Significance of Pupil Diameter Measurements for the Assessment of Affective State in Computer Users

Armando Barreto[1], Jing Zhai[1], Naphtali Rishe[2], and Ying Gao[1]

Electrical and Computer Engineering Department[1] and School of Computer and Information Sciences[2]
Florida International University
Miami, FL 33174 USA

Abstract- **The need to provide computers with the ability to distinguish the affective state of their users is a major requirement for the practical implementation of Affective Computing concepts. The determination of the affective state of a computer user from the measurement of some of his/her physiological signals is a promising avenue towards that goal. In addition to the monitoring of signals typically analyzed for affective assessment, such as the Galvanic Skin Response (GSR) and the Blood Volume Pulse (BVP), other physiological variables, such as the Pupil Diameter (PD) may be able to provide a way to assess the affective state of a computer user, in real-time. This paper studies the significance of pupil diameter measurements towards differentiating two affective states (stressed vs. relaxed) in computer users performing tasks designed to elicit those states in a predictable sequence. Specifically, the paper compares the discriminating power exhibited by the pupil diameter measurement to those of other single-index detectors derived from simultaneously acquired signals, in terms of their Receiver Operating Characteristic (ROC) curves.**

I. INTRODUCTION

New developments in human-computer interaction technology seek to close the communication gap between the human and the machine. A key component needed to meet these requirements is the ability of computer systems to address user affect. Picard and others have described the importance of the emotional and affective factors in human-computer interaction [1]. The knowledge of a user's affect can provide useful feedback regarding the degree to which a user's goals are being met, enabling dynamic and intelligent adaptation. Since physiological variables in humans are inherently controlled by their autonomic nervous system, these expressions of emotion are less susceptible to environmental interference or voluntary masking than others, such as, for example, facial expression or speech activity. Previous attempts to recognize emotions from physiological changes have analyzed a variety of autonomic activities such as the Electroencephalogram (EEG), the Electrocardiogram (ECG), the Electromyogram (EMG), Blood Pressure (BP), Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Skin Temperature (ST), Heart Rate Variability (HRV), etc. Many of these physiological variables have been chosen because they can be monitored in non-invasive and non-intrusive ways. However, one physiological variable that has not been studied extensively for the purpose of affect recognition is the pupil dilation. In an isolated fashion, it has been verified that variations of the Pupil Diameter (PD) reflect the emotional changes driven by auditory emotional stimulation [2].

From human physiology studies, it is known that the Sympathetic Division of the Autonomic Nervous System (ANS) significantly influences these physiological variables. The sympathetic division prepares the body for heightened levels of somatic activity. When fully activated, this division readies the body for a crisis that may require sudden, intense physical activity, which is known as the "fight or flight" response. Generally, an increase in sympathetic activity stimulates tissue metabolism and increases alertness. The heart rate, skin resistance, blood pressure and pupil diameter are all affected by branches of the sympathetic division of the ANS. In this study, we monitored four physiological variables (GSR, BVP, ST and PD) simultaneously and compared the significance of signals derived from these measurements towards the detection of sympathetic activation associated with a multifaceted emotional state — 'Stress'.

When a subject experiences stress and nervous tension, the palms of his/her hands become moist. Increased activity in the sympathetic nervous system will cause increased hydration in the sweat duct and on the surface of the skin. The resulting drop in skin resistance (increase in conductance) is recorded as a change in electrodermal activity (EDA), also called Galvanic Skin Response (GSR). So, in everyday language, electrodermal responses can indicate 'emotional sweating'. The GSR is measured by passing a small current through a pair of electrodes placed on the surface of the skin and measuring the conductivity level. In spite of its simplicity, GSR measurement is currently considered one of the most sensitive physiological indicators of psychological phenomena. GSR is also one of the signals used in the polygraph or 'lie detector' test. A GSR2 module, by Thought Technology LTD (West Chazy, New York) was used in our research to measure GSR. The resistance found in between its two electrodes determines the oscillation frequency of a square-wave oscillator inside the device. We have used a "frequency-to-voltage-converter" integrated circuit (LM2917N) to obtain output voltages that are proportional to instantaneous skin conductance. This modified device was calibrated by connecting several resistors of known resistance to it and measuring the output voltage of the frequency-to-voltage converter in each case.

