbaumer, N., Donchin, E., Wolpaw, E.W., Wolpaw, J.R., 2003. Brain-computer interface technology: A review of the second international meeting. IEEE Transactions on Neural Systems and Rehabilitation Engineering 11(2), 94-109.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech Communication 48(9), 1162-1181.

Wagner, J., Kim, J., André, E., 2005. From physiological signals to emotions: implementing and comparing selected methods for features extraction and classification. IEEE International Conference on Multimedia & Expo.

Wu, T.F., Lin, C.J., Weng, R.C., 2004. Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975-1005.

Xu, Z., John, D., Boucouvalas, A.C., 2006. Expressive image generation: Towards expressive Internet communications. Journal of Visual Languages and Computing 17(5), 445-465.

Zeng, Z., Tu, J., Pianfetti, B.M., Huang, T.S., 2008. Audio–Visual Affective Expression Recognition Through Multistream Fused HMM. IEEE Trans. on multimedia 10(4), 570-577.

Zhang, H., Malik, J., 2005. Selecting Shape Features Using Multi-class Relevance Vector Machine, EECS Department, University of California, Berkeley.

List of captions

Table 1: List of publications on emotion assessment from physiological signals. Signals acronyms are: Electromyography (EMG), Electrocardiogram (ECG), Galvanic Skin Response (GSR), Electroencephalography (EEG), Blood Volume Pulse (BVP).Classification acronyms are : Sequential Floating Forward Search (SFFS), Linear discriminant analysis (LDA), Mean Square Error (MSE), Multi Layer Paerceptron (MLP), K-Nearest Neighbors (KNN).

Table 2: features extracted from peripheral signals.

Table 3: Average confusion matrices across participants for peripheral features and different classifiers: LDA (a), Linear SVM (b), RBF SVM(c) and Linear RVM (d).

Figure 1: Emotion assessment in human computer interfaces, adapted from the execution / evaluation model (Norman, 1990).

Figure 2: (left) Different emotional classes in the valence-arousal space and their associated image; (right) schedule of the protocol and detail of a trial.

Figure 3: Complete process of trial acquisition, classification, fusion and rejection. Classification tasks acronyms are Calm Positive Negative (CPN), Calm Excited (CE), Negative Positive (NP), Calm Negative (NP) and Calm Positive (CP).

Figure 4: Mean classifier accuracy across participants for EEG STFT features and the different classification schemes. The bars on top of each column represents the standard deviation across participants. Classification tasks acronyms are Calm Positive Negative (CPN), Calm Excited (CE), Negative Positive (NP), Calm Negative (NP) and Calm Positive (CP).

Figure 5: Mean classifier accuracy across participants for EEG MI features and the different classification schemes. The bars on top of each column represents the standard deviation across participants. Classification tasks acronyms are Calm Positive Negative (CPN), Calm Excited (CE), Negative Positive (NP), Calm Negative (NP) and Calm Positive (CP).

Figure 6: Mean classifier accuracy across participants for peripheral features and the different classification schemes. The bars on top of each column represents the standard deviation across participants. Classification tasks acronyms are Calm Positive Negative (CPN), Calm Excited (CE), Negative Positive (NP), Calm Negative (NP) and Calm Positive (CP).

Figure 7: Mean classifier accuracy across participants for different modalities and their associated classifiers, as well as for fusion of the two EEG and the three physiological modalities. Classification tasks acronyms are Calm Positive Negative (CPN), Calm Excited (CE), Negative Positive (NP), Calm Negative (NP) and Calm Positive (CP).

Figure 8: Relation between the δ threshold value, classification accuracy and the amount of eliminated samples for the CPN (Calm Positive Negative) classification task.

