Automatic stress classification with pupil diameter analysis

Marco Pedrotti*[1,5], Mohammad Ali Mirzaei[2], Adrien Tedesco[3], Jean-Rémy Chardonnet[2], Frédéric Mérienne[2], Simone Benedetto[4,5], Thierry Baccino[4,5]

[1]Université Paris 6 - Pierre et Marie Curie, 4 Place Jussieu, 75005, Paris, France
[2]Arts et Métiers ParisTech, CNRS, Le2i, Institut Image, 2 rue Thomas Dumorey, 71100 Chalon-sur-Saône, France
[3]Scientific Brain Training, 66 Boulevard Niels Bohr, BP 52132, F - 69603, Villeurbanne, France
[4]Université Paris 8 - Vincennes St.Denis, 2 rue de la Liberté, 93200, Saint-Denis, France
[5]CHART/LUTIN (EA 4004), Cité des Sciences et de l' Industrie de la Villette, 30 Avenue Corentin Cariou, 75930, Paris, France
*Corresponding author. *E-mail address :* marco.pedrotti@hotmail.it

This article proposes a method based on wavelet transform and neural networks for relating pupillary behavior to psychological stress. We tested the proposed method by recording pupil diameter and electrodermal activity during a simulated driving task. Self-report measures were also collected. Participants performed a baseline run with the driving task only, followed by three stress runs where they were required to perform the driving task along with sound alerts, the presence of two human evaluators, and both. Self-reports and pupil diameter successfully indexed stress manipulation, and significant correlations were found between these measures. However, electrodermal activity did not vary accordingly. After training, our four-way parallel neural network classifier could guess whether a given unknown pupil diameter signal came from one of the four experimental trials with 79.2 % precision. The present study shows that pupil diameter signal has good discriminating power for stress detection.

1. INTRODUCTION

Stress detection and measurement are important issues in several Human-Computer Interaction domains such as Affective Computing, Adaptive Automation, Ambient Intelligence. In general, researchers and system designers seek to estimate the psychological state of operators in order to adapt or re-design the working environment accordingly (Sauter, 1991). The primary goal of such adaptation is to enhance overall system performance, trying to reduce workers' psychophysical detriment (e.g. Czaja & Sharit, 1993; Dennerlein et al., 2003; Fujigaki & Mori, 1997). One key aspect of stress measurement concerns the recording of physiological parameters which are known to be modulated by the autonomic nervous system (ANS). However, despite technological progress in biological signal acquisition, inferring psychological significance from physiological signals is still a major challenge as biological signal analysis has progressed less intensively (Cacioppo & Tassinary, 1990), and it can be stated that affect recognition has not reached a satisfying level yet (Mauss & Robinson, 2009; Van den Broek et al., 2009).

This study describes a new method for stress measurement using pupil diameter (PD) signal analysis. It is well-known that pupillometry is a reliable tool for studying cognitive and emotional processes (Granholm & Steinhauer, 2004; Kuhlmann & Böttcher, 1999). The pupil is the aperture of the iris, the pigmented structure containing two antagonistic muscle groups - the sphincter and the dilator muscles. The former constrict the pupil; the latter dilate the pupil. The human pupil is known to reflect the activity of the autonomic nervous system: in particular, it has been shown that the pupil enlarges (mydriasis) as a consequence of mental effort exertion and various sources of psychological stress (see Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Bradley et al., 2008; Goldwater, 1972; Partala & Surakka, 2003; Vo et al., 2008). After dilation, the pupil naturally tends to constrict (myosis) back to previous diameter. Thus, we formulated the hypothesis that overall pupillary activity (i.e. pupil diameter oscillations) should be more intense under stressful conditions: phasic changes should follow stressful events. Moreover, mean signal amplitude should also increase, indicating higher tonic level.

However, the use of PD as a dependent variable in psychological studies has important methodological implications. The main concern stems from the primary function of the pupil itself, i.e. the regulation of the amount of light that enters the retina. The so-called *light reflex* occurs in order to avoid overexposure and retinal damage (Loewenfeld, 1993, p. 136). Such a constriction is rapid (latency within 250ms from stimulus onset) and proportional to stimulus intensity, and it is affected by individual differences. The return to pre-stimulus diameter (*dark reflex*) is much slower (see Beatty, 1986; Bergamin & Kardon, 2003; Ellis, 1981; Lanting et al., 1990). The *near reflex* (or accommodation response) - i.e. a near object

This research was supported by the French Minister of Economy, Finances and Industry, under contract n. F1105041V (MASSAI).
We thank the anonymous reviewers for their helpful comments.

requiring foveal focusing - causes pupil constriction, accompanied by eyes convergence and lenses curvature. While this phenomena could regulate human pupil size from 1 up to 10mm, very small (up to 0.01mm) pupillary dilations can be elicited by various psychological manipulations (see Beatty & Lucero-Wagoner, 2000): the *psychosensory reflex* denotes pupil dilations evoked by sensory stimulation, whether auditory, tactile, gustatory, olfactory or noxious. Beatty (1986) pointed out that this phenomenon is a sort of bridge between the well-understood *light* and *near* reflexes, and the more complex pupillary variations associated with cognitive processing. Many other factors can lead to pupil diameter variations: Janisse (1977) identified 23 sources of pupil variation, to which the effect of verbalization (Bernick & Oberlander, 1968) could be added. Therefore, extreme care should be taken in order to control as many potential bias sources as possible. Among these, illumination requires particular attention, although most studies fail to report such measurements.

Bearing these considerations in mind, there is a substantial interest and potential benefit in using PD for stress detection in applied studies: unlike other physiological measures (e.g. cardiovascular activity, electrodermal activity, electroencephalography, etc.), the pupil can be measured unobtrusively. With modern devices, one camera is sufficient for pupil tracking, and there is no need for physical contact. Remote eye trackers have such properties, and recent research has demonstrated their suitability for pupillometry studies (Klingner et al., 2008; Palinko et al., 2010, 2011, 2012). Besides PD, other eye-movement metrics (e.g. saccade parameters) are known to reflect stress-related variations (see Benedetto et al., 2011; Di Stasi et al., 2013). Efforts are also being made to unobtrusively measure skin temperature and other ANS measures with imaging techniques (see for example Nhan et al., 2010; Shastri et al., 2009).

The present study concerns stress detection in a simulated driving task. With the aim of validating our results by means of triangulation (Van den Broek et al., 2009), we recorded - besides PD - participants' self-reported stress levels and electrodermal activity (EDA, see section 3.5), a sensitive psychophysiological indicator of stress (Boucsein, 2012, p.459).

The article is organized as follows: existing PD data analysis techniques are introduced in section 1.1. Experimental set-up and hypotheses of the present study are described in section 2. Section 3 describes the whole process of data acquisition and analysis. Statistical results are presented in section 4. The framework of automatic stress detection based on Wavelet-Neural Network is outlined in section 5: Wavelet multi-resolution decomposition and Neural Networks will be used for feature extraction and classification respectively. Classification results and future challenges are discussed in section 6.

1.1 Methods for PD data analysis

Over the last decades, several methods have been used for analyzing PD data. The signals coming from pupil diameter recordings have been analyzed in both the time and frequency domains: state-of-the art is briefly reviewed in this section. Regardless of the method employed for the analysis, eye-blink artifacts represent a common problem in video-pupillography: most systems measure pupil size upon eye image processing (see Holmqvist et al., 2010; Wyatt, 2010). During eye-blinks the lid covers the eye, and the camera cannot detect the pupil. Since eye tracking systems deal with this problem (loss of information) in a variety of ways (Gitelman, 2002), it is impossible to create a universal procedure to recover missing information. Several algorithms for eye-blink detection have been proposed, by both researchers directly interested in the eye-blink phenomenon and researchers faced with eye-blink artifacts (Pedrotti et al., 2011). Once blink onset and offset have been identified, missing/corrupted pupil data are usually estimated using linear (or cubic) interpolation, or even more sophisticated techniques such as moving average or support vector regression (see Nakayama et al., 2012). Overall, pupil data preprocessing is necessary since it is known that eye-blink artifacts have an impact on analysis results, both in the time and frequency domains (Nakayama, 2006; Nakayama & Shimizu, 2002, 2004).