K. Elleithy (ed.), Advances and Innovations in Systems, Computing Sciences and Software Engineering, 59–64.

The measurements of Blood Volume Pulse (BVP) in this project were obtained using the technique called photoplethysmography (PPG), to measure the blood volume in skin capillary beds, in the finger. PPG is a non-invasive monitoring technique that relies on the light absorption characteristics of blood. Traditionally, the Blood Volume Pulse has been used to determine the heart rate only. However, if measured precisely enough, it can be used to extract estimates of the heart rate and its variability. In our experiment, the sampling rate used to record the BVP signal was 360 samples/second.

Changes of acral skin blood flow are also a commonly used indicator for sympathetic reflex response to various stimuli. When sympathetic stimuli are applied to a person, the blood volume in the finger vessels is expected to decrease due to the vasoconstriction in the hairless areas of the hand but not in the hairy skin of the hand. If this assumption is true, the finger temperature should transiently decrease according to this effect. A thermistor can be attached to the subject's finger to sense the temperature changes. In our experiment, the subject's skin temperature was measured with an LM34 integrated circuit that provided a linear output between –50 and 300 degrees Fahrenheit. The output of the sensor was buffered and fed into a differential amplifier (with a gain of 31 V/V) to amplify the temperature changes in the range of 75-100 ˚F. This sensor was attached to the distal phalanx of the left thumb finger with the help of Velcro. The signal was recorded at the sampling rate of 360 samples/second. The experiments were performed in an air-conditioned room, to minimize the potential impact of environmental temperature changes on this experimental variable.

The diameter of the pupil is determined by the relative contraction of two opposing sets of muscles within the iris, the sphincter and dilator pupillae, and is determined primarily by the amount of light and accommodation reflexes [3]. The pupil of the human eye can constrict and dilate such that its diameter can range from 1.5 to more than 9mm. The pupil dilations and constrictions in humans are governed by the ANS. Several researchers have established that pupil diameter increases due to many factors. Anticipation of solving difficult problems, or even thinking of performing muscular exertion will cause slight increases in pupil size. Hess [4] indicated that other kinds of anticipation may also produce considerable pupil dilation. Previous studies also have suggested that pupil size variation is also related to cognitive information processing. This, in turn, relates to emotional states (such as frustration or stress) since the cognitive factors play an important role in emotions [5]. Partala and Surakka have found, using auditory emotional stimulation, that the pupil size variation can be seen as an indication of affective processing [2]. All these previous results found in the literature prompted us to attempt to use the pupil size variation to detect affective changes during human-computer interactions. There are several techniques available to quantify pupil size variations [5]. Currently, automatic instruments, such as infrared eye-tracking systems, can be used to record eye-related information, including pupil diameter and point of gaze. In our study, the subject's left eye was monitored with an Applied Science Laboratories series 5000 eye tracking system running on a PC computer to extract the values of pupil diameter. The sampling rate of the system was 60 samples/second. To minimize the potential impact of illumination changes on the subject's pupil diameter, the lighting of the experimental environment and the average brightness of the stimulus computer were kept constant during the complete experimental sequences and across all the subjects.

II. Methodology

A. Stress Elicitation

Our aim in this research is the detection of mental stress, as physical stressors occur far less frequently in the context of human-computer interaction. Therefore, in order to elicit mental stress at controlled intervals a computerized "Paced Stroop Test" was used. The Stroop Color-Word Interference Test [6], in its classical version, requires that the font color of a word designating a different color be named. In our research, the classical Stroop Test was adapted into an interactive version that requires the subject to click on the correct answer rather than stating it verbally. Since adding task pacing to the Stroop Test might intensify the physiological responses [7], each trial was designed to only wait 3 seconds for a user response. If the subject could not make a decision within 3 seconds, the screen automatically changed to the next trial. This modified version was implemented with Macromedia Flash® and also programmed to output bursts of sinusoidal tones through the sound system of the laptop used for stimulation, at selected timing landmarks through the protocol to time-stamp the recorded signals at those critical instants.