Ref.	Number of part.	Elicitations	Time aspects	Signals / sensors	Emotion classes	Classifiers / classification	Best results
(Kim, 2004)	3 user specific	Music, AuDB (Augsburger database of bio-signals)	Time of a trial not specified 25 recordings over 25 days	Skin conductance, EMG, Respiration, ECG	Joy, anger, relaxation, sadness Positive / negative High / low arousal	SFFS feature selection, LDA with MSE SFFS feature selection, LDA with MSE SFFS feature selection, LDA with MSE	84% 84% 94%
(Lisetti and Nasoz, 2004)	29 not user specific	Film clips	70-231 s	GSR, heart rate, temperature	Sadness, amusement, fear, anger, frustration, surprise	Neural network with Marquardt backpropagation	84%
(Rainville et al., 2006)	43 not user specific	Self Induction	90 s	ECG, respiration, skin conductance, EMG (zygomatic, masseter, corrugator)	Anger (15 part.), fear (15 part.) happiness (15 part.), sadness (17 part.)	Step wise discriminant analysis	49%
(Picard et al., 2001)	1	Self induction	100 to 250 s 20 different days of recording	EMG, GSR, respiration, BVP, ECG	Neutral (no-emotion), anger, hate, grief, platonic love, romantic love, joy, reverence High / low arousal Positive / negative	SFFS-Fisher projection	81% 84% 87%
(Kim et al., 2004)	50 children not user specific	Combination of story telling, visualisation and audio stimulus	50 s	Skin temp Elctro dermal activity Heart rate	Sadness, anger, stress Sadness, anger, stress, surprise	Subjects for train and other for test SVM	78% 62%
(Wagner et al., 2005)	1	Music chosen by the participant	2 min 25 recordings over 25 days	EMG, ECG, GSR, respiration	Anger, sadness, joy, pleasure Valence Arousal	LDA, KNN, MLP SFFS, Fisher, ANOVA	92% 86% 96%

Ref.	Number of part.	Elicitations	Time aspects	Signals / sensors	Emotion classes	Classifiers / classification	Best results
(Haag et al., 2004)	1	images from IAPS	2 s several days	EMG, GSR, Skin temperature, BVP, ECG, respiration	Arousal Valence	Neural network for regression Accury is computed as the number of samples that fall in a 20% bandwidth of the correct value	97% 90%
(Sinha and Parsons, 1996)	27 not user specific	Self induction	60 s 2 recording sessions on different days	ECG, GSR, finger temperature, blood pressure, EOG, EMG (zygomatic, corrugator, masseter, depressor muscles)	Fear, Anger, neutral	LDA (first session as training set, second as test set)	67%
(Takahashi, 2004)	12 not user specific	Film clips	Time of a trial not specified	EEG, BVP, GSR	Joy, anger, sadness, fear, relaxation	Linear SVM one vs. all	42%
(Chanel et al., 2006)	4 user specific	images from IAPS	6 s	EEG, GSR, BVP, respiration, finger temperature	2 levels of arousal (low, high) 3 levels of arousal (low, medium, high)	Naïve-Bayes Naïve-Bayes	72% 58%
(Leon et al., 2007)	9 not user specific	images from the IAPS	6 s	Heart rate, GSR, blood volume pressure	Neutral, Negative, Positive	Autoassociative neural networks 1 participant for testing others for training	71%
(Sakata et al., 2007)	16	Pictures	3 s	EEG (results presented) heart rate	6 emotions	LDA	29%
(Rani et al., 2006)	15 user specific	solving anagrams playing pong	3-4 min 6 sessions on different days	ECG, GSR, bio-impedance, EMG (corrugator, zygomatic, trapezius), temp., BVP, heart sound	3 levels of intensity for: engagement, anxiety, boredom, frustration, anger Results are the average accuracy across participants and affective states	KNN Regression tree Bayes network SVM	79% 83% 78% 86%

Table 1:

Table 2:

Peripheral signal	Extracted features	Comments
GSR	Mean skin resistance over the whole trial	Estimate of general arousal level
	Mean of derivative over the whole trial	Average GSR variation
	Mean of derivative for negative values only	Average decrease rate during decay time
	Proportion of negative samples in the derivative vs. all samples	Importance and duration of the resistance fall
Blood pressure	Mean value over the whole trial	Estimate of general pressure
Heart rate	Mean of heart rate over the whole trial	-
	Mean of heart rate derivative Standard deviation of heart rate	Estimations of heart rate variability
Respiration	Power in the 0Hz-1.5Hz ($\Delta f = 0.25$Hz) bands (6 features)	-
	Mean respiration signal over the whole trial	Average abdomen expansion
	Mean of derivative over the whole trial Standard deviation	Variation of the respiration signal
	Maximum value minus minimum value	Dynamic range

Table 3:

(a)

Truth \ Classified	Calm	Excited
Calm	65%	35%
Excited	33%	67%

(b)

Truth \ Classified	Calm	Excited
Calm	31%	69%
Excited	11%	89%

(c)

Truth \ Classified	Calm	Excited
Calm	13%	87%
Excited	3%	97%

(d)