The Task-Evoked Pupillary Response (TEPR) is a useful tool for PD signal analysis in the time domain. The main contributions on TEPR come from cognitive psychology. The rationale underlying this method is the same as the Event-Related Potential (ERP) in electroencephalography (EEG) measurements. Since the magnitude of psychologically-induced pupillary responses can be in the order of tenths - even hundredths - of mm, we can be more confident in associating such responses to a given stimulus if we know the exact time point of stimulus appearance. Like for ERPs, a time window (e.g. 500ms) before stimulus onset is used as baseline, i.e. the average PD $\bar{x}$ for the pre-stimulus time window is calculated. Subsequently, $\bar{x}$ is subtracted from each data point in the post-stimulus time window (e.g. 5s after stimulus onset). The resulting waveform - usually an average of several measurements (e.g. Fig 3-A) - indicates the pupillary reaction to the stimulus, and parameters such as peak dilation and latency to the peak can be calculated. Beatty (1982) demonstrated that such parameters are sensitive indicators of processing load for different cognitive tasks. Although this technique is appropriate for laboratory studies implying short and simple tasks, several constraints might undermine its application in more complex and dynamic situations. Klingner (2010) introduced the fixation-aligned pupillary response averaging, in which eye fixations - instead of experimental stimuli - are used to temporally align PD time windows. Additionally, TEPRs have been analyzed using Principal Component Analysis (PCA) and Independent Component Analysis (ICA) in an attempt to reduce the large number of time points to a smaller set - usually two or three factors (see Granholm & Verney, 2004; Jainta & Baccino, 2010; Kuchinke et al., 2007; Verney et al., 2004).

To the authors' knowledge, Lüdtke et al. (1998) first introduced the use of Fast Fourier Transform (FFT) for pupil signal analysis in the frequency domain. Analyzing a signal in such domain allows to know - e.g. - whether significant changes happen within specific frequency bands. With the aim of detecting low-frequency fatigue-related pupillary oscillations, Lüdtke and colleagues evaluated the mean value of the amplitude spectrum for all the frequencies ≤0.8Hz. Nakayama & Shimizu (2004) found that the power spectrum density (PSD) of pupil signals increases within certain band intervals (0.1-0.5 Hz and 1.6-3.5Hz) as a function of cognitive task difficulty. Lew et al. (2008) analyzed PD signals using the Short-Time Fourier Transform (STFT), which allows to extract the frequency information yet preserving the time domain.

One promising technique for reducing data complexity in recorded PD signals is wavelet analysis. Marshall (2000, 2002) first proposed the use of wavelet analysis for analyzing PD time-series, and associated the occurrence of high-frequency variations (faster than 20ms, i.e. >50Hz) to instances of cognitive load. Since then - to our knowledge - few studies have applied wavelet transforms to PD signals: Shi et al. (2006) analyzed pupillary behavior in relation to user' s visual ability; Pinzon-Morales & Hirata (2012) evaluated PD oscillations to estimate sleepiness levels. In the present study, we used the Discrete Wavelet Transform (DWT) as a tool to extract relevant signal features (i.e. low-frequency approximation), discarding the noise that appears in the high-frequency part of the signal (see section 3.4.4). In this respect, our approach is opposite to the one proposed by Marshall.

2. METHOD

2.1. Stimuli

Our rationale for stress manipulation implies the repeated performance of a simple driving task, to which we added different external stressful stimuli (see section 2.3). The protocol was approved by the French National Board of Informatics and Freedom (declaration n. 0727429; www.cnil.fr). Participants performed a simulated Lane Change Test (LCT), which consists of driving on a traffic-free straight three-lane road (see Fig. 1-A), changing lanes according to the information appearing on two identical road signs displayed concurrently every 150m, on both sides of the road (ISO, 2010; Minin et al., 2011; Mitsopoulos-Rubens et al., 2011). The driving simulator consisted of seat (Playseat Evolution), steering wheel and pedals (Logitech G25, no gear-shift was used), and a 32" LCD monitor (Thomson 32LB220B4, 70x39cm, 1366x768px). The LCT software (www.corys.com) limited the maximum speed at 60km/h, so that participants could maintain this speed by simple flat-out. Each trial consisted of 18 lane changes, accomplished over 180s (ISO, 2010). The average distance between participant and screen was 130cm.

2.2. Participants

Thirty-three healthy people (all with valid driving license) participated in the experiment. Seventeen people were allocated to the *experimental* group - i.e. the group that underwent stress manipulation - and the remaining sixteen were assigned to the *control* group (see section 2.3). The *experimental* group contained 9 women (mean age = 38 years, *SD* = 15) and 8 men (mean age = 43 years, *SD* = 9). Eight women (mean age = 42 years, *SD* = 8) and 8 men (mean age = 41 years, *SD* = 13) were assigned to the *control* group. All participants read and signed an informed consent, and received a reward for every hour spent inside the laboratory. Participants were informed that they could leave the experiment at any time and for any reason. One participant from the *control* group quit the experiment during t_2 because of simulator sickness.

Participants' stress trait was measured with the State-Trait Anxiety Inventory (STAY-B, Spielberger et al., 1983, translated in French by Schweitzer et al., 1990). In order to disclose possible social desirable responding, we asked participants to fill in the Social Desirability Scale (DS36, Tournois et al., 2000). The *experimental* and *control* groups did not differ in terms of STAY-B scores (t(30) = 1.03, n.s.), nor on DS36 scores (t(30) = 0.1, n.s. for *self-deception*, and t(30) = 0.23, n.s. for *other-deception*).

2.3. Experimental design and procedure

Upon arrival at the lab, participants sat in a quiet room wherein they installed the electrodes for electrodermal measurement on their forehead (see section 3.5), read and signed an informed consent, and filled in the STAY-B and DS36 questionnaires (see section 2.2).

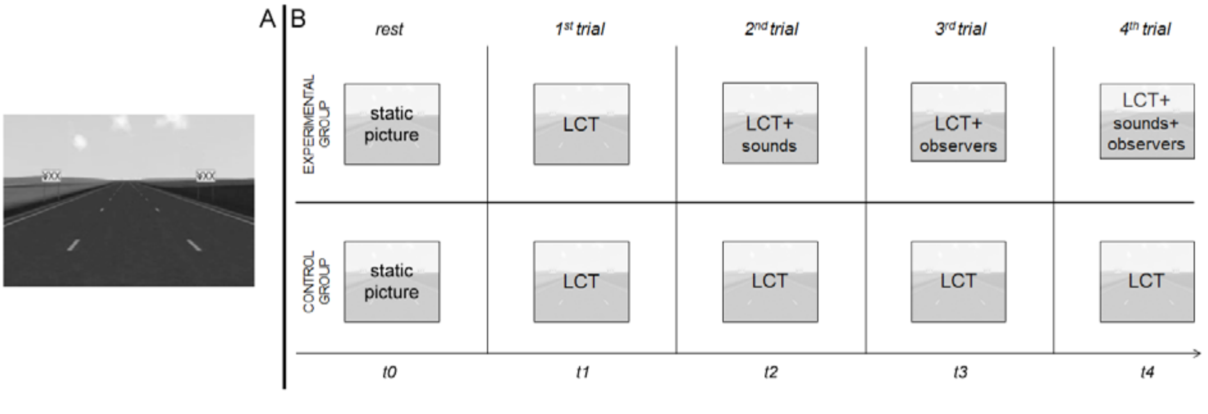

FIG. 1. (A) Lane Change Test scenario. (B) Experimental design.

Thereafter, participants moved to the simulation room, where they received on-screen instructions for the LCT (see ISO, 2010, Annex A) and performed a one-minute LCT training to disclose any potential issue (misunderstandings, simulator sickness, technical equipment, etc.).

Before beginning the LCT sessions, we recorded baseline physiological signals during a 90s rest period (t_0) in which participants looked at a static picture of the LCT scenario. Subsequently, participants carried out four LCT trials (Fig. 1-B). The *control* group performed the four driving trials without any disturbing factor. The *experimental* group underwent two types of stressors, announced by means of screen instructions displayed before the LCT trial started.

Before the 2nd trial (t_2), participants were informed that they would hear a sound if their driving behavior was not appropriate (i.e. beginning a lane change as soon as the symbols appear on a sign, but not before - see ISO, 2010, Annex A). Indeed, a sound was presented every 20s, regardless of driving behavior. In order to avoid excessive habituation, three different sounds were used: a 8.6Hz white noise, a police siren, and a 4KHz tone (.*wav* files are available from the corresponding author). All sounds lasted 1s, and they were presented in a pseudo-random order: each sound was presented 3 times, totaling 9 sound presentations in the 2nd trial.

Before the 3rd trial (t_3), participants were informed that their driving performance would have been evaluated by two experts. After this announcement, two experimenters entered the test room. These “fake” experimenters (1 man and 1 woman) wore white lab coats, they held a copybook which they used for taking notes, and they stood on the participant’ s right (90° of visual angle) so that their presence could be perceived without disturbing the execution of the driving task. Moreover, a previously floor-pointed camera was turned towards the participant, so that she/he would believe she/he was being filmed. No stressful sounds were presented during the 3rd trial.