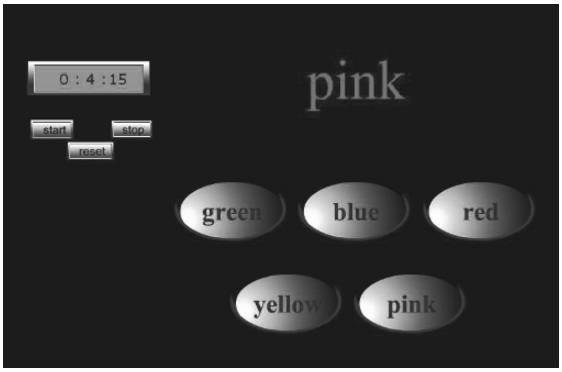

Figure 1. Sample Stroop Test interface

Figure 2 is the audio output schedule for the experiment, from the beginning of the session to its end. The complete experiment comprises three consecutive sections. In each section, we have four segments including: 1) 'IS' - the Introductory Segment to let the subject get used to the task environment, in order to establish an appropriate initial level for his/her psychological state, according to the law of initial values (LIV) [8]; 2) 'C' – is a Congruent segment, comprising 45 Stroop congruent word presentations (font color matches the meaning of the word), which are not expected to elicit significant stress in the subject; 3) 'IC' – is an Incongruent segment of the Stroop Test in which the font color and the meaning of the 30 words presented differ, which is expected to induce stress in the subject; 4) 'RS' – is a Resting Segment to let the subject relax for some time. The binary numbers shown in Figure 2 represent the de-multiplexed output of the audio signaling used in the system to time-stamp the four physiological signals, BVP, GSR, PD and ST. Our previous report on the instrumental setup [9] provides more details on this audio scheme.

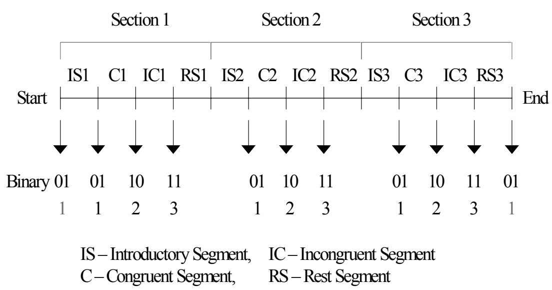

Figure 2. Audio output schedule.

B. Physiological Recording Setup

The complete instrumental setup developed for this research is illustrated in Figure 3. The stimulus program (interactive paced Stroop Test) described above runs in a laptop PC. While performing the Stroop Test, the subject has the GSR, BVP and ST sensors attached to his/her left hand. These three signals are digitized, using a multi-channel data acquisition system, NI DAQPad-6020E for USB, a product of National Instrumentation Corp, and the samples are read into Matlab® directly at rate 360 samples/sec. Additionally, the eye gaze tracking system (ASL-504) records PD data to a file on its own interface PC, at a rate of 60 samples/sec. The software for this system allows the extraction of selected variables (in this case the pupil diameter and the marker channel) to a smaller file, which in turn can be read into Matlab® also, where it can be aligned with the BVP, GSR and ST signals, thanks to their common timing marks for the start and stop events. At this point the pupil diameter data can be upsampled (interpolated) by six, to achieve a common sampling rate of 360 samples/sec for all four measured signals.

Figure 4 shows an example of the four signals recorded from a subject through the complete length of the experimental session, after all of the signals have been synchronized (at a sampling rate of 360 samples per second). The gaps in the pupil diameter signals, due to blinking, have been compensated by automatic interpolation.

Signals from 32 experimental subjects were collected and divided into 192 data entries, since each participant generated data under three relaxed (congruent Stroop) segments and three stressed (incongruent Stroop) segments.

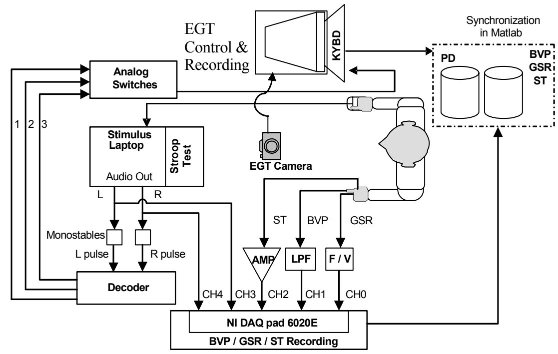

Figure. 3. .Instrumental Setup.