Truth \ Classified	Calm	Excited
Calm	33%	67%
Excited	10%	90%

Figure 1

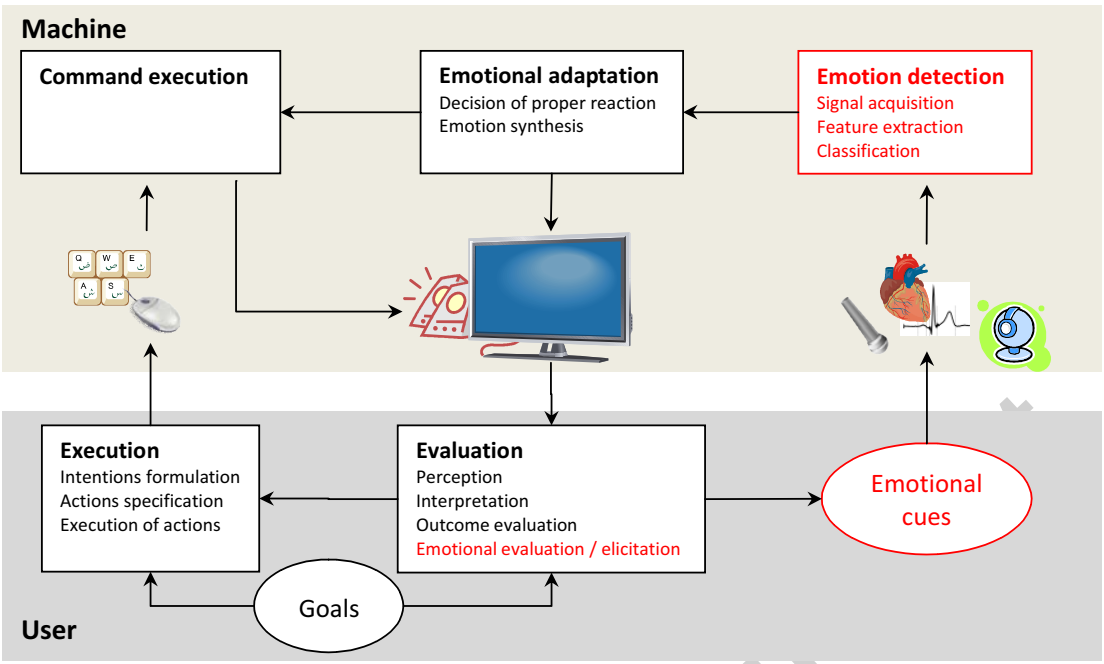


Figure 2

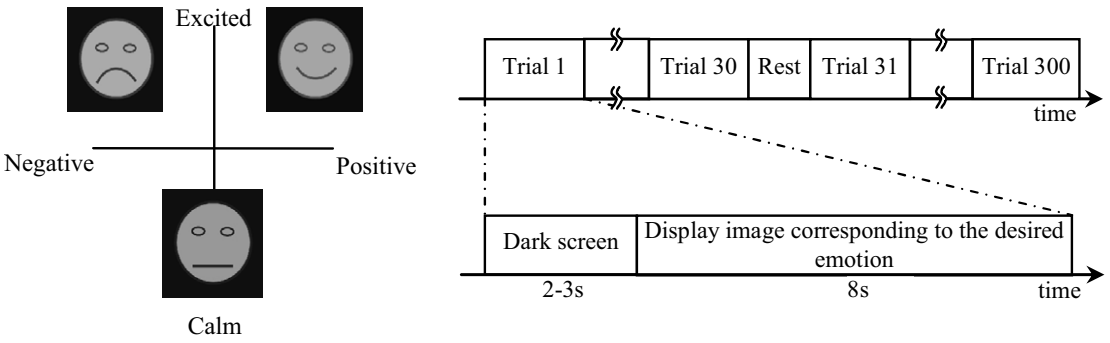
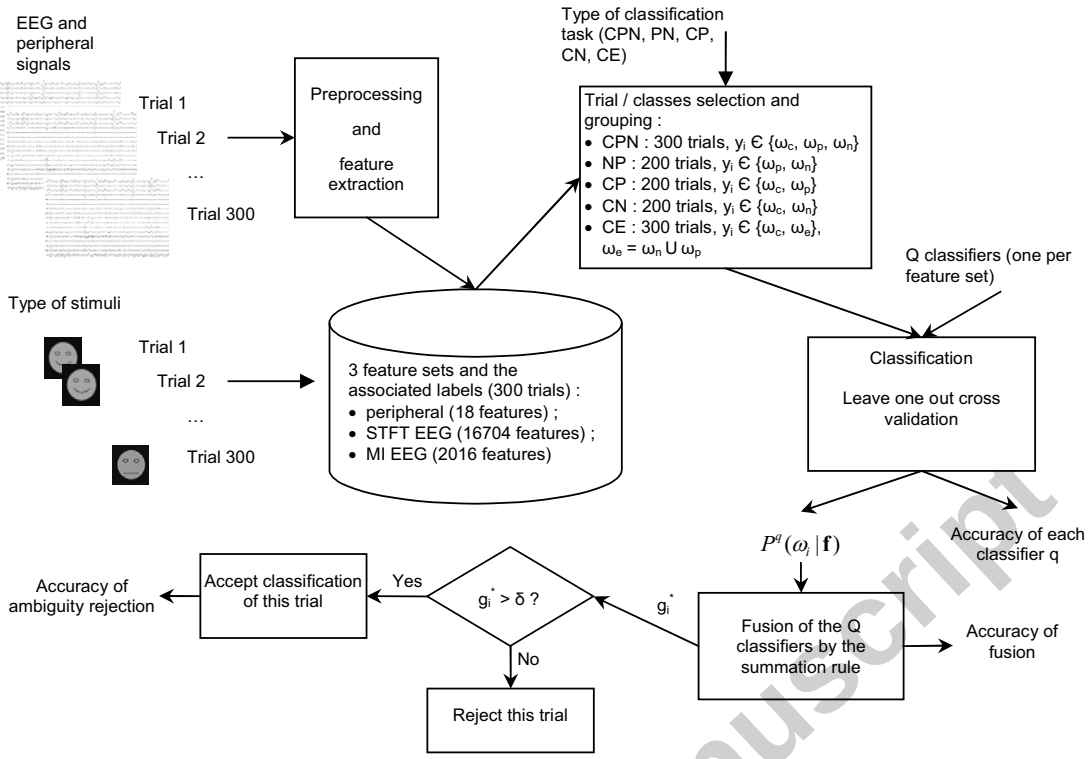