Before the 4th trial (t_4), participants were informed that they would have been evaluated by the experimenters and alerted with sounds in case of incorrect driving behavior.

After each LCT trial, participants reported their perceived stress level (see section 3.3). Each trial (t_1, t_2, t_3, t_4) lasted three minutes.

2.4. Hypotheses

For the *experimental* group, stress level should be lowest at t_1 - where simple LCT performance is demanded - and highest at t_4, where two concurrent stressors are associated with the LCT task. We cannot predict whether stress level will significantly differ between t_2 and t_3, i.e. we do not know *a priori* whether alert sounds are more stressful than observers (or vice versa). What we know, is that they are two different types of stressors, and they could be then classified.

For the *control* group, we expect stress level to significantly decrease from t_1 to t_4 as an effect of habituation.

The two groups should exhibit the same stress level at t_1, since they bear exactly the same conditions until that point. We expect to find significant between-groups differences at t_2, t_3, t_4.

We expect PD changes to reflect increased stress levels for the experimental group, and decreased stress levels for the control group. Electrodermal and self-report measures are expected to correlate with pupillary behavior measures.

3. DATA ACQUISITION AND ANALYSIS

3.1. Apparatus synchronization

The equipment used in the present study includes an eye tracker (RED 4, www.smivision.com) for PD measurement, an A/D converter (MP36R, www.biopac.com) for measuring illumination and EDA, a PC for stimulus (LCT) presentation, and a PC running Matlab (www.mathworks.com) for synchronization and auditory stimulus delivery (psychtoolbox.org; Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). All the systems were connected on a local area network (LAN), and synchronization was achieved using a combination of custom software written in Matlab and free software (synergy-foss.org). Each single experimental event (e.g. beginning of a LCT trial, presentation of a LCT sign, presentation of a sound, etc.) was marked - via TCP/IP messages - in the eye tracker’s and A/D converter’s log files for conjoint offline analysis. A detailed technical description goes beyond the scope of the present article, and interested readers may refer to the corresponding author for further information.

3.2. Illumination

Illumination was measured with an Extech 403125 luxmeter (www.extech.com). With the aim of recording the amount of light that impacted on the participant’ s eyes, the light sensor was attached to a ceiling-mounted holder, placed 5cm above participant’ s head, laterally centered with respect to her/his nose. The luxmeter was connected to a Biopac MP36R A/D converter which stored illumination values in *lux* at 50Hz sampling rate.

3.3. Visual Analog Scale (VAS)

After each LCT trial (t_1, t_2, t_3, t_4) participants rated their perceived stress level using 3 on-screen Visual Analog Scales. The VAS ranged from 0 (not at all) to 100 (maximum). The three dimensions were *stress* (“how much stressed were you feeling during task performance?”), *anxiety* (“how much anxious were you feeling during task performance?”), and *avoidance* (“during task performance, to what extent were you willing to leave the situation?”). Correlation analysis revealed strong correlations between the three dimensions, except between *stress* t_1 and *avoidance* t_1, and between *anxiety* t_1 and *avoidance* t_1 (see Table 1).

TABLE 1
Correlations Between VAS Dimensions

	anxiety t_1	anxiety t_2	anxiety t_3	anxiety t_4	avoidance t_1	avoidance t_2	avoidance t_3	avoidance t_4
stress t_1	.93**				.13			
stress t_2		.87**				.76**		
stress t_3			.95**				.51**	
stress t_4				.61**				.74**
anxiety t_1					.18			
anxiety t_2						.66**		
anxiety t_3							.52**	
anxiety t_4								.37*

Note. N = 32 (one participant quit the experiment after t_2 because of simulator sickness). *$p < .05$. **$p < .005$.

3.4. Pupil diameter

Pupil diameter was recorded at 50Hz sampling rate using a SMI RED 4 remote video eye tracker. This system measures pupil size and eye movements by means of pupil and corneal reflection tracking (see Holmqvist et al., 2010), and has a precision of 0.01mm for pupil diameter. PD signal was treated according to the following procedure:

Step 1) Preprocessing: eye-blink artifacts were identified and replaced by linear interpolation (see section 3.4.1).

Step 2) TEPRs extraction: pupillary responses following sound presentations (in t_2 and t_4) were evaluated to confirm the existence of pupillary reactions to stressful stimuli (see section 3.4.2).

Step 3) Normalization: PD values in t_1, t_2, t_3, t_4 were normalized for each participant, according to her/his average PD at rest (see section 3.4.3).

Step 4) Analysis of Variance (ANOVA): we tested the hypothesis that our stress manipulation had an effect on average PD (see section 4.2).

Step 5) Signal approximation extraction: the normalized PD signals from t_1, t_2, t_3, t_4 were transformed by means of Discrete Wavelet Transform (DWT). The Haar wavelet was used to decompose and transfer the signal into multi-resolution representation (see section 3.4.4).

Step 6) Classification with neural networks: PD signal approximation was used as an input vector (feature) during the training and test stages (see section 5).

The first step (Preprocessing) is generally necessary regardless of the aim of any study. Concerning our study, we consider steps 2 and 4 (TEPRs extraction, ANOVA) mandatory for theoretically justifying the use of PD as a stress index, since they demonstrate the sensitivity of PD to stress manipulations. Normalization (step 3) is necessary since each person has her/his own PD at rest. Steps 5 and 6 (Signal approximation extraction and Classification) are an attempt to use PD for automatic stress measurement in applied contexts. With the aim of being able to measure stress levels over relatively short time periods, we extracted and analyzed (in steps 4, 5 and 6) only the first 80s of PD data from each trial.

3.4.1. Pupil diameter preprocessing

The RED 4 provides two types of pupil measurement, (a) in pixel and (b) in millimeters (Fig. 2). In the present study, mm values were preferred for PD signal analysis since they are calculated taking into account the distance between participant and camera, providing more reliable data than raw pixel measurement. Nonetheless, the RED 4 outputs useful information for blink detection in the pixel data. When a blink occurs, zeros - along with other physiologically impossible values - are recorded in the pixel output (see Fig. 2, bottom line, right-hand scale). In a statistical perspective, such values can be considered as outliers of the PD distribution of a given data set. Blinks were detected as contiguous sets of outliers. Moreover, other blink markers in the eye tracking protocol - such as the momentary loss of gaze position during blink - were combined in order to foster correct blink detection percentage (for a detailed description of the algorithm, see Pedrotti et al., 2011). Blink onset was defined as the 3rd sample (60ms) preceding the first zero observation: at this point, the lid starts its descent until the pupil is covered (in 79% of blinks, see Pedrotti et al., 2011). Blink offset was defined as the first valid sample after a blink: at this point, the pupil is visible to the eye tracker camera.

After identifying blink onsets and offsets in the pixel data, we replaced blink data in the mm output by means of linear interpolation, using blink onset as starting point, and two samples after blink offset as ending point. An example of the result of this preprocessing procedure is shown in Fig. 2 (top line, left-hand scale).

3.4.2. Task-Evoked Pupillary Responses (TEPRs)

TEPRs were extracted - from the *experimental* group - for each sound presentation in t_2 and t_4 (see Fig. 1-B and section 2.3). Baseline pupil diameter was computed as the mean PD in the 200ms pre-stimulus time window. Fig. 3-A shows the average pupillary response (solid line, left-hand scale) from 288 waveforms (16 participants[1] x 9 sounds x 2 trials). The dashed line (right-hand scale) shows the average

[1] One participant' s data were excluded because of poor recording quality.

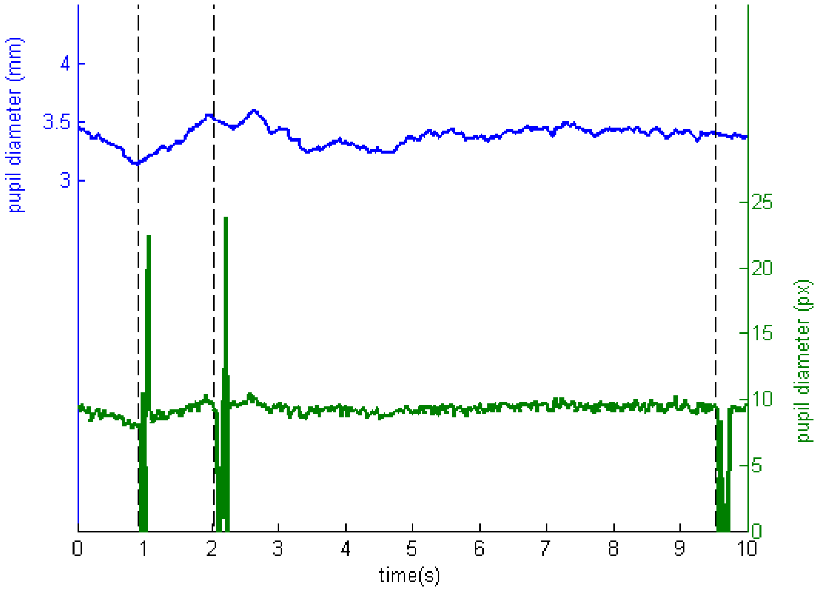

FIG. 2. Pupil output in mm (top line, left-hand scale) and px (bottom line, right-hand scale). *Note.* Dashed vertical lines indicate blink onsets.