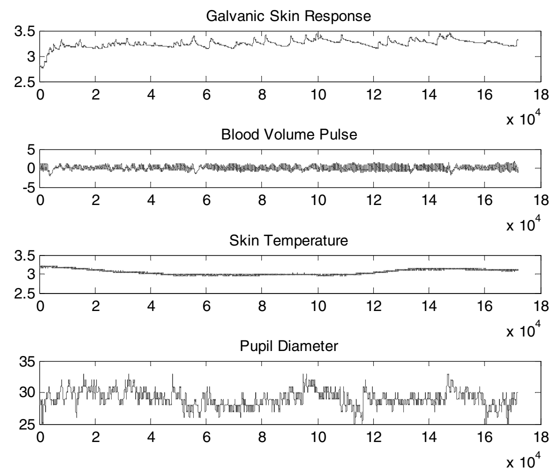

Figure. 4. Four physiological signals after synchronization. From top to bottom: GSR, BVP, ST, and PD.

C. Definition and Normalization of Individual Detection Signals

In this study our goal is to compare the potential of single-index indicators derived from the four physiological signals measured in terms of their individual discriminating power to differentiate between the congruent Strop segments (associated with a "relaxed" affective state in the user) and the incongruent Stroop segments (which are assumed to have caused a "stressed" state in the subject). The following paragraphs describe how the sample values of each of the signals were consolidated in a single feature value for each congruent or incongruent segment in the test.

The average value of the GSR samples collected during the whole extent of a congruent or incongruent Stroop segment was used as a representative response for this variable for each segment: *GSRmean*. Increased sweat production during "stressed" segments would predict a noticeable change of this average value during those segments.

From the BVP signal the interbeat interval (IBI), defined as the time in milliseconds between two normal, consecutive peaks in the BVP signal was defined for each two consecutive beats. The inverse of the IBI, expressed in beats per minute (BPM) is the heart rate, which is known to be altered by autonomic activation. The single index value defined from the BVP signal for each (congruent or incongruent) segment was the average of the IBI values in the segment: *BVPIBImean* .

For the skin temperature signal, it was expected that the temperature in the finger surface would display transient decreases when the stressor stimuli occur. To extract this information, the amplified ST signal was first filtered to remove recording noise. The slope of the filtered skin temperature in each segment was then used as a feature element of this signal. We found that the patterns of temperature slope provided more indicative information than the patterns of mean value of this signal. One possible explanation for this finding is that the skin temperature seems to obey much longer time constants in its variation, and, as such, its instantaneous value does not necessarily reflect well the "current" affective status of the subject, at any given time. In a protocol that included alternation between two types of stimuli (congruent and incongruent Stroop), the ST level during one given interval may still reflect the response to the previous interval. However, the derivative of the changing signal showed an interesting pattern. When the stressor stimuli occur, the slope of the temperature signal was generally negative. The slope of the ST signal was estimated using the digital low pass differentiation algorithm 1f3, as defined in [10], to yield the detection signal *STslope*.

The raw pupil diameter (PD) signal was recorded separately, as previously described. The artifact gaps due to blinking were automatically detected and filled by interpolation. The single-index signal extracted from the pupil diameter samples in each segment was simply the average value of PD, which we have labeled: *PD*. According to previous knowledge from the literature, we expected the mean PD should increase during the stress segments.

Prior to attempting to use these single-index signals to identify "stressed" (incongruent) and "relaxed" (congruent) experimental segments, they underwent a process of normalization. Let Xs represent the feature value for any of the raw features defined from the signal sample values during congruent and incongruent segments of the experiment. Let Xr represent the corresponding feature value extracted from the signals samples that were recorded during the relaxation period, prior to the first congruent Stroop segment. To eliminate the initial level due to the individual differences, Equation (1) was first applied to get the corrected feature signals (Y_s) for each of the subjects.

$$Y_s = \frac{X_s}{X_r} \tag{1}$$

For each subject, there were three congruent segments and three incongruent segments. Therefore, six values of any of the features were obtained from the signals recorded during these segments. Equation (2) normalizes each feature value dividing it by the sum of all six segment values.