Figure 3



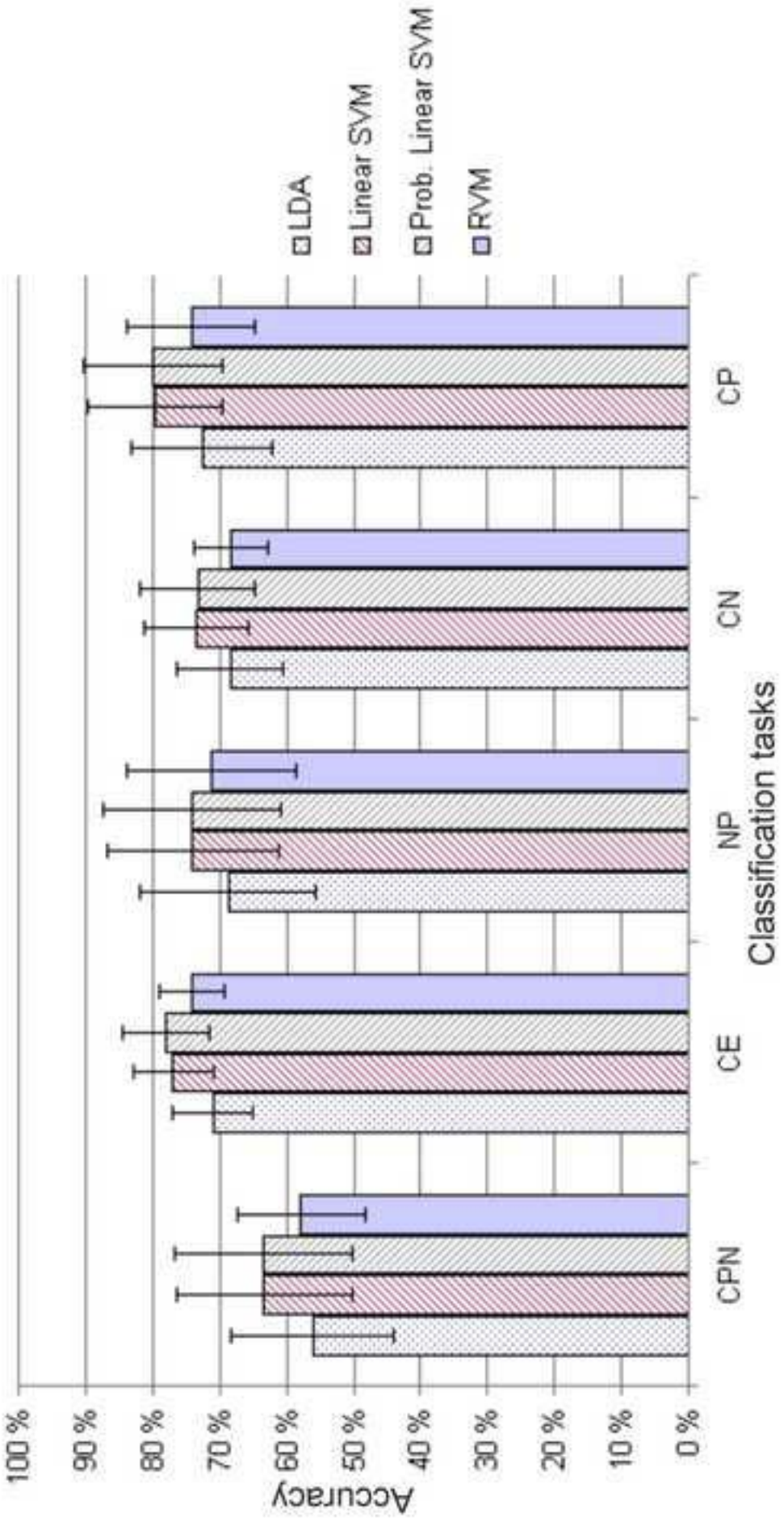

LDA
Linear SVM
Prob. Linear SVM
RVM
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
Accuracy
CPN
CE
NP
CN
CP
Classification tasks