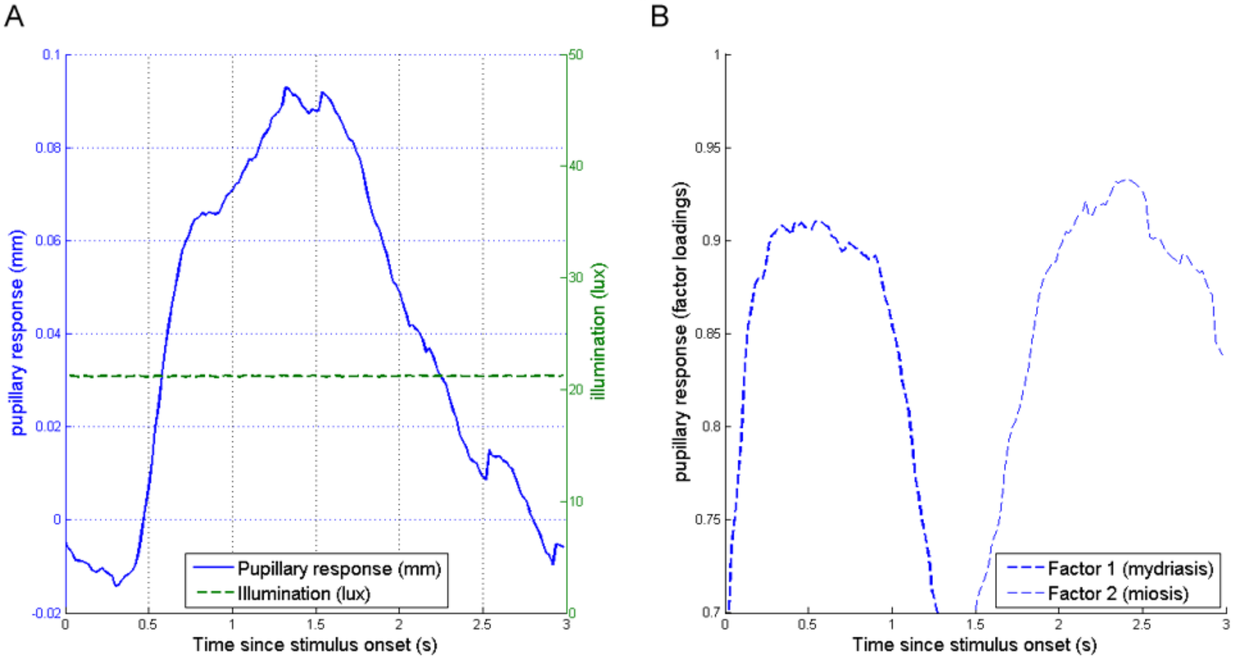

FIG. 3. (A) Pupillary response to stressful sounds (solid line, left-hand scale) under constant illumination (dashed line, right-hand scale). (B) Pupillary response factor loadings (only loadings >0.7 are considered for interpretation).

illumination, measured in synchrony with PD (see section 3.1). A typical phasic pupillary reaction - dilation (mydriasis) followed by constriction (myosis) - occurred. With the aim of reducing the 150 data points (50 points x 3s, 50Hz sampling rate) to a smaller set of factors, the data from the 288 TEPRs extracted were analyzed with a Factor Analysis using Statistica (www.statsoft.com). The 150 data points were treated as dependent variables serving as input for Factor Analysis. Factor loadings were extracted following a varimax rotation. Two factors could explain 82.34% of variance in the data, i.e. pupillary response shapes are consistent across individuals and trials. Factor loadings plotted against time (Fig. 3-B) clearly show the separation between the rising (factor 1, mydriasis, 62.91% of explained variance) and falling (factor 2, myosis, 19.43% of explained variance) part of the pupillary response depicted in Fig. 3-A. Moreover, the absence of measured relevant illumination changes allows us to associate the recorded waveform to the stress elicited by experimental manipulation (sound delivery associated with poor driving performance).

3.4.3. Pupil diameter normalization

The TEPR does not require previous data normalization, since the baseline PD is re-calculated for each event, based on a short pre-stimulus time window (200ms in our case): this procedure can be viewed as a sort of normalization, in that a participant- and moment-specific PD value is subtracted from absolute PD values. However, before performing any inter-individual comparison (such as ANOVA), PD values should be normalized, since it is known that PD at rest differs between people. For each participant, we calculated mean PD (μPD_{t0}) at rest (t_0 in Fig. 1-B). Subsequently, μPD_{t0} was subtracted from each PD data point in t_1, t_2, t_3, t_4. This procedure allows for later comparisons of PD between participants.

3.4.4. Signal Approximation Extraction

Combination of Wavelet and Neural Networks has been accepted as an accurate method for feature extraction and classification (Minu et al., 2010). Any noisy signal imposes some uncertainty to the calculation and - consequently - to the results. Therefore, before using PD signal as input for the neural network classifier, we need to remove noise from the signal. De-noising could improve classification performance, since it would increase the signal-to-noise ratio. Moreover, it would reduce computational costs, i.e. shorter time to obtain results: this latter aspect is important in a real-time application perspective, although we focus on offline analysis for the present study.

Several mathematical approaches could be used for this purpose: we chose the Discrete Wavelet Transform (DWT) because *a)* it allows removing noise yet preserving the original shape of the signal and *b)* it encompasses a down-sampling procedure which reduces computation time.

Mathematically, a signal (time series) $x(t) \in L(R)$ can be decomposed into linear combination of a set of *n* base functions $\{\phi_0, \phi_1, \cdots, \phi_n\}$ if the signal is in the space spanned by the basis. Then, $x(t)$ can be decomposed into a linear combination of the base functions (Mallat, 1989):

$$x(t) = \sum_k a_k \phi_k(t) \ \ k \prec n \ (k \prec \infty), k \in \mathbb{Z} \qquad (1)$$

Where *k* is an integer index of the finite or infinite sum and $a_k, \phi_k(t)$ are expansion coefficients and functions respectively. This representation is the most common form of multi-spectrum decomposition. Consider two sets of base functions:

1. $\phi_{\ j,k}(t) = \ 2^{-\frac{j}{2}}\varphi(2^{-j}t - kx), \mathrm{j} \succ 0, \mathrm{k} \in \mathbb{Z}$
2. $\phi_k(t) = e^{\frac{2fktj}{T}}$

If item 1 or 2 are substituted in Equation 1, the wavelet and Fourier decompositions will be achieved respectively. These are two well-known examples of decomposition of a signal into primitive or fundamental constituents of their spaces. In fact, the Fourier series decomposes a signal into a set of sine and cosine functions. By DWT in multi-resolution analysis a signal is represented by a sum of a set of more flexible functions called *mother wavelet* which are localized both in time and frequency.

Any wavelet decomposition consists of two parts: approximation and detail. Approximation refers to the overall, general form of the signal (e.g. low frequency component), while detail better explains the high frequency information such as edges, discontinuities, sharp points, etc.