$$Y_s' = \frac{Y_{si}}{\sum_{i=1}^{6} Y_{si}} \tag{2}$$

These two stages of normalization aimed at minimizing the impact of individual subject responses on the affective state identification process. After this pre-processing, all features (GSRmean, BVPIBImean, STslope and PD) were normalized to the range of [0, 1] using max-min normalization, as shown in Equation (3), to be considered as detection signals and compared against a threshold that spanned a uniform range of possible values: [0,1], for a fair comparison.

$$Y_{norm} = \frac{Y'_s - Y'_{s\min}}{Y'_{s\max} - Y'_{s\min}} \tag{3}$$

III. Results And Disscusion

A. Comparison of single detection signals

The four physiological signals monitored in our experiments are expected to exhibit different characteristics during the intervals when the subject was not under stress (i.e., during congruent Stroop segments) and during the intervals in which the subject was being stressed by incongruent Stroop word presentations. It is possible to summarize the information contained in each physiological signal by extracting one or several numerical features from each. In previous studies we have developed affective state classifiers that combine the information from several features extracted from each of the four physiological signals monitored, by means of machine learning systems [11][12][13].

In this study, however, our goal was to compare the discriminant power of information derived from the Pupil Diameter mean in a given interval (PD), with respect to other single detection signals GSRmean, BVPIBImean and STslope). Therefore, these three signals, as well as the PD measurements, were normalized as indicated by equations (1), (2) and (3).

B. Receiver Operating Characteristic (ROC) Curves

Next we present the comparison of the Receiver Operating Characteristic (ROC) curves for the four chosen normalized detection signals (PD, GSRmean, BVPIBImean and STslope), as a way to compare the levels of affective state discrimination power associated with them.

Receiver Operating Characteristic (ROC) curves show graphically the trade-off that a classifier must make between its "false positive rate" (which reflects the false alarm level, i.e., fraction of negative cases incorrectly classified as positive) and its "true positive rate" (i.e., the fraction of all positive cases correctly classified), by means of adjusting a threshold. The ROC is a plot of false positive rate vs. true positive rate for a classifier as different settings for the threshold are considered. At the lowest sensitivity level (i.e., setting the threshold at the highest possible value of the detection signal) the classifier produces no false alarms but detects no positive cases. This is represented by the origin of the coordinate axes in the ROC plot. As the sensitivity is increased, (i.e., as the threshold is lowered) the classifier detects more positive examples but may also start generating false alarms (false positives). Eventually the sensitivity may become so high (threshold set at the lowest possible value of the detection signal) that the classifier always claims each case is positive. So the classifier gets all positive cases right (true positive rate = 1), but it gets all the negative cases wrong, because it raises a false alarm on each negative case (false positive rate = 1). This corresponds to the top right-hand corner of the ROC. While all ROC curves "start" at the coordinate origin, (0,0) and "end" at the upper-right corner (1,1), the trajectory between these points followed by a given ROC, and consequently, the "Area under the ROC" are indicators of the discriminant power of the classification signal being thresholded. A "random classifier" (i.e., a process that produces uniformly distributed random numbers, without any relation to the input which is supposedly being classified) would display a ROC that follows approximately a 45^o diagonal ascent from (0,0) and (1,1). The "area under the ROC" (AUROC) would, therefore, be close to 0.5 (half the area of the 1.0-by-1.0 square). On the other hand, a classification system that produces a highly discriminating detection signal will have one or several threshold levels that map close to the upper-left corner of the ROC, at (0,1) indicating a high sensitivity (large true positive rate) and also a high specificity (low false positive rate). If that is the case, the AUROC will come close to encompassing the full 1.0-by-1.0 square. That is, the AUROC will approach the ideal value of 1.

C. ROC comparison for Pupil Diameter and other signals

In the light of the background provided by the previous sub-section, our interest is to compare the discriminant power of the Pupil Diameter with those of the other 3 normalized physiological measures chosen to represent each (congruent or incongruent) interval (GSRmean, BVPIBImean and STslope).