Figure 4

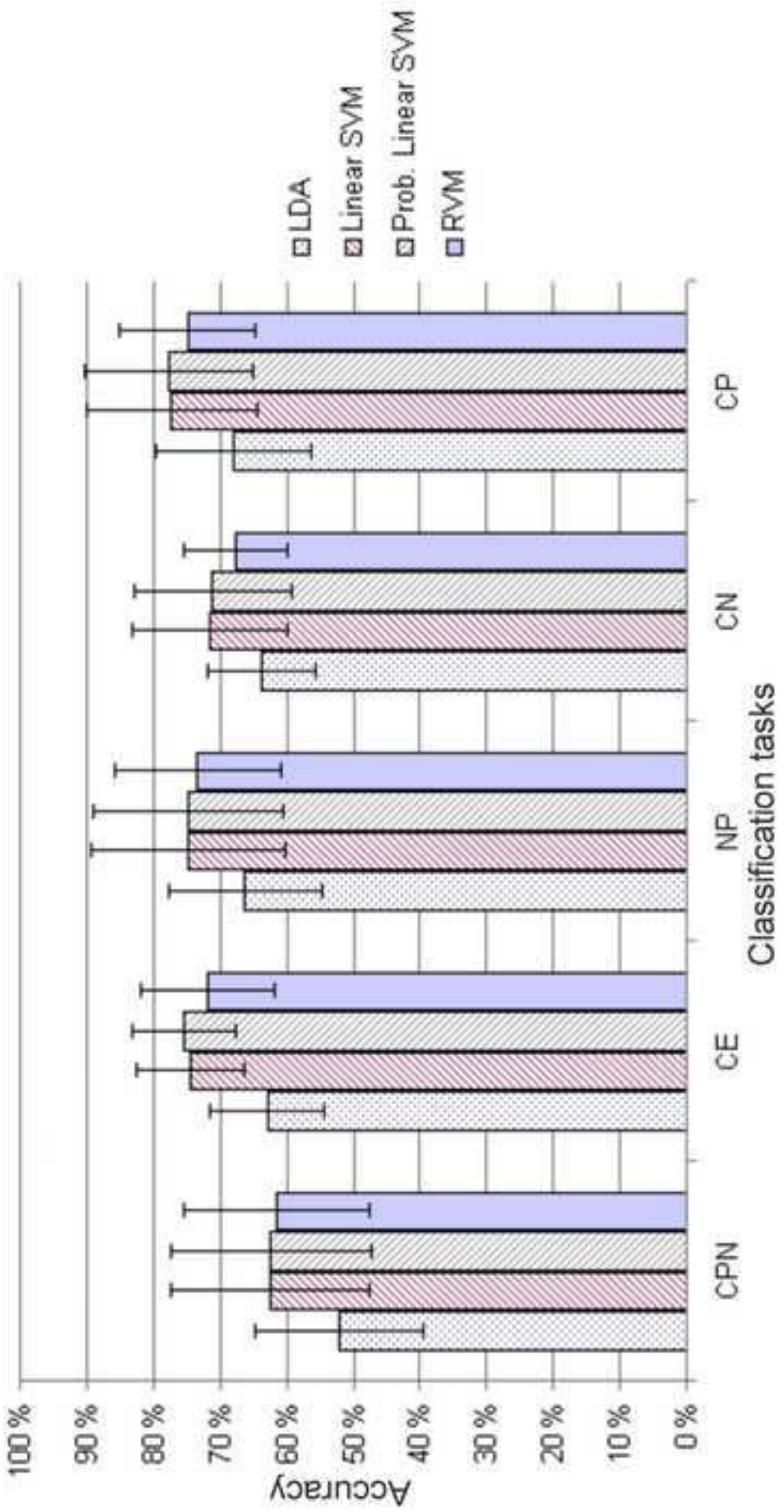

LDA
Linear SVM
Prob. Linear SVM
RVM
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
Accuracy
CPN
CE
NP
CN
CP
Classification tasks

Figure 5