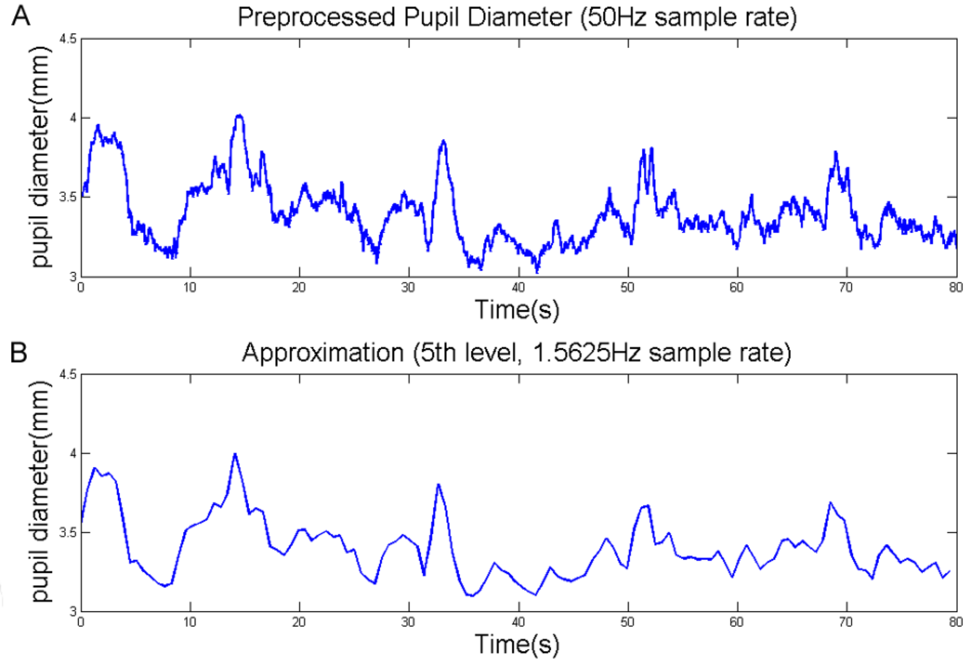

FIG. 4. (A) Preprocessed pupil diameter from trial t_2 of participant #7. (B) Signal approximation after 5 wavelet (Haar) decomposition levels.

Approximation and detail coefficients of a given discrete signal *x[n]* can be extracted by low-pass and high-pass filtering respectively. Figure 4-B shows an example of signal approximation extraction using the Haar wavelet as mother wavelet, and 5 decomposition levels (i.e. the output of level$_n$ is used as input for level$_{n+1}$). It seems evident that approximation preserves the original shape of the signal, while noise is discarded: in most cases, noise resides in the high-frequency part (Hamid et al., 2011). In the original signal (Fig. 4-A) 4000 data points (coefficients) are needed to describe 80s of pupil diameter (sample rate is 50Hz). Each wavelet decomposition level encompasses a down-sampling by a 2 factor. After five decompositions, sampling rate is reduced from 50Hz to 1.5625Hz, and 125 coefficients are sufficient to describe 80s of pupil diameter. The vectors containing the 125 coefficients will be used as input for a neural network classifier (see section 5).

3.5. Electrodermal Activity (EDA)

Skin conductance (SC) was recorded using the exosomatic method with Direct Current (Boucsein, 2012). Two Biopac EL507 disposable circular electrodes (Ag/AgCl, 1cm diameter circular contact area, 0.5% Chloride) were attached to the participant' s forehead. Although palmar and plantar zones would be preferable for EDA recording - since they have higher sweat glands density (Dawson et al., 2000; Sato et al., 1989) - the driving task employed in the present study required both hands and feet to be completely free. The electrodes were fixed to the skin upon participant' s arrival at the lab, assuring a minimum delay of 15min before the recordings started. This time is sufficient to allow good electrical contact between the skin and the electrode surface. An elastic headband was used to prevent artifacts due to wire movements (Boucsein et al., 2012). The electrodes were connected to the Biopac MP36R A/D converter (50Hz sampling rate). For the EDA recording channel, a low-pass filter was applied (35Hz cutoff frequency). Skin Conductance signals were treated with the following procedure:

Step 1) Filtering: a low-pass filter was applied, with 2Hz cutoff frequency.

Step 2) Down-sampling: sampling rate was reduced from 50Hz to 10Hz.

Step 3) Transformation: data were transformed with the formula $y = log(1 + x)$ (Boucsein, 2012). The respective units are labeled as *log µS*.

Step 4) Skin Conductance Response (SCR) extraction: SCRs were extracted following sound presentations in t_2 and t_4 (*experimental* group). The average SC in the 200ms pre-stimulus time window was subtracted from each SC data point in the 10s post-stimulus time window. Fig. 5-A shows the average SCR from 288 sound presentations (16 participants[2] x 9 sounds x 2 trials). Like for PD data (see section 3.4.2), we used SCR time-series as input for Factor Analysis: two factors could explain 75.52% of variance (54.24% Factor 1, 21.28% Factor 2). Fig. 5-B shows factor loadings plotted against time. Unlike for TEPRs, the temporal separation between the two factors (roughly at 4s) does not match the peak of the average SCR (Fig 5-A): factor interpretation is harder in this case, however the overall proportion of explained variance tells that SCRs - like TEPRs - have a uniform structure across participants and trials.

Step 5) Non-Specific Electrodermal Response Frequency (NS.EDR freq.) extraction (see section 4.3.1)

Step 6) EDA Area extraction (see section 4.3.2)

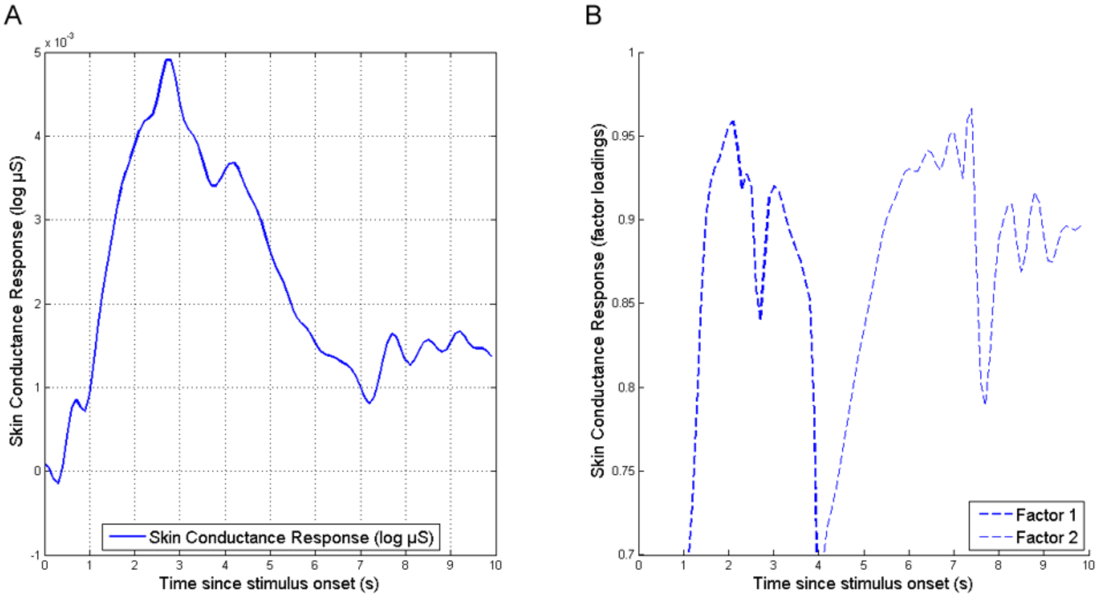

FIG. 5. (A) Skin Conductance Response to stressful sounds. (B) Skin Conductance Response factor loadings (only loadings >0.7 are considered for interpretation).

[2] One participant was excluded because of too many artifacts in the SC record.

4. RESULTS

4.1. Visual Analog Scale (VAS)

VAS *stress* scores from 32 participants[3] were analyzed with a repeated-measures ANOVA (rmANOVA) using *trial* (t_1, t_2, t_3, t_4) as within-factor and *group* (*experimental*, *control*) as between-factor (see Fig. 6). A *trial**group* interaction effect was found (F(3,90) = 5.41, p<.005, η^2_p = .15). The effect of *group* is also significant (F(1,30) = 7.25, p<.05, η^2_p = .19), with higher scores for the *experimental* group.

The difference on VAS *stress* scores between the *experimental* and *control* group is not significant at t_1 (F(1,30) = 0.03, n.s.). VAS *stress* scores are significantly higher for the *experimental* group at t_2 (F(1,30) = 10.06, p<.005, η^2_p = .25), t_3 (F(1,30) = 7.79, p<.01, η^2_p = .21), t_4 (F(1,30) = 5.64, p<.05, η^2_p = .16).

A *trial* effect (F(3,48) = 4.6, p<.01, η^2_p = .22) was found within the *experimental* group: VAS *stress* scores are significantly higher at t_2 with respect to t_1 (F(1,16) = 8.72, p<.01, η^2_p = .35), at t_3 with respect to t_1 (F(1,16) = 9.78, p<.01, η^2_p = .38), at t_4 with respect to t_1 (F(1,16) = 4.46, p=.05, η^2_p = .22). The differences between t_2 and t_3, t_2 and t_4, t_3 and t_4 are not significant. Within the *control* group, there is no *trial* effect.