The ROC curve for each of these detection functions has been estimated using the values for the 6(segments) x 32(subjects) = 192 segments analyzed in our experiments. Only half of these segments correspond to "stressed" states induced by incongruent Stroop stimulation (ideal classifier output = 1), while the other half are known to be associated with "relaxed" (congruent Stroop) intervals (ideal classifier output = -1). Each point of the ROC curves is determined by comparing the value of the detection signal to a specific threshold level and determining which portion of the "positive classifier outputs" match the ideal (1) and which portion of the "negative classifier outputs" are in disagreement to the ideal output (-1). These "portions", expressed as fractional numbers, yield the coordinates of the ROC point (false positive rate, true positive rate) for the threshold value tested. The process was carried out using the ROC Matlab ® scripts provided by Dr. Gavin C. Cawley (University of East Anglia, Norwich, UK) in his web site http://theoval.sys.uea.ac.uk/matlab/default.html. These scripts not only sweep the complete range of normalized threshold values, [0,1], and draw the ROC, but additionally estimate a "convex hull" that fits the actual ROC points calculated. The convex hull is shown with dashed lines in the following plots.

Figures 5 through 8 show the ROC computed for the GSRmean, BVPIBImean, STslope and PD detection signals, respectively. The area under the ROC curve computed in each case is stated in the caption for each figure. It should be noted that both the ROC curves for GSRmean and for BVPIBImean show a convexity that makes them depart from the random classification diagonal to some extent. However, their areas under the ROC are only moderately better than 0.5: AUROC_GSRmean = 0.6519 and AUROC_BVPIBImean = 0.6455 .

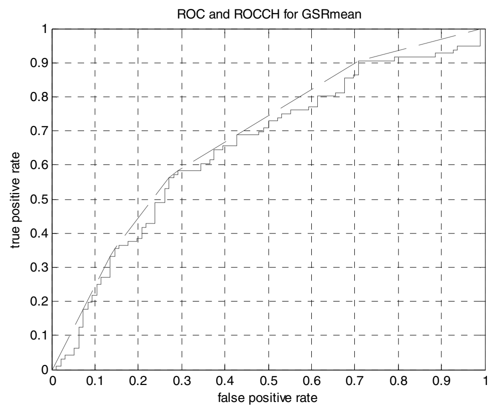

Figure 5. GSRmean ROC curve (AUROC = 0.6519)

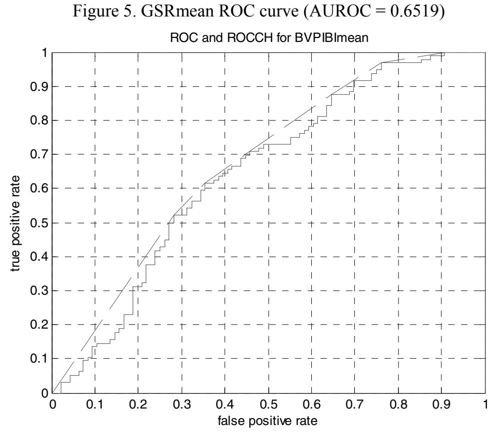

Figure 6. BVPIBImean ROC curve (AUROC = 0.6455)

The ROC curve for STslope shown in Figures 7 is actually very close to the random classification diagonal and, in general terms, follows a straight line at an angle just slightly larger than 45^o. As such, the area under this curve is not much higher than ½ : AUROC_STslope = 0.5849 .

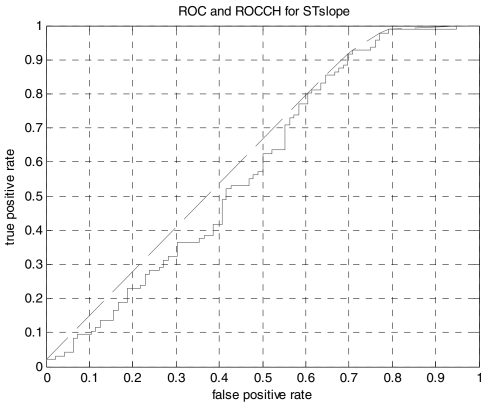

Figure 7. STslope ROC curve (AUROC = 0.5849)

In contrast, the ROC curve for PD, shown in Figure 8, exhibits a sharp slope from the coordinate origin, almost immediately reaching into high levels of true positive rate. Then the curve exhibits a number of intermediate points (threshold levels) for which the true positive rate is better than 0.8 while simultaneously having a false positive rate of less than 0.2. Accordingly, the area under this curve is large: AUROC_PD = 0.9647.