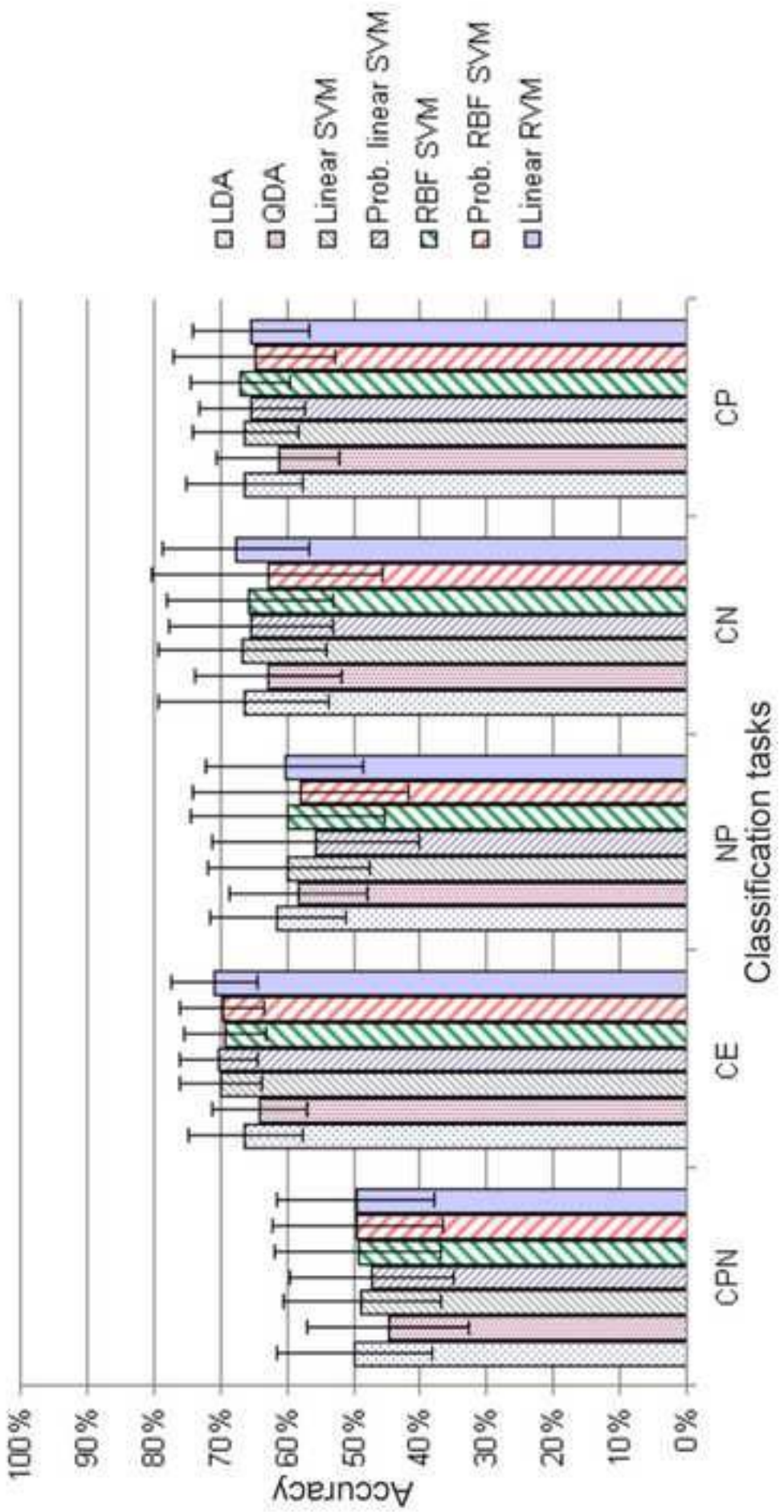

LDA
QDA
Linear SVM
Prob. linear SVM
RBF SVM
Prob. RBF SVM
Linear RVM
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
Accuracy
CPN
CE
NP
CN
CP
Classification tasks

Figure 6

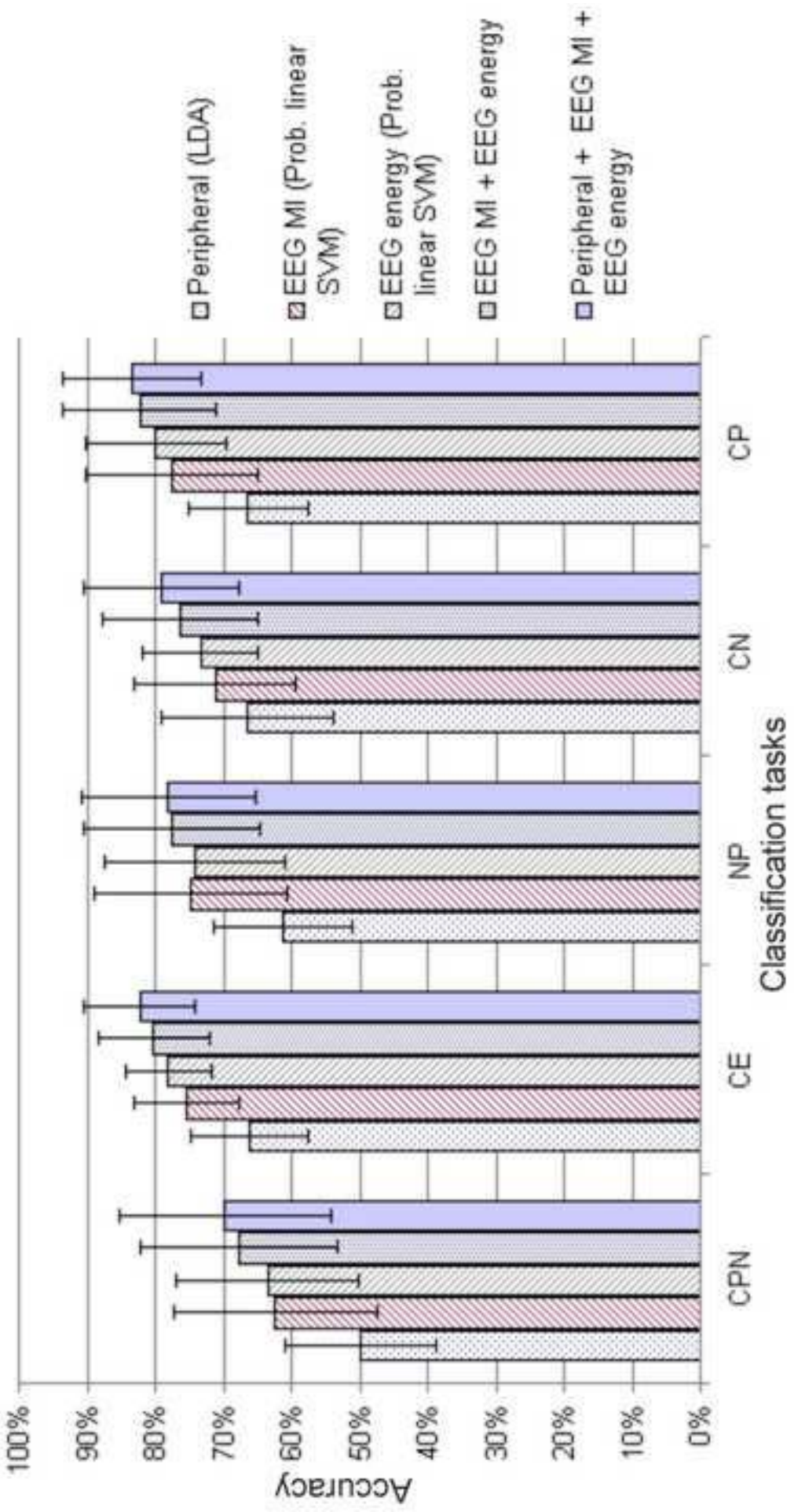

Peripheral (LDA)
EEG MI (Prob. linear SVM)
EEG energy (Prob. linear SVM)
EEG MI + EEG energy
Peripheral + EEG MI + EEG energy
Accuracy
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
CPN
CE
NP
CN
CP
Classification tasks