4.2. Pupil Diameter (PD)

Before carrying out any between-groups comparison at t_1, t_2, t_3, t_4, we verified that average PD at rest (t_0) did not differ between the *experimental* and *control* groups. Preprocessed average PD was calculated for sixteen participants of the *experimental* group (one participant was excluded because of poor recording quality) and for thirteen participants of the *control* group (two participants were excluded because of poor recording quality, one participant was excluded because she quit the experiment after t_2). The data were analyzed in an independent samples t-test, which confirmed no difference on average PD at t_0 between the two groups (*t*(27) = 1.52, *n.s.*).

We then tested the hypothesis that stress manipulation had an effect on mean PD across the experimental trials t_1, t_2, t_3, t_4. Normalized average pupil diameters were calculated for the *experimental* group (16 participants, one was excluded because of poor recording quality) and the *control* group (13 participants, two were excluded because of poor recording quality, one was excluded because she quit the experiment after t_2) at t_1, t_2, t_3, t_4.

Data were analyzed with a rmANOVA, using *trial* (t_1, t_2, t_3, t_4) as within-factor and *group* (*experimental*, *control*) as between-factor. A significant *trial**group* interaction effect was found (F(3,81) = 9.14, p<.001, η^2_p = .25, see Fig. 7).

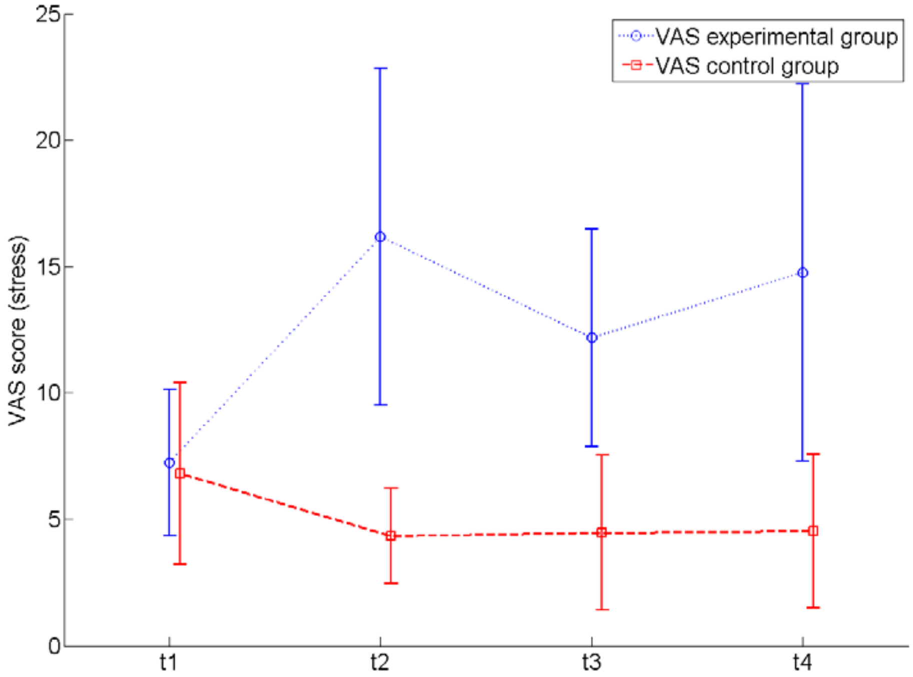

FIG. 6. Average VAS scores (*stress* dimension) for each group and experimental trial. *Note.* Vertical bars denote 95% confidence intervals (mean ±2SE). *N* = 32 (17 experimental + 15 controls).

[3] One participant quit the experiment after t_2 because of simulator sickness.

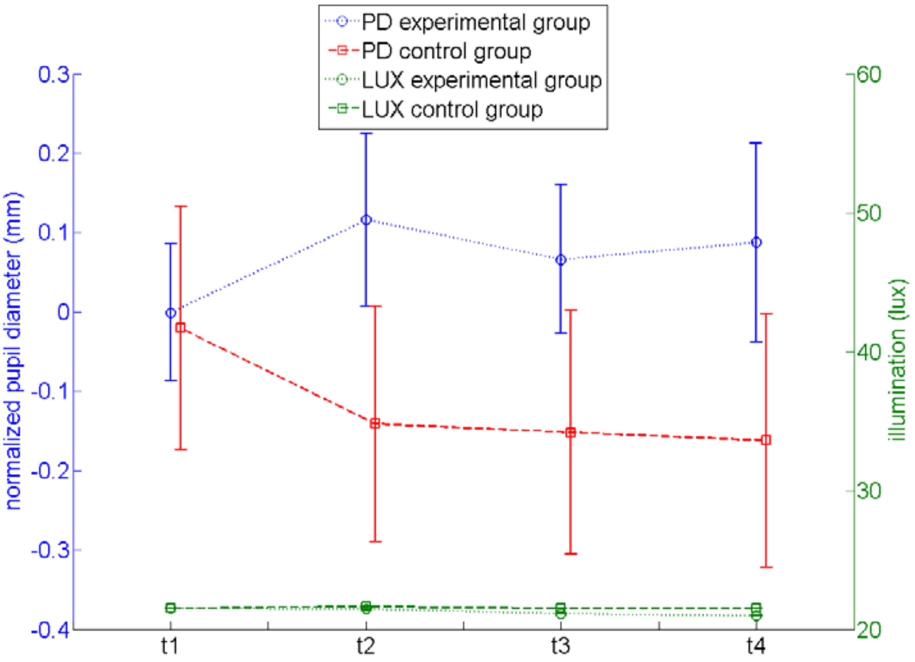

FIG. 7. Normalized average pupil diameter (left-hand scale) and average illumination (right-hand scale) for each group and experimental trial. *Note.* Vertical bars denote 95% confidence intervals (mean ±2SE). *N* = 29 (16 experimental + 13 controls).

The effect of *group* was also present (F(1,27) = 4.01, p=.05, η^2_p = .13), in that normalized average pupil diameter was higher for the *experimental* with respect to the *control* group. When t_1 was removed from the within-factors (since stress manipulation effectively started at t_2), the effect of *group* emerged more clearly (F(1,27) = 6.31, p<.05, η^2_p = .19).

There was no difference between the *experimental* and *control* group at t_1 (F(1,27) = 0.05, n.s.), while PD was significantly larger for the *experimental* group at t_2 (F(1,27) = 6.93, p<.05, η^2_p = .2), t_3 (F(1,27) = 5.35, p<.05, η^2_p = .16), t_4 (F(1,27) = 5.31, p<.05, η^2_p = .16).

The effect of *trial* was significant (F(3,45) = 3.45, p<.05, η^2_p = .19) within the *experimental* group: PD was significantly larger at t_2 with respect to t_1 (F(1,15) = 12.63, p<.005, η^2_p = .46), t_3 with respect to t_1 (F(1,15) = 7.42, p<.05, η^2_p = .33). The differences between t_1 and t_4, t_2 and t_3, t_2 and t_4, t_3 and t_4 were not significant.

Within the control group, the effect of *trial* was significant (F(3,36) = 7.22, p<.001, η^2_p = .37): PD significantly decreased at t_2 with respect to t_1 (F(1,12) = 7.7, p<.05, η^2_p = .39), at t_3 with respect to t_1 (F(1,12) = 13.74, p<.005, η^2_p = .53), at t_4 with respect to t_1 (F(1,12) = 12.83, p<.005, η^2_p = .52). The differences between t_2 and t_3, t_2 and t_4, t_3 and t_4 were not significant.

4.3. Electrodermal Activity (EDA)

4.3.1. Non-Specific Electrodermal Response Frequency (NS.EDR freq.)

NS.EDR freq. - i.e. the number of SCRs in absence of apparent stimulation - is thought to be an indicator of negatively tuned emotional states such as stress (Boucsein, 2012), in that NS.EDR freq. should increase under stressful conditions. NS.EDR freq. scores were analyzed with a rmANOVA using *trial* (t_1, t_2, t_3, t_4) as within-factor and *group* (*experimental*, *control*) as between-factor. No significant effects were found.

4.3.2. EDA Area

The area (computed as time-integral) below a SC waveform can be used as a measure of emotional arousal (see Bach et al., 2010; Boucsein, 2012). Since Skin Conductance Level (SCL) has great inter-individual variability, we subtracted the estimated tonic level from each SC time-series before computing area measures. Tonic level was estimated by means of deconvolution using the Ledalab package (Benedek & Kaernbach, 2010; www.ledalab.de). Fig. 8-A shows an example of the tonic level estimation used in this study. The SC waveform - after subtraction of the estimated tonic level - shows a zero baseline (Fig. 8-B), making it possible to compare SC between individuals. These SC vectors were used as input for calculating area scores on a trial-by-trial basis. Areas were calculated by trapezoidal numerical integration using the Matlab *trapz* function. Area scores were then analyzed with a rmANOVA using *trial* (t_1, t_2, t_3, t_4) as within-factor and *group* (*experimental*, *control*) as between-factor. No significant effects were found.