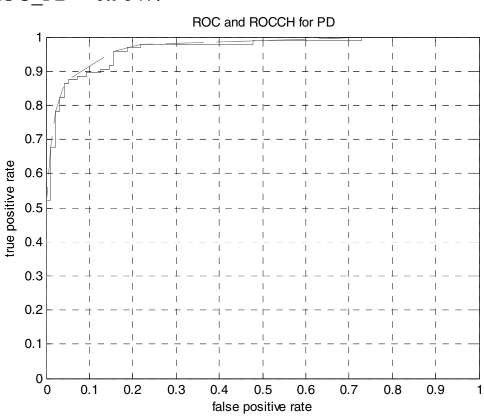

Figure 8. PD ROC curve (AUROC = 0.9647)

If a figure of merit indicating the *Discriminating Potential* of a given detection signal, X, calculated as:

$$DP = (AUROC(X) - 0.5) \times 200\% \quad (4)$$

is considered (such that a random classifier will yield 0% and a detection signal for which AUROC approaches 1.0 will yield approximately 100%), we would find that: DP(GSRmean) = 30.38%; DP(BVPIBImean)=29.10%; DP(STslope)=16.98% and, significantly, DP(PD)=92.94% . This indicates that, at least in terms of its ROC curve, the PD detection signal has significantly more potential to help identify one state from the other, while STslope shows particularly limited discriminating potential.

IV. Conclusion

We have investigated the potential of four detection signals derived from physiological measurements, GSRmean, BVPIBImean, STslope and PD, to act as individual classification signals for the differentiation between stress and relaxation in computer users. Our results indicate that two of the signals (GSRmean and BVPIBImean) derived from two of the physiological variables most commonly monitored for affective sensing exhibited only moderate discriminating potential. The STslope signal exhibited limited potential for this detection problem. In contrast, the mean value of the pupil diameter, PD, displayed a strong potential for single-signal discrimination between relaxed and stressed user states. Of course, this analysis has only addressed the *potential* of these signals for discrimination, and the actual performance of a detector based on any of these signals will be strongly influenced by the definition of the actual threshold used.

Acknowledgment

This work was partially sponsored by NSF grants IIS-0308155, CNS-0520811, HRD-0317692 and CNS-0426125.

References

[1] Picard, R.W. 1997. Affective computing. MIT Press, Cambridge, Mass.

[2] Partala, T., Surakka, V. 2003. Pupil size variation as an indication of affective processing. Int. J. of Human-Computer Studies 59:185-198

[3] Beatty, J., Lucero-wagoner, B. 2000. The Pupillary System. In: Handbook of Psychophysiology. J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson, editors. pp. 142-162. Cambridge University Press

[4] Hess, E.H. 1975. The tell-tale eye : how your eyes reveal hidden thoughts and emotions. Van Nostrand Reinhold, New York

[5] Grings, W.W., Dawson, M.E. 1978. Emotions and Bodily Responses A psychophysiological Approach. Academic Press, Inc.

[6] Stroop, J.R. 1935. Interference in serial verbal reactions. Journal of Experimental Psychology 18:643-661

[7] Renaud, P. and Blondin, J.-P., "The stress of Stroop performance: physiological and emotional responses to color-word interference, task pacing, and pacing speed," International Journal of Psychophysiology, vol. 27, pp. 87-97, 1997.

[8] Stern, R.M., Ray, W.J., Quigley, K.S. 2001. Psychophysiological Recording. Oxford University Press.

[9] Barreto, A. and Zhai, J., "Physiological Instrumentation for Real-time Monitoring of Affective State of Computer Users," WSEAS Transactions on Circuits and Systems, vol. 3, pp. 496-501, 2003.

[10] Usui, S. and Amidror, I., "Digital Low-Pass Differentiation for Biological Signal Processing," IEEE Trans. BME, 29: 686-693, 1982.

[11] Zhai J., and Barreto A., "Stress Detection in Computer Users Through Noninvasive Monitoring of Physiological Signals ", Biomedical Science Instrumentation, vol. 42, pp. 495-500, 2006..

[12] Zhai, J., and Barreto, A., "Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables", Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006, pp. 1355 – 1358.

[13] Zhai, J., and Barreto, A., "Stress Recognition Using Non-invasive Technology", Proceedings 19th Int. Florida Artificial Intelligence Research Society Conference (FLAIRS 2006), pp. 395 – 400, 2006.