Figure 7

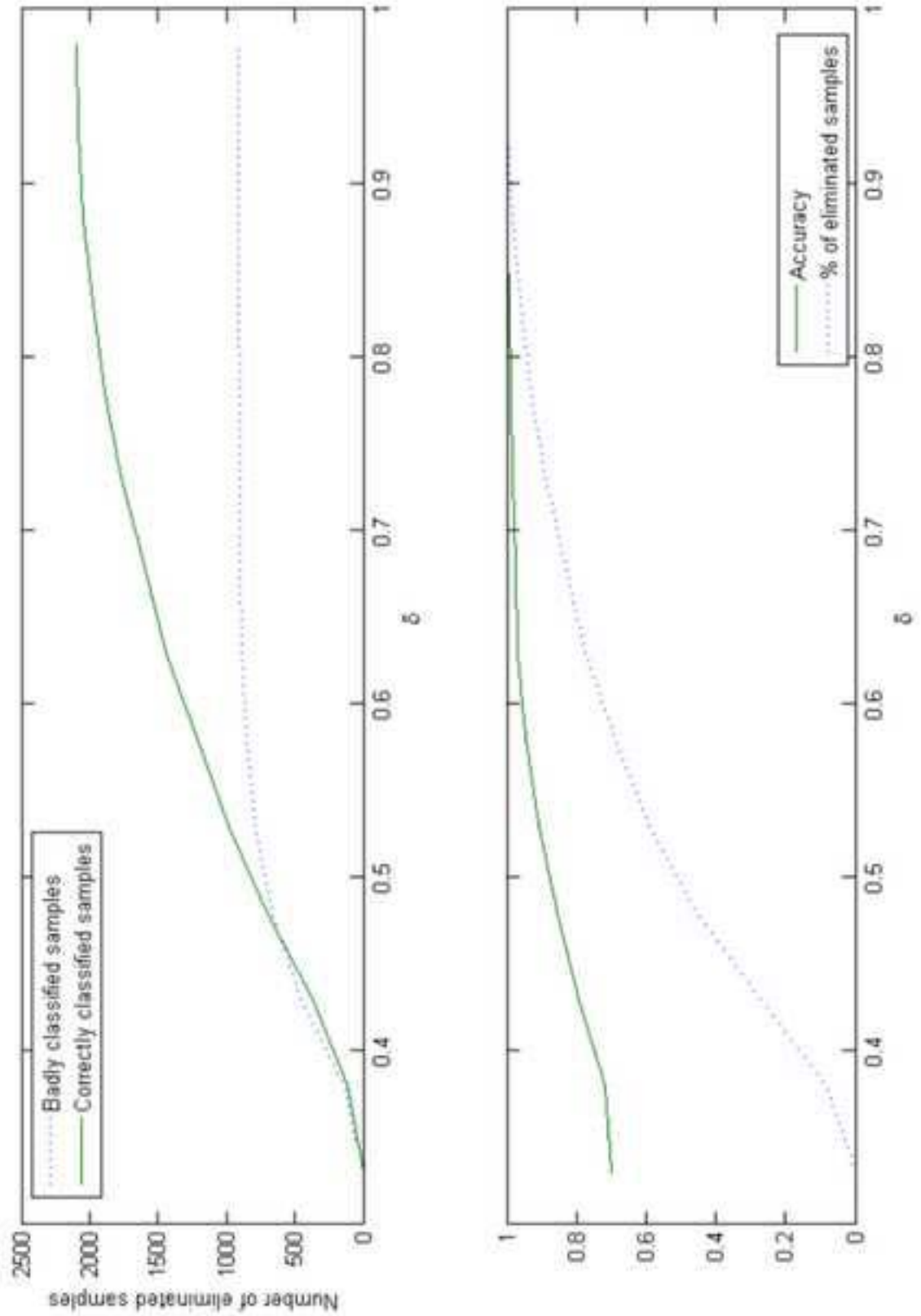

Badly classified samples
Correctly classified samples
Number of eliminated samples
δ
Accuracy
% of eliminated samples
δ

Figure 8