4.4. Bivariate analysis

4.4.1. Correlation between subjective and physiological measures

It appears that normalized average VAS and PD scores have a similar pattern across the experimental trials, for both the *experimental* and *control* groups (see Fig. 6 and Fig. 7). For the *experimental* group, both stress measures show a steep increase at t_2 with respect to t_1, followed by a decrease at t_3.

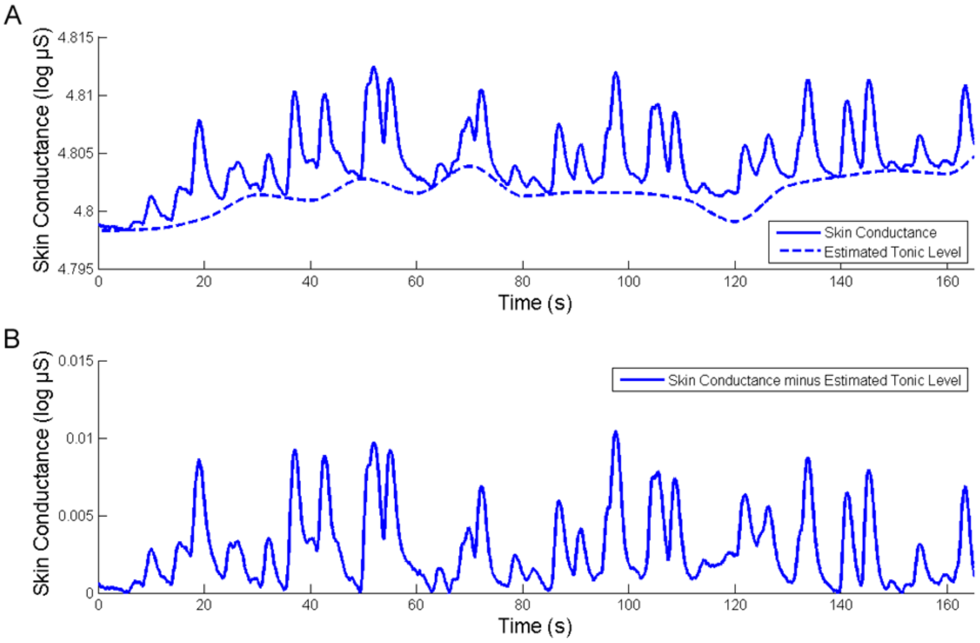

FIG. 8. (A) Skin conductance (solid line) and Tonic Level (dashed line) estimated by deconvolution (Benedek & Kaernbach, 2010). (B) Skin conductance minus Tonic Level, showing a zero baseline. The time-integral of this time series is used as EDA area measure.

Finally, another increase occurs at t_4. For the *control* group, both measures decrease at t_2, and maintain a relatively stable level until t_4. Significant correlations were found between normalized average PD and VAS *stress* scores at t_1 (*r* = .4, p<.05), t_2 (*r* = .41, p<.05), t_3 (*r* = .44, p<.05), t_4 (*r* = .4, p<.05).

4.4.2. Correlation between PD and illumination

Illumination is an important factor influencing PD (see section 1). In this experiment, both stimulus and ambient illumination were kept constant during the whole experiment. Despite, illumination data analysis revealed a slight illumination decrease for the *experimental* group at t_3 and t_4 (see Fig. 7). Post-experiment investigations revealed that the cause could be attributed to the presence of the "fake" experimenters in the test room. Any displacement in the test room - even apparently insignificant ones like removing a chair - could cause an illumination change, indexed by the high sensitivity of the luxmeter. Although such subtle changes are below visual threshold and could not influence pupil size (Loewenfeld, 1993), we tested whether normalized average PD was correlated with average illumination at t_1, t_2, t_3, t_4. No significant correlations were found.

5. CLASSIFICATION OF PD WITH NEURAL NETWORKS

After verifying the sensitivity of PD to stress manipulations (see section 4.2), we used PD signal approximation as input for a classifier. Statistical analyses support our choice of PD as a stress measure, in that average PD is significantly larger for the *experimental* group - with respect to the *control* group - at t_2, t_3, t_4, while there is no difference at t_1. This is in line with our predictions, since there was no stress manipulation until the end of t_1 , i.e., the two groups were exactly in the same conditions before t_2. Further support for this consideration comes from the subjective stress ratings (VAS, see section 4.1).

The aim of this analysis stage is automatic stress classification using normalized pupil diameter as the only information source: for this purpose, we use only PD data from the *experimental* group, i.e. the group that underwent stress manipulation.

Following our experimental plan (Fig. 1-B), four classes should be used, i.e. one class for t_1, one class for t_2, one class for t_3, and one class for t_4. The hypothesis underlying the experimental plan was that participants in the *experimental* group would feel more stressed as the experiment went on, with t_4 being the most stressful trial because of the cumulative effect of stressful sounds and human observers. Indeed, statistics revealed that PD significantly increased only at t_2 and t_3 with respect to t_1, that is, differences between the different types of stressors (sound at t_2, observation at t_3, and their combination at t_4) could not be revealed using statistical linear models (see section 4.2). Such statistics rely on mean and variance as basic features for discrimination. In contrast, neural networks have a non-linear characteristic which is imposed by non-linear transfer functions such as *logsig*, *tansig*, etc. Such a more sophisticated classifier could improve discrimination performance using the whole signal (or its approximation) as input. Specifically, PD signal approximations (see section 3.4.4) were used as input features for classification.

The classification procedure involves two stages. In stage 1 (training), four binary neural network classifiers are trained. Each of these classifiers operates in *one-versus-all* mode, i.e. the aim of the training here is to maximize recognition precision of one class with respect to all the other classes (e.g. maximize recognition of t_1 with respect to t_2, t_3 and t_4).

In stage 2 (test), the four binary classifiers are put in parallel. An unknown, unlabeled PD signal approximation $\vec{x}'$ is given as input to each of the four binary classifiers. Each classifier returns a score y (between 0 and 1) which can be interpreted as the probability that $\vec{x}'$ belongs to a certain class (i.e. the degree to which an instance is a member of a class, see Fawcett, 2006). The final decision is made according to the highest score attributed to $\vec{x}'$ by each of the four binary classifier.

Data from 10 participants (randomly selected) were used in the training stage (10 participants x 4 classes, totaling 40 signals). Data from the remaining 6 participants were used in the test stage (6 participants x 4 classes, totaling 24 signals). Implementation details are outlined in the following sections.

5.1. Binary classifiers architectures

Figure 9 shows a schematic representation of the *one-versus-all* classification procedure: a 80s artifact-free PD signal $\vec{x}$ (preprocessed normalized PD) is decomposed by means of Discrete Wavelet Transform (DWT, see section 3.4.4) using the Haar mother wavelet. Signal approximation $\vec{x}'$ is extracted and given as input to a binary neural network classifier. The classifier returns a score y. In an ideal situation, the binary classifier "t_1 vs t_2,t_3,t_4" (i.e. the classifier specialized for recognizing PD signals coming from t_1 trials) assigns a score $y=1$ to an input signal $\vec{x}'$ recorded during a t_1 trial, and a score $y=0$ to an input signal $\vec{x}'$ recorded during either a t_2,t_3 or t_4 trial.

Table 2 summarizes architectural details of the four *one-versus-all* binary classifiers.

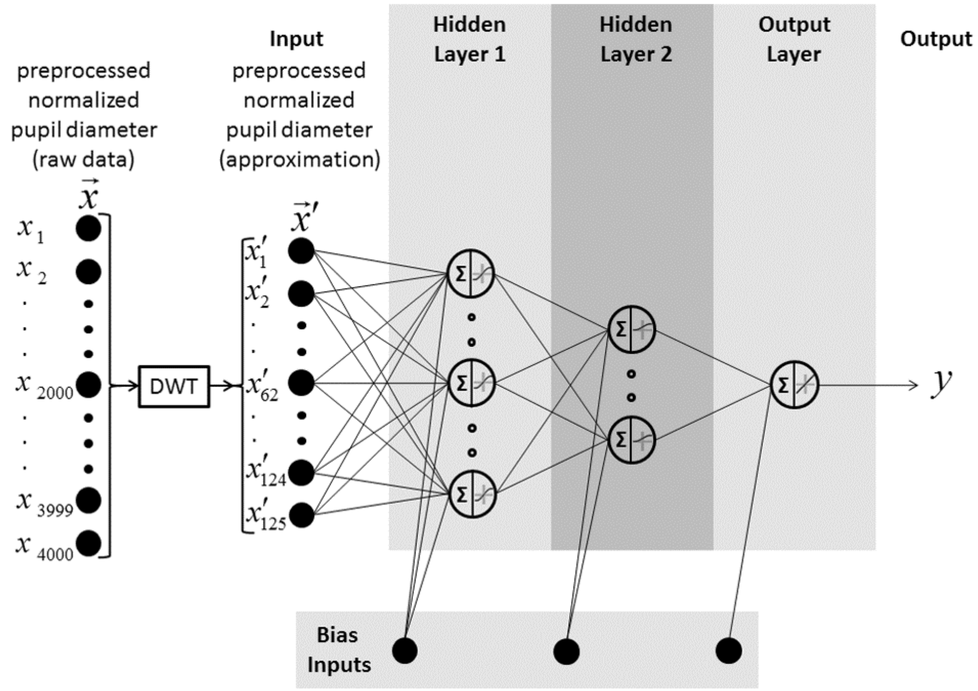

FIG. 9. Schematic representation of the *one-versus-all* binary neural network classifier.

TABLE 2
Neural Network Classifiers Architectures

Network	Hidden Layer I # of neurons	Hidden Layer I transfer function	Hidden Layer II # of neurons	Hidden Layer II transfer function	Output Layer # of neurons	Output Layer transfer function
t1 vs *t2,t3,t4*	15	tan-sigmoid	8	log-sigmoid	1	pureline
t2 vs *t1,t3,t4*	14	tan-sigmoid	7	log-sigmoid	1	pureline
t3 vs *t1,t2,t4*	11	tan-sigmoid	6	log-sigmoid	1	pureline
t4 vs *t1,t2,t3*	11	tan-sigmoid	6	log-sigmoid	1	pureline

Note: Architecture details of the four binary classifiers (see Fig. 9). Each classifier is designed to maximize recognition of one class with respect to all the other classes (e.g. maximize recognition of t_1 with respect to t_2, t_3, t_4) .

5.2. Binary classifiers training

Each of the four binary classifiers was trained separately using the Matlab Neural Network Training tool (*nntool*). The Levenberg-Marquardt algorithm - which updates weight and bias values according to gradient descent and other conjugate gradient methods (Moré, 1978) - was selected. Parameter values are reported in Table 3.

5.4. Four-way parallel classifier test

Data from 6 participants were used for test, totaling 24 (6 participants x 4 classes) signal approximations $\vec{x}'$. The four-way parallel classifier has a precision of 79.2%, i.e. 5 misclassifications out of 24 signals. Detailed confusion matrix is shown in Table 4.

TABLE 3
Parameter Values of the Neural Network Training Algorithm

Parameter	epochs	time	goal	grad $_{min}$	μ	μ $_{dec}$	μ $_{inc}$	μ $_{max}$
Value	1000	infinite	0	1e-08	0.001	0.1	10	1e10

Note: epochs = maximum number of iterations; *time* = time limit before algorithm stops; *goal* = target gradient value; *grad$_{min}$* = minimum gradient magnitude; *μ* = convergence factor (see Barman & Chowdhury, 2012).

5.3. Four-way parallel classifier architecture

Fig. 10 depicts the scheme of the four-way parallel classification procedure. An unknown, unlabeled PD signal approximation $\vec{x}'$ (i.e. $\vec{x}'$ has never been used in the training stage) is given as input to each of the four binary classifiers. Each classifier returns a score y. Scores are stored in the 4-D vector $\vec{y}$. In the example in Fig. 10, the binary classifier "t_1 vs t_2,t_3,t_4" assigned a score of 0.9 to $\vec{x}'$. All the scores assigned to $\vec{x}'$ from the other binary classifiers are lower than 0.9, thus we conclude that $\vec{x}'$ comes from a t_1 trial.

TABLE 4
Four-Way Classifier Confusion Matrix

		PREDICTED			
		t1	**t2**	**t3**	**t4**
ACTUAL	**t1**	66.7% (4)	0	33.3% (2)	0
	t2	0	83.3% (5)	0	16.7% (1)
	t3	0	0	83.3% (5)	16.7% (1)
	t4	0	16.7% (1)	0	83.3% (5)

Note. Number in parentheses indicate frequencies.

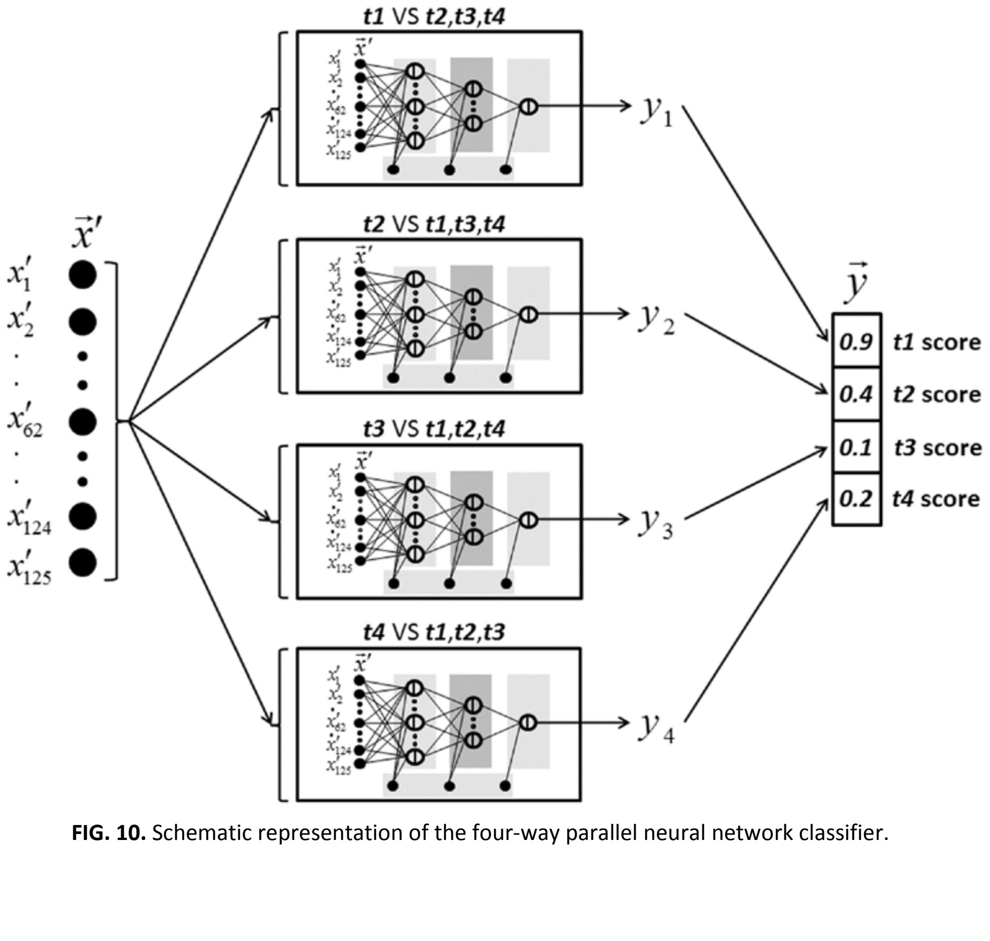

FIG. 10. Schematic representation of the four-way parallel neural network classifier.

6. DISCUSSION

Among several psychophysiological correlates of stress - such as cardiovascular activity, electrodermal activity, respiration - we focused on PD since it can be measured in a completely unobtrusive manner. This makes PD particularly attractive in a real-life implementation perspective, where stress level could be measured automatically, by using video cameras. We proposed a method for relating PD behavior to psychological stress and tested its validity in a simulated driving experiment. For ethical reasons, experimental stressors were conceived to elicit moderate stress levels, as confirmed by subjective ratings (VAS) results (see Fig. 6). The fact that PD could index such mild variations is encouraging for further developments in which higher stress levels could be detected.

We proved the sensitivity of PD as a stress concomitant, in both event-related and general state paradigms. For the event-related part, we showed how pupillary responses (TEPRs) follow the presentation of auditory stimuli associated with poor task performance: participants in the *experimental* group were told that they would hear a sound alert if their driving behavior was not appropriate. Because of the simplicity of the driving task (LCT), participants were expected to feel disoriented, irritated, frