Chapter 2
Engineering Issues in Physiological Computing

Domen Novak

Abstract Prototypes of physiological computing systems have appeared in countless fields, but few have made the leap from research to widespread use. This is due to several practical problems that can be roughly divided into four major categories: hardware, signal processing, psychophysiological inference, and feedback loop design. This chapter explores these issues from an engineering point of view, discussing major weaknesses and suggesting directions for potential solutions. Specifically, some of the topics covered are: unobtrusiveness and robustness of the hardware, real-time signal processing capability, different approaches to design and validation of a psychophysiological classifier, and the desired complexity of the feedback rules. The chapter also briefly discusses the challenge of finding an appropriate practical application for physiological computing, then ends with a summary of recommendations for future research.

Introduction

Prototypes of physiological computing systems have appeared in countless fields, from critical applications such as stress and fatigue monitoring to home entertainment solutions such as physiology-based music selectors. However, despite the wealth of publications and fascinating prototypes, few physiological computing systems have made the jump from research to widespread use. We may wonder why this is so: is there no need for them or are they not yet ready for consumers?

While there is certainly a need for computers that could recognize and adapt to human psychological states, several practical problems have prevented physiological computing from achieving widespread use. In the author's personal experience, these issues are frequently raised by both engineers and potential

D. Novak (✉)
Sensory-Motor Systems Lab, ETH Zurich, Tannenstrasse 1, 8092 Zurich, Switzerland
e-mail: domen.novak@hest.ethz.ch

S. H. Fairclough and K. Gilleade (eds.), *Advances in Physiological Computing*, Human–Computer Interaction Series, DOI: 10.1007/978-1-4471-6392-3_2,

end-users, but do not yet have reliable solutions. They can be roughly divided into four major categories corresponding to the main components of a physiological computing system. In such a system, physiological data is first recorded by a sensor or range of sensors in next section. The raw data is then processed using algorithms that remove artefacts and extract potentially relevant features (Signal processing). A set of inference rules is used to convert the processed physiological data into an estimate of the user's psychological state (Inferring psychological states). Finally, the system acts on the inferred user state (Feedback loop) and the process begins again with a new recording.

This chapter explores engineering issues related to each of the above four components of a physiological computing system. It is not intended to cover all existing concerns, but to point out and describe some of the major problems that people may be unaware of as well as suggest directions for potential solutions. Since many people involved in physiological computing are either psychologists or computer scientists, we felt that an engineering perspective could be beneficial. This perspective attempts to be broad, progressing from electrical and mechanical engineering (hardware) to computer engineering and computer science (signal processing and machine learning) with a particular emphasis on evaluation and validation of all components. After covering the main specific topics, Finding the appropriate application explores the general challenge of finding an appropriate practical application for physiological computing. Finally, the last section summarizes the main issues and presents some recommendations for future research in the field.

Hardware

Physiological measurement equipment is the basic building block of physiological computing, and low-quality measurements invalidate the entire system. All sensors thus need to be carefully calibrated and validated. Commercial solutions are generally well-validated in laboratory conditions, but are geared toward researchers and often inappropriate for widespread use. As an alternative, ambulatory systems have been developed to measure physiological data in varied conditions such as walking and driving. The common feature of such systems is that they sacrifice accuracy in favour of increased unobtrusiveness. A, by now slightly outdated, list of ambulatory hardware was compiled by Ebner-Priemer and Kubiak (2007).

The Trade-Off Between Unobtrusiveness and Accuracy

The main weakness of research-grade physiological sensors is their obtrusiveness, as complex setups and controlled conditions are required to achieve good results. As an example, consider electroencephalography (EEG), which requires the

subject to wear a cap with electrodes. For proper measurement, the head needs to be measured and the cap needs to be properly applied to ensure proper electrode positioning. As it is often not clear which electrode sites are the most informative, researchers commonly use as many electrodes as possible. Furthermore, electrode gel is generally applied to improve signal quality and the electrooculogram (EOG) is measured to remove ocular artifacts from the EEG. The preparation time for an EEG measurement is thus around 30 min when using a cap with ~ 15 signal electrodes, a reference, ground, and EOG electrodes. Few people are willing to spend so much time applying a cap and ruining the appearance of their hair unless absolutely necessary.

If we wish to make the whole experience faster and more pleasant for the user, a number of things can be changed: the number of electrodes can be reduced, dry electrodes with no gel can be used, and the EOG can be omitted. All of these approaches have been implemented in consumer hardware. Devices such as the Emotiv EPOC allow EEG to be measured unobtrusively and at a far lower cost, but with also an obviously lower accuracy (Duvinage et al. 2012). The question is then: how much accuracy must we sacrifice to obtain a consumer-friendly device?

Research-grade sensors are usually made for a variety of possible situations. Consumer solutions are likely to be application-specific and can be made very specialized (as noted by e.g. Brunner et al. 2011). Some sensors can be made contactless: for instance, temperature could be measured using infrared cameras. Others can be built into the user interface or the surrounding environment. Lin (2011), for example, built their sensors into the steering wheel of a car while Wilhelm et al. (2006) built them into clothing. These sensors have great potential, but need to be validated to ensure that factors such as intermittent contact with the skin do not invalidate the measurements. Furthermore, real-time capability needs to be ensured, as sensors such as those developed by Wilhelm et al. (2006) only allow data to be stored on a memory card and analyzed later. While this is fine for initial research, practical applications require the data to be either wirelessly transmitted to a central computer in real time or analyzed with e.g. microcontrollers placed near the sensors.

Validating Ambulatory Physiological Sensors

Major progress has already been made in the validation of ambulatory equipment, especially dry and wireless EEG systems. Three approaches have commonly been used:

- An ambulatory system is used, and the study evaluates whether its accuracy is sufficient for a particular application (e.g. Berka et al. 2004).
- An ambulatory system is used together with a reference laboratory system, and the study evaluates whether the two systems infer significantly different psychological information such as stress level (e.g. Estepp et al. 2010).

- An ambulatory system is used together with a reference system, and the study evaluates whether the two systems record significantly different raw physiological signals such as ECG (e.g. Chi et al. 2012).

The third option is by far the most general, as guaranteeing high-quality raw data guarantees usability of an ambulatory system in a variety of applications. It can, however, be problematic for physiological computing developers, as it can be very difficult to ensure that the ambulatory and reference system are measuring the same data. For instance, as electrodes from two sensors cannot be placed in the exactly same spots at the same time, it is impossible to measure the exactly same signals even using identical sensors (Chi et al. 2012). Nonetheless, it has been shown that it is possible to measure raw EEG (Chi et al. 2012), respiration (Grossman, Wilhelm and Brutsche 2010) and ECG (Chi, Jung and Cauwenberghs 2010) using ambulatory sensors with practically the same accuracy as using laboratory hardware. Of course, this has only been demonstrated for some particular models of hardware.

Since validating the quality of raw data from ambulatory sensors is technically demanding, it would be optimally left to the manufacturer, who would publish evaluations of ambulatory hardware as compared to a reference, similarly to how traceability is performed in metrology. The evaluations would ideally be done in both ideal operating conditions (which, for an ambulatory system, are still worse than laboratory conditions) and poor operating conditions (e.g. many motion artefacts). With such hardware validation, physiological computing could utilize ambulatory systems without worrying about their performance.

When such data is not easily accessible, developers can nonetheless use ambulatory systems and compare the obtained information (extracted features or inferred psychological states) to either a reference device used in the same conditions or results obtained by other studies in similar situations. Such comparisons are useful not only to ensure acceptable accuracy, but also to ensure that a device is measuring the quantity it should measure. For instance, if stress in a task is correlated with physical activity, a poorly designed ambulatory sensor may actually measure motion artefacts and successfully infer increased stress from them despite not actually measuring any physiological processes.

Robustness

Many commercial sensors (ambulatory or not) are not at all robust. They are adversely affected by factors such as movement, temperature, humidity etc. These factors do not prevent the sensor from outputting a value; rather, they affect the output value either directly (e.g. motion artefacts cause electrode movement and thus incorrect readings of skin conductance) or indirectly via human physiology (e.g. increased environmental temperature causes sweating, increasing skin conductance for nonpsychological reasons). While indirect effects are not the fault of hardware, sensors do need to be made more robust to direct effects.

The primary problem are motion artefacts, which shift electrodes on the skin and cause incorrect readings. Even if the subject is perfectly still, artefacts can occur due to movement of the cables between the electrodes and the analog-digital converter. This can partially be compensated for by signal processing (Signal Processing), but not always. The motion artefacts in the skin conductance signal, for instance, can be very difficult to distinguish from actual skin conductance changes. Motion artefacts in the electrocardiogram (ECG) can be noticed easily, but are still difficult to remove since their frequency range partially overlaps the frequency range of the ECG.

Ambulatory sensors may be the solution for at least some environmental factors, as they are specifically built for robustness in rough conditions. Even researchers who are only interested in laboratory studies could thus benefit from ambulatory sensors. Other factors, however, will likely remain a problem (e.g. sensor output can vary with operating temperature). In such cases, physiological computing experts should keep up to date with discoveries in fields such as metrology and biomedical engineering while remaining aware of sensors' shortcomings and potentially compensating for them using signal processing (Signal Processing).

Standardization and Measurement Guidelines

Though physiological sensors can be made very application-specific, a certain degree of standardization would nonetheless allow easier implementation of practical solutions as well as allow results to be more easily compared between studies. First, this could involve very simple components. For example, Brunner et al. (2011) suggested the standardization of connectors between EEG caps and signal amplifiers. Similar steps could be taken for other sensors, making it easier and cheaper to build a complete system.

In addition to standardizing the components themselves, measurement procedures could also be standardized. As an example, let's look at skin conductance measurements, where several attempts at standardization have been made with no major success. Scarpa Scerbo et al. (1992) showed that the highest skin conductance values are obtained when measuring at the distal phalanges of the fingers. Nonetheless, many studies still place electrodes on the medial phalanges, proximal phalanges or palm of the hand. Proper preparation of the skin is also uncertain: Boucsein (2011) summarizes various preparation strategies, with no clear advantage of any method. This becomes doubly important since some skin conductance sensors (such as the g.GSR from g.tec Medical Engineering GmbH, Austria) are not meant to be used with electrode gel. This makes it very difficult to compare results from different studies.

Standardization could even be extended from low-level issues of obtaining a good signal to high-level issues such as obtaining good psychophysiological information. Strict standardization cannot be expected at higher levels since

different goals and conditions require different hardware and measurement procedures, but it should be possible to at least develop guidelines for particular goals. For instance, laboratory studies commonly use every available sensor to infer psychological states. Features are extracted from all the sensors, and machine learning algorithms are commonly used to determine which features are the most important (Novak et al. 2012). However, as the field moves toward downscaled and cheaper measurement solutions, it becomes necessary to know just which sensors are important in a given situation and which can be ignored. Researchers should thus, if possible, report which signals most contributed to psychophysiological inference. This can be done using the same machine learning methods used for psychophysiological inference. For instance, in our previous study, we used stepwise linear discriminant analysis to identify the features that had the largest effect on classification (Novak et al. 2011). Though we used a respiration sensor in the study, no respiration features had a large effect on classification, and a downscaled system for the same application could thus omit respiration altogether. Similarly, Wilson and Russell (2007) began their work on adaptive assistance using a full set of EEG electrodes, but later downscaled their system to five electrodes since they found them to be the most important.If researchers consistently report the most important features and sensors in their studies, a meta-analysis would be able to produce guidelines on which sensors to use in which situations (e.g. mental workload assessment in office tasks, fun assessment in physically demanding tasks), allowing application-specific sensors to be made simpler and cheaper with little loss in accuracy.

Signal Processing

The raw data collected from physiological sensors must generally be processed before it can be used for psychophysiological inference. In general, this process consists of filtering the signals to remove irrelevant low- and high-frequency information, removing any noise (due to e.g. motion) and calculating psychophysiologically relevant features from the cleaned signals (e.g. band power from the EEG). The required methods are fairly well-known and for the most part not limited to physiological computing: ECG processing is based on decades of clinical ECG analysis while EEG processing uses essentially the same methods for both physiological computing and active brain-computer interfaces. Nonetheless, some issues still need to be addressed.

Real-Time Noise Removal

Physiological computing systems must be able to quickly detect and respond to changes in the inferred psychophysiological state. While physiological quantities

Table 2.1 Measures of heart rate variability calculated from the ECG over a 2-minute rest period when a single additional heartbeat is erroneously detected due to noise

	no noise	at 50 %	at 20 %
SDNN (% of true value)	100.0	123.6	136.5
RMSSD (% of true value)	100.0	116.1	143.5
pNN50 (% of true value)	100.0	100.4	99.0
HF power (% of true value)	100.0	226.7	4156.2
LF power (% of true value)	100.0	153.5	1319.2
LF/HF ratio (% of true value)	100.0	67.7	31.8

The additional heartbeat is added either halfway between two real heartbeats (column 'at 50 %') or at 20 % of the interval between two real heartbeats (column 'at 20 %')

can be measured in real time, real-time processing is significantly more challenging. First of all, the raw data often contains noise that is not at all related to the measured physiological response. Examples include motion artefacts, speech artefacts in respiration signals, eye artefacts in the EEG, and so on. While some of these can be removed using simple bandpass filtering, this is not always possible. For instance, the frequency bands of the EEG, electrooculogram and electromyogram all partially overlap, so bandpass filtering does not remove all eye and motion artefacts from the EEG (Vaughan et al. 1996).

As an example of how noise can affect measurements, let's look at a 2-minute ECG recording from a 25-year-old healthy subject resting without performing any activity. As a very small change, an additional peak is added between two consecutive R-peaks in the signal, simulating an extra heartbeat erroneously detected due to an artefact. It is first added exactly halfway between the two R-peaks, then 20 % of the distance from the first peak to the second. Standardized measures of heart rate variability are computed according to recommendations of the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996): standard deviation of NN intervals (SDNN), square root of the mean squared differences of successive NN intervals (RMSSD), percentage of differences of successive NN intervals greater than 50 ms (pNN50), total power in the high-frequency heart rate band (HF power), total power in the low-frequency heart rate band (LF power), and the ratio of LF and HF power. Results are shown in Table 2.1.

A single erroneously detected heartbeat can thus cause huge changes in calculated heart rate variability. Sufficiently large errors could be automatically detected even online. For instance, if no R-peak occurs for more than 2 seconds, this could automatically be declared an error by the preprocessing algorithm. More complex criteria have been evaluated for heart rate in a classic psychophysiological paper by Berntson et al. (1990). In a similar vein, during online feature extraction, it would be possible to set acceptable ranges for individual features. For instance, expected values of SDNN and RMSSD in a given population could be obtained from the literature, and any value outside this range would be automatically declared an error. A larger problem is presented by smaller errors (e.g. 30 %

increase), which may be incorrectly interpreted as a change in psychological state and cause inappropriate reactions by the computer.

Such small errors require more complex approaches to detect online. The first popular approach uses a secondary reference sensor that gauges the quality of the primary sensor's output. With EEG, for instance, it is common to detect noise due to eye movements by measuring the EOG. Signal processing algorithms can then remove noise from the EEG by using the EOG as a reference (Croft and Barry 2000). Similarly, motion artefacts can be detected using sensors such as accelerometers. A second popular approach uses processing methods such as principal or independent component analysis to remove artefacts without the need for an additional sensor. It has been successfully used to remove motion artefacts from ambulatory EEG (Gwin, Gramann, Makeig, and Ferris 2010) and ambulatory ECG (Wartzek et al. 2011) and thus has high potential, but has not yet seen widespread adoption in physiological computing.

As a final note, though modern real-time artefact removal algorithms are quite advanced, it is not unreasonable to expect occasional errors due to artefacts, and the physiological computing system should plan for this (for instance, by acting conservatively).

Feature Extraction

The features commonly extracted from physiological responses for the purpose of psychophysiological inference are relatively well-defined, with lists of common features available in the work of e.g. Kreibig (2010) for autonomic nervous system responses. Nonetheless, some problems remain. The chief problem again has to do with real-time use: how often should features be extracted and over what kind of time periods?

The frequency of feature extraction depends both on the needs of the study and the real-time processing capabilities. While the raw data must be recorded with a high sampling frequency, it is not necessary to calculate features with the same frequency if we only wish to perform psychophysiological inference every few minutes. It is also difficult with current hardware, as many physiological features (e.g. spectral analysis of heart rate or EEG) require significant computing power to calculate. In general, it seems most appropriate to perform feature extraction once per instance of psychophysiological inference.

This feature extraction should be performed over a time period ('window') spanning from a point in the past to the present moment. It is unclear, however, what the best length of the feature extraction window is. The upper bound is likely the time between instances of psychophysiological inference: since (we assume) an action is performed by the physiological computing system after each inference, measurements taken before the action should be irrelevant to the current state. Of course, we may wish to make the window even shorter. Immediately after an action is taken by the computer, the user is not in a steady state since he/she must

get used to the effects of the action. We could thus only include data from the steady state. There are also some theoretical considerations; for instance, some heart rate variability features should only be extracted from a steady-state period of at least 2 min (Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology 1996). Nonetheless, we also do not want to make the window too long, as the magnitude of physiological responses to a stimulus diminishes over time. Long windows may thus make it difficult to extract stimulus-related information from background noise. As some physiological signals respond more quickly to stimuli than others (EEG in less than a second, skin conductance in a few seconds, and skin temperature in up to a minute), different windows may be needed for different features.

Finally, regardless of real-time use, we should consider the definitions of the features themselves. Although most features are well-defined, some definitions seem somewhat arbitrary and stem from literature published before physiological computing ever got started. Consider, for example, the skin conductance signal. A common skin conductance feature is the number of skin conductance responses, which are defined as sufficiently large and rapid changes from the baseline value. A commonly used amplitude threshold for a skin conductance response is 0.05 microsiemens. But why this specific value? As Boucsein (2011) explains, this threshold originally largely depended on the skin conductance signal's expected range and amplification. Old recording devices with paper output did not use thresholds below 0.5 microsiemens, but values as low as 0.01 microsiemens have been suggested for modern sensors (Boucsein 2011). The 0.05 microsiemens value seems to be used today mainly because it is popular. However, given the myriad of possibilities regarding sensor placement, use of gel, sensor amplification, and filtering, all of which affect the range of the signal, it makes little sense to always use the same threshold. In fact, we may wonder whether counting the number of skin conductance responses itself may not be a relic. The practice originated in the time of recorders with paper output, when manual analysis was required, but in the era of personal computers it may be more sensible to use e.g. the central moments of the signal (i.e. variance, skewness or kurtosis).

The evaluation of different windows and the evaluation of potential new features are both better suited for basic psychophysiological research than for physiological computing, but an easy first step would be to identify a set of potential window lengths and/or new features, then calculate features on old data from several published studies and evaluate how well-correlated they are with psychological information.

Inferring Psychological States

Inferring the subject's psychological state from measured physiological responses represents a major challenge in physiological computing, and requires knowledge of both psychology and computer science. As stated by Cacioppo and Tassinary

(1990), connections between physiology and psychology are rarely one to one (a single psychological element affecting a single physiological response), but are more likely to be one-to-many (one psychological element affects many physiological responses) or many-to-one (many psychological elements affect one physiological response). Researchers have already raised the issue of whether physiological responses even contain enough psychological information to allow practical implementations of physiological computing (e.g. Fairclough 2009).

Engineers may be frustrated by the lack of standardized methods for the interpretation of psychophysiological responses. Asked about his dislike of physiological computing, a colleague with over a decade of engineering experience remarked:

> When I'm doing sensor fusion, I know that the Kalman filter is great. In control engineering, I always know the basic approaches to build on. But with psychophysiology, it seems like you start over with every new application.

To a degree, his concern is valid: due to the inherently subjective nature of psychophysiological responses, there can probably never be a 'standard' recipe for a physiological computing system. Nonetheless, at the moment it can be assumed that most physiological computing systems can share the same general structure and deal with the same issues. In their seminal work on psychophysiological inference, Picard et al. (2001) demonstrated several steps that have now become commonplace: feature extraction, dimension reduction and classification. A recent review (Novak et al. 2012) shows that most studies that perform data fusion with autonomic nervous system responses in psychophysiology still perform feature extraction and classification, with many incorporating dimension reduction. The used classifiers range from linear discriminant analysis to neural networks, but generally each study only uses a single static classifier (i.e. one that does not take temporal relations into account). Classification is usually performed on a prerecorded dataset that contains roughly equal numbers of physiological data examples from each possible class. While this approach is perfectly valid, it has several weaknesses.

Context-Awareness

Psychophysiological responses are affected by a huge number of confounding factors. Age, gender, disease, time of day, physical activity, external temperature, ingested substances (coffee, medicine…) and many other factors can completely obscure any physiological changes due to psychological factors. Laboratory studies generally try to control as many of these factors as possible, creating very artificial conditions that are not feasible in the real world (Wilhelm and Grossman 2010).

In real-world applications, we need to take into account both the nonpsychological context (temperature, physical activity…) as well as psychological context (situation-specific demands of enacting a given psychological state) of

physiological responses in order to accurately interpret them (Kreibig 2010; Wilhelm and Grossman 2010). As an illustration of how problematic this can be, Picard et al. (2001) analyzed the psychophysiological responses of a single subject who expressed eight emotional states daily over 20 days. They were able to classify the eight emotions with an accuracy of 81 %. They then attempted to classify the measured emotions according to the day they were evoked and were able to classify the day with an accuracy of 83 %. It is therefore easier to determine the day on which an emotion was expressed than the type of emotion from physiology! This is even more startling since there were 8 possible emotions and 20 possible days, making day classification much more challenging in principle.

Picard et al. (2001) also found that the subject's daily mood (long-term psychological state) affects what emotion (short-term psychological state) can be expressed and to what degree. Kreibig (2010) thus emphasized the importance of separating moods from emotions, but Wilhelm and Grossman (2010) noted that it is very difficult to capture mood alterations with physiology in most studies.

Luckily, many confounding factors can be measured using various sensors. As reviewed by Wilhelm and Grossman (2010), it is possible to measure e.g. physical activity using accelerometers, speech using respiration sensors, food intake using electronic diaries or circadian rhythms using clocks. The measured confounding factors could thus be included and accounted for in a sufficiently complex psychophysiological inference algorithm. This is what we refer to as context-awareness. The end goal should be to establish reliability of inference across a range of representative test conditions, test environments and individual differences (Fairclough 2009).

Unfortunately, psychophysiological studies only rarely distinguish between different contexts (Wilhelm and Grossman 2010), a problem that has also been noted in affective computing fields such as speech and gesture recognition (Zeng et al. 2009). This is not because the idea itself is new; the psychophysiological inference algorithm of Picard et al. (2001) already included multiple measures that should correspond to physical activity and circadian rhythms. Context-awareness does seem to be gaining popularity, especially in fatigue studies. Ji et al. (2006) and Yang et al. (2008) both combined physiological measurements of fatigue with user conditions (e.g. sleep quality, workload) as well as environmental conditions (e.g. weather). Nonetheless, context-awareness is in its infancy, and should represent a major avenue of new physiological computing research. In the beginning, context could represent only an additional input feature to the psychophysiological inference algorithm, but more complex approaches should be possible later. For instance, the context could represent the prior probability for a probabilistic inference algorithm, or the system could switch between different inference algorithms depending on the context.

Dynamic and Ensemble Classification Algorithms

A recent review of data fusion algorithms using autonomic nervous system responses in physiological computing shows that most existing algorithms are static single-level classification algorithms (Novak et al. 2012). Essentially, features from the current time period are input into a single static classifier that outputs the inferred psychological state among a limited number of possibilities. Even ignoring context-awareness (Context-awareness), this is a relatively simple approach.

First of all, a single static classifier ignores the dynamic nature of physiological measurements and emotions by treating each measurement as independent of previous ones. However, current cognitive workload, for instance, is not independent of cognitive workload felt a few minutes ago, and current skin conductance is not independent of recent skin conductance. While most physiological features are averaged over a period of time (up to a few minutes), thus reducing temporal relations in the data, dynamic classifiers that take temporal relations into account could potentially increase accuracy. This is especially important since different physiological signals have different response times to stimuli, so information from different periods of time should be taken into account.

The most promising dynamic classifiers are dynamic Bayesian networks. Kalman filters, which are commonly used in general sensor fusion and have been shown to improve psychophysiological inference with autonomic nervous system responses by 'learning' about a subject over time (Koenig et al. 2011; Novak et al. 2011), are theoretically a simple dynamic Bayesian network, though not well-validated in physiological computing. More advanced dynamic Bayesian networks have been tested, some of them incorporating context-awareness (Ji et al. 2006; Lee and Chung 2012; Yang et al. 2008). Besides Bayesian networks, alternate dynamic classifiers include e.g. Long Short Term Memory recurrent neural networks (Wöllmer et al. 2011).

In addition to dynamic classifiers, one possibility that has remained relatively unexplored are ensemble classifiers: combining several classifiers (of the same type or different types) to obtain a final result. For instance, each measurement 'category' (autonomic nervous system, central nervous system, nonphysiological) could have its own classifier, and the outputs of the individual classifiers would then be combined to obtain the final result. This was performed, among others, by Chanel et al. (2009). Another possibility is the so-called decision cascading, where one classifier makes a rough first estimate and a second classifier then confirms or discards that estimate (e.g. Picot et al. 2012).

Both dynamic and ensemble classifiers could in principle improve inference accuracy, providing a more complex and intelligent inference algorithm. However, they are also likely to require a large amount of training data, which is not always available in physiological computing. They thus need to be properly compared to simple classifiers in order to evaluate their effectiveness.

Detecting Brief Critical States

Most psychophysiological inference algorithms assume that all psychological states are equally probable. This is often true in laboratory studies, but not in the real world. When trying to detect stress and fatigue during driving, for instance, the vast majority of the measured psychological states would involve normal driving, and a small number of brief high-stress or high-fatigue periods would need to be detected. This challenge was mentioned as early as Picard et al. (2001).

The same problem is well-known in gesture and speech recognition, where the majority of recordings consist of 'garbage' and actual events occur only briefly. Hidden Markov models, for instance, can try to deal with the problem using a 'garbage' model where one possible class is dedicated specifically to various types of meaningless measurements (Bernardin et al. 2005; Wilpon et al. 1990). For physiology, Kreibig (2010) suggested tackling the problem by first using unspecific physiological responses (which distinguish between neutral and nonneutral conditions) to detect periods of interest, then using specific physiological responses to determine the exact psychological state experienced—a type of ensemble classification. Wilhelm and Grossman (2010) similarly suggested using abrupt changes in physiology to detect nonneutral conditions. Nonetheless, the problem is currently unsolved and likely requires the development of new types of inference algorithms.

Inference Validation

Once we have created a system capable of inferring a person's psychological state, we need to test it and see if it works correctly. For this, it is necessary to compare inferred psychological states to another, reference measurement that we assume is correct. Possibilities include self-report questionnaires, observable behavior, or simply using standardized stimuli that are expected to always induce the same psychological state. All of these have their own weaknesses.

Self-report questionnaires measure conscious processes only, while physiological responses are based on both conscious and unconscious processes. Thus, conditions may arise when subjects are unaware of their psychological states despite observable physiological and behavioral indicators (Fairclough 2009; Kreibig 2010). This is especially likely in subjects who are not healthy young adults, such as severe stroke victims (Koenig et al. 2011). Analysis of observable behavior provides an alternative, but psychophysiological changes can occur in the absence of any corresponding expression of overt behavior (Fairclough 2009). Finally, induction of psychological states using standardized stimuli such as media or standard tasks is very context-specific and does not generalize well (Fairclough 2009). Frustration induced with an extremely difficult mental arithmetic task, for instance, may not evoke the same physiological responses as frustration due to a

traffic jam. At this time, finding appropriate reference measurements is likely to remain an application-specific affair. Some engineers may actually be content not to separately validate the inference part and simply consider the system a success if it accomplishes its overall goal such as higher performance or user satisfaction.

Assuming that we find a reference measure, we can then calculate the accuracy of the psychophysiological inference: the percentage of times that the psychological state inferred by physiology is the same as the state inferred by the reference measurement. Accuracy remains the primary and often only quantitative way to validate psychophysiological inference (as reviewed by Novak et al. 2012). However, it is not always clear what the target accuracy in a certain setting should be. With active brain-computer interfaces, which are closely related to physiological computing, users were found to expect and accept approximately 75 % accuracy in recognition of four possible desired movements (Ware et al. 2010), but the finding is very application-specific.

In critical situations such as driver fatigue monitoring, physiological computing systems should be very accurate, as any mistake would either cause harm (potential problem not detected) or annoy the user (alarm or automated assistance engaged inappropriately). In casual applications such as computer games where the difficulty is regularly adjusted, accuracies of around 70 % for a two-class problem (increase/decrease difficulty) may be acceptable since the general trend would lead the player toward the optimal difficulty given enough time. On the other hand, such low accuracies would likely not be very useful since the same level of information can likely be obtained from simple performance measures. In any case, physiological measurements should always be compared to other (non-reference) measures to determine whether they can provide sufficient accuracy both on their own and when combined with other sources of information.

Of course, overall accuracy is not the only important factor in validation. Accuracy is frequently described using confusion matrices, which state how often each particular psychological state is misclassified as a different one (e.g. Healey and Picard 2005). This can help us better evaluate the practical usefulness of the classifier. For instance, if we have three possible levels of stress (low, medium and high), we may accept a classifier that regularly confuses low and medium stress, but always correctly detects high stress, as we would only wish to react to high stress. Other measures such as confidence values are also sometimes used together with accuracy (e.g. Picard et al. 2001), but not very frequently in physiological computing.

Finally, classification (which selects one psychological class out of many) is not the only possible method of psychophysiological inference. An admittedly less popular alternative is estimation, which uses methods such as linear regression or fuzzy logic (as reviewed by Novak et al. 2012) to output a continuous value of a particular psychological dimension (e.g. stress value 4.1 out of 10). This is intuitively useful when the goal of physiological computing is to adjust continuous variables such as the amount of lighting in a room or the speed of opponents in a computer game. However, the problem of estimation is that commonly accepted validation metrices practically do not exist in physiological computing. While it

should be possible to determine, for instance, the mean squared error, variance, or bias of an estimator, this is rarely done in psychophysiological studies, and results are instead often described only qualitatively. An important exception is the work of Mandryk and Atkins (2007), who validate the accuracy of a fuzzy estimator using mean squared error and a questionnaire as the reference measure. Since estimation could represent a useful alternative to classification in specialized applications where we only wish to monitor one psychological dimension, this work could serve as the starting point for development of more advanced validation methods.

Feedback Loop

Once the psychological state of the user has been inferred, the physiological computing system must respond to undesirable user states, thus closing the biocybernetic feedback loop. Three broad categories of feedback exist: offering assistance to a frustrated user, adapting the level of challenge if the user is bored or discouraged by a task, or adding an emotional display to encourage positive and mitigate negative emotions (Gilleade et al. 2005). However, while the theory of physiological feedback is well-developed and many possible feedback stimuli have been identified, implementations have proven challenging. It appears to be unclear just when and how feedback should be provided in order to achieve the desired goals.

Feedback Complexity and Speed

We should first ask ourselves: in a given system, how many actions should the system have at its disposal to achieve the desired goals? A larger selection of possible actions could increase the potential precision and helpfulness of the system, as it could then respond to more specific issues or simply perform a variety of actions so that the user does not constantly experience the same feedback. However, it is questionable whether the specificity of psychophysiological inference is sufficient to allow more than a small number of user states to be reliably identified. Furthermore, adding more and more actions could make it harder to analyze the performance of the system, as it becomes unclear just which of many possible actions made a contribution to the user's psychological state. This is especially problematic since there is no guarantee that contributions are additive; performing two actions consecutively may have a wildly different effect than the sum of the effects of each individual action.

At the moment, a small number of discrete actions or a small number of continuous variables (e.g. game difficulty level) should be sufficient for physiological computing. If the developer of the system has time, it may be best to first

analyze the response of the user to each individual action and then to likely combinations of actions, leading to the best possible understanding of the system.

Once we have defined the possible actions the system can take, it is also necessary to determine how often feedback should be provided. This depends on the physiological measurements used, the application, and the actions themselves. Firstly, different physiological signals have different stimulus response times, from less than a second for EEG to more than ten seconds for peripheral skin temperature. This sets the upper boundary for feedback frequency: once an action is taken by the system, its effect should become visible in the physiological response before a new action is taken.

The feedback frequency should also take into account the intrusiveness and explicitness of the individual actions (Fairclough 2009; Ju and Leifer 2008). Very explicit actions should be taken only occasionally, as they may otherwise upset the users. Consider the example of changing the difficulty of the game: if the difficulty changes every ten seconds, users may become annoyed at the inability to enjoy a stable game experience. Conversely, if the system offers assistance every ten seconds, users may become upset as attention is drawn to their poor performance. Implicit, unobtrusive actions such as changing the lighting of the room can be taken more often, but have two disadvantages. First, they may not be able to evoke large changes in the user's psychophysiological state since they are by definition weaker than explicit actions. Second, even if implicit actions can evoke changes, they may be difficult to both design and validate due to their unobtrusive nature. An example is the application of Ritter (2011), which aims to improve task performance by subtly changing the visual appearance of items on the screen. While improved performance is shown, the entire system is basically a 'black box' and it is very hard to determine just what factors led to improved performance.

User- and Situation-Specific Feedback Rules

We have already mentioned that each person's physiological responses are unique and that psychophysiological inference should take into account factors such as age and gender. Similarly, the feedback loop should also take each person's characteristics into account, as two people in the same situation will not necessarily respond to the same stimulus in the same way. As an example, studies of socially assistive robotics have shown that, in stroke rehabilitation, some users will perform best when a robot provides nurturing statements ("I know it's hard, but it's for your own good!") while others will perform best when the robot provides challenging statements ("Oh come on, you can do it!") (Tapus et al. 2008).

The feedback loop should also consider previously taken actions. On one hand, this would allow the physiological computing system to learn what actions 'work' (as suggested by Serbedzija and Fairclough 2012). On the other hand, it would also allow it to better gauge what the situation is like, thus identifying actions that

should not be taken since a certain opposing action or a very similar (perhaps even the same) action was taken shortly beforehand.

Unfortunately, feedback in existing applications is generally limited to a handful of predefined rules independent of either user or situation (see Novak et al. 2012, for a review of feedback loops using autonomic nervous system responses). Exceptions do, however, already exist. One promising example is the work of Liu et al. (2008), where players need to throw baskets through a basketball hoop controlled by a robotic arm. The hoop is constantly moved in different directions according to the measured psychophysiological state. The movement rules themselves are gradually adapted as the robot tries different patterns and discovers the effect each pattern has on the current player's enjoyment of the game. This application shows the promise of context-awareness in physiological computing.

Since context-awareness is a major field of study in other fields, it should be possible to adapt many lessons on context-aware feedback for physiological computing. As the feedback rules themselves do not necessarily depend on how the user's psychological state is obtained, the same feedback rules could, for instance, be used with a system that infers cognitive workload from physiology and a system that infers cognitive workload from movement patterns. This perhaps makes the task easier than context-aware psychophysiological inference, where the effects of context on physiology are very specific. On the other hand, context-aware feedback rules may include a physiology-specific factor: the reliability of the psychophysiological inference itself. A physiological computing system could provide strong feedback when it is confident that the psychological state has been correctly identified while taking only minor actions or even explicitly querying the user when the inferred psychological state is uncertain. This was done by e.g. Gruebler et al. 2012, whose robot performs actions slowly when uncertain, giving the user more time to intervene. It could be a welcome approach to also partially dealing with psychophysiological inference issues, which are likely to remain problematic for quite some time.

Finding the Appropriate Application

In the previous four sections, we examined the main issues that physiological computing currently needs to address. While most of these will undoubtedly be dealt with in time, we should think about what applications would benefit from physiological computing in its current state. The purpose of this section is not to discourage research in applications with no immediate practical benefit, but rather to help researchers consider the presented topics in a more 'applied' manner.

It might seem almost trivial that any product will only be successful if it provides a useful service: improved performance, higher pleasure, or something else. But at the same time, it must be better at providing this service than alternative solutions. In physiological computing, it must essentially provide the computer with enough additional information to justify the added cost and the

obtrusiveness of the sensors. Furthermore, this additional information should not be more easily obtained by other, nonphysiological means. End-users are usually very aware of this, leading to questions such as:

> But why would I want to wear a cap and gel on my head just for that?
>
> Why don't I just tell the computer what I want myself?

As an example, consider the physiology-guided music selector, which has been explored by numerous authors (the first being Healey et al. 1998). The premise is that the device selects an appropriate song for the listener based on the current psychological state, which is inferred from various physiological measurements (usually skin conductance and/or heart rate). However, we might ask whether an expensive music player with potentially unreliable inference of psychological state would really be preferable to a simple music player where we can select the playlist ourselves.

Similarly, numerous studies have shown prototypes of games or tasks where the difficulty can be dynamically adjusted using physiological measurements to provide a moderate challenge to the user. As with the music player, we can ask whether expensive, obtrusive and potentially unreliable sensors are preferable to either letting users change the difficulty themselves or having the system change difficulty based on some measure of task performance. For example, the Director of the Left 4 Dead video games by Valve Software is an artificial intelligence that dynamically adjusts the challenge posed by the game in response to the players' performance, which is assumed to be proportional to their emotional intensity (Booth 2009). Physiological computing would thus be more appropriate for tasks where performance measures do not exist or are not necessarily connected to psychological state (e.g. a task where the user is likely to maintain high performance while becoming excessively stressed).

Critical situations such as fatigue monitoring in vehicles may be the most promising application of physiological computing at the moment. The potential costs of failing to detect fatigue are high, making the investment reasonable. A vehicle is unlikely to be used by a large amount of people, so it would be possible to tailor the data fusion algorithms to each user. Physiological sensors built into the seat and steering equipment (e.g. Lin 2011) could reduce the obtrusiveness of the system, and sensors already existing in vehicles such as speedometers and clocks could provide context information. Less critical applications such as home entertainment are unlikely to see widespread use until physiological computing has been made less expensive, less intrusive, and better at performing psychophysiological inference. An expensive or obtrusive device will never be used by consumers, and an inaccurate one will serve more as a novelty than anything else.

Recommendations

Physiological computing currently faces many issues, which we roughly divided into four categories: hardware, signal processing, psychophysiological inference, and feedback design. At the moment, they are severe enough to hamper practical use of physiological computing in most applications, but they should not be thought of as insurmountable. A brief summary of practical steps that physiological computing experts and psychophysiologists in general can take today is as follows.

In hardware, the most important issue is the development and validation of ambulatory physiological sensors. It is crucial that the such systems are robust and properly validated with comparison to either laboratory hardware or previous literature. The field has advanced far enough that simply presenting a prototype is no longer sufficient; it should be proven to work so that future studies do not need to worry about low-level problems. At the same time, even researchers not working with ambulatory sensors can contribute by reporting which specific sensors and features were found to be most useful for a given situation. This would immensely assist hardware developers in 'pruning' unnecessary electrodes or sensors from ambulatory solutions. Finally, as hardware improves, manufacturers should consider greater standardization so that different devices can be used together more easily.

In signal processing, the most important issue is real-time artefact detection and removal. While offline methods are common, advanced online approaches have not yet seen widespread use in physiological computing. At the moment, physiological computing should adopt existing online approaches and test their effectiveness. A second important issue is the real-time extraction of features from raw data. Here, it may be beneficial to take data from several already published studies and explore different extraction methods (different window lengths, normalization approaches or even new features) to see how they affect psychophysiological inference.

Psychophysiological inference remains somewhat mired in the 'classic' laboratory approach of classifying different psychological state using a single static classifier with no regard for context or temporal trends. Especially context-awareness is a promising avenue of research that could provide useful knowledge not only for physiological computing but human-machine interaction in general. Continuous estimation of psychological dimensions should also be explored as an alternative to classification. Finally, there should be a major focus on algorithms that detect brief critical states, unlike the classic approach where all emotions are equally likely and the baseline is the only neutral state.

Feedback design is still in its infancy, and its practicality must be better explored. User experience studies are needed to determine just how complex feedback rules can be and how often feedback should be provided. At the moment, such studies could work with very basic feedback rules and evaluate not only how accurate the psychophysiological inference is, but also how satisfied the user is

with the system. As more complex feedback rules are designed, studies could evaluate whether the complexity is beneficial to the user. In particular, user-specific and context-aware feedback rules could have a great effect on the performance of a feedback loop.

While the above challenges do not cover every issue in physiological computing, they can hopefully serve as an overview of the more technical side of the field. As they are gradually overcome, physiological computing will mature and spread into different facets of everyday life. But even in its current state, physiological computing could already gain acceptance in applications such as fatigue monitoring, helping to popularize the field and pave the way for the future.

References

Berka C, Levendowski DJ, Cvetinovic MM, Petrovic MM, Davis G, Lumicao MN et al (2004) Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. Int J Hum-Comput Int 17:151–170

Bernardin K, Ogawara K, Ikeuchi K, Dillmann R (2005) A sensor fusion approach for recognizing continuous human grasping sequences using hidden Markov models. IEEE Trans Rob 21:425–430

Berntson GG, Quigley KS, Jang JF, Boysen ST (1990) An approach to artifact identification: application to heart period data. Psychophysiology 27:586–598

Booth M (2009). The AI Systems of Left 4 Dead. Artificial Intelligence and Interactive Digital Entertainment Conference at Stanford

Boucsein W (2011). Electrodermal Activity (2nd ed.)

Brunner P, Bianchi L, Guger C, Cincotti F, Schalk G (2011) Current trends in hardware and software for brain-computer interfaces (BCIs). J Neural Eng 8(2):025001

Cacioppo JT, Tassinary LG (1990) Inferring psychological significance from physiological signals. Am Psychol 45:16–28

Chanel G, Kierkels JJM, Soleymani M, Pun T (2009) Short-term emotion assessment in a recall paradigm. Int J Hum Comput Stud 67:607–627

Chi YM, Jung T, Cauwenberghs G (2010) Dry-contact and noncontact biopotential electrodes: methodological review. IEEE Rev Biomed Eng 3:106–119

Chi YM, Wang Y-T, Wang Y, Maier C, Jung T-P, Cauwenberghs G (2012) Dry and noncontact EEG sensors for mobile brain-computer interfaces. IEEE Trans Neural Syst Rehabil Eng 20:228–235

Croft RJ, Barry RJ (2000) Removal of ocular artifact from the EEG: a review. Neurophysiol Clin 30:5–19

Duvinage M, Castermans T, Dutoit T, Petieau M, Hoellinger T, De Saedeleer C, Seetharaman K, et al. (2012). A P300-based quantitative comparison between the emotiv Epoc headset and a medical EEG device. In: Proceedings of the 9th Iasted conference on biomedical engineering

Ebner-Priemer UW, Kubiak T (2007) Psychological and psychophysiological ambulatory monitoring. Eur J Psychol Assess 23:214–226

Estepp J, Monnin J, Christensen J, Wilson G (2010). Evaluation of a dry electrode system for electroencephalography: applications for psychophysiological cognitive workload assessment. In: Proceedings of the 2010 human factors and ergonomics society annual meeting, pp 210–214

Fairclough SH (2009) Fundamentals of physiological computing. Interact Comput 21:133–145

Gilleade, K., Dix, A., & Allanson, J. (2005). Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In: Proceedings of DiGRA

Grossman P, Wilhelm FH, Brutsche M (2010) Accuracy of ventilatory measurement employing ambulatory inductive plethysmography during tasks of everyday life. Biol Psychol 84:121–128

Gruebler A, Berenz V, Suzuki K (2012) Emotionally assisted human-robot interaction using a wearable device for reading facial expressions. Adv Robot 26:37–41

Gwin JT, Gramann K, Makeig S, Ferris DP (2010) Removal of movement artifact from high-density EEG recorded during walking and running. J Neurophysiol 103:3526–3534

Healey JA, Picard RW (2005). Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems, *6*, 156–166

Healey JA, Picard RW, Dabek F (1998). A new affect-perceiving interface and its application to personalized music selection. In: Proceedings of the 1998 workshop on perceptual user interfaces. San Francisco, USA

Ji Q, Lan P, Looney C (2006) A probabilistic framework for modeling and real-time monitoring human fatigue. Sys Man Cybern Part A Syst Hum 36:862–875

Ju W, Leifer L (2008) The design of implicit interactions: Making interactive systems less obnoxious. Des Issues 24:72–84

Koenig A, Novak D, Omlin X, Pulfer M, Perreault E, Zimmerli L, Mihelj M et al (2011) Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. IEEE Trans Neural Syst Rehabil Eng 19:453–464

Kreibig SD (2010) Autonomic nervous system activity in emotion: a review. Biol Psychol 84:394–421

Lee B-G, Chung W-Y (2012) Driver alertness monitoring using fusion of facial features and bio-signals. IEEE Sens J 12:2416–2422

Lin Y (2011) A natural contact sensor paradigm for nonintrusive and real-time sensing of biosignals in human-machine interactions. IEEE Sens J 11:522–529

Liu C, Conn K, Sarkar N, Stone W (2008) Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE Trans Rob 24:883–896

Mandryk RL, Atkins MS (2007) A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int J Hum Comput Stud 65:329–347

Novak D, Mihelj M, Munih M (2012) A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. Interact Comput 24:154–172

Novak D, Mihelj M, Ziherl J, Olenšek A, Munih M (2011) Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation. IEEE Trans Neural Syst Rehabil Eng 19:400–410

Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans Pattern Anal Mach Intell 23:1175–1191

Picot A, Charbonnier S, Caplier A (2012) On-line detection of drowsiness using brain and visual information. IEEE Trans Syst Man Cybern Part A Syst Hum 42(3):764–775

Ritter W (2011) Benefits of subliminal feedback loops in human-computer interaction. Adv Hum-Comput Interact 2011:346492

Scarpa Scerbo A, Freedman LW, Raine A, Dawson ME, Venables PH (1992) A major effect of recording site on measurement of electrodermal activity. Psychophysiology 29:241–246

Serbedzija N, Fairclough SH (2012). Reflective pervasive systems. ACM Transactions on Autonomous and Adaptive Systems *7*(1), article 12

Tapus A, Tapus C, Matarić M (2008) User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. Intel Serv Robot 1:169–183

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996) Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Eur Heart J 17:354–381

Vaughan TM, Wolpaw JR, Donchin E (1996) EEG-based communication: Prospects and problems. IEEE Trans Rehabil Eng 4:425–430

Ware MP, McCullagh PJ, McRoberts A, Lightbody G, Nugent C, McAllister G, Mulvenna MD et al. (2010). Contrasting levels of accuracy in command interaction sequences for a domestic

brain-computer interface using SSVEP. 5th Cairo international biomedical engineering conference, pp 150–153

Wartzek T, Eilebrecht B, Lem J, Lindner H-J, Leonhardt S, Walter M (2011) ECG on the road: robust and unobtrusive estimation of heart rate. IEEE Trans Biomed Eng 58:3112–3120

Wilhelm FH, Grossman P (2010) Emotions beyond the laboratory: theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. Biol Psychol 84:552–569

Wilhelm FH, Pfaltz MC, Grossman P (2006) Continuous electronic data capture of physiology, behavior and experience in real life: towards ecological momentary assessment of emotion. Interact Comput 18:171–186

Wilpon JG, Rabiner LR, Lee C-H, Goldman ER (1990) Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Trans Acoust Speech Signal Process 38:1870–1878

Wilson GF, Russell C (2007) Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Hum Factors 49:1005–1018

Wöllmer M, Blaschke C, Schindl T, Schuller B, Färber B, Mayer S, Trefflich B (2011) Online driver distraction detection using long short-term memory. IEEE Trans Intell Transp Syst 12:574–582

Yang G, Lin Y, Bhattacharya P (2008) Multimodality inferring of human cognitive states based on integration of neuro-fuzzy network and information fusion techniques. EURASIP J Adv Signal Process 2008:371621

Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 38(1):39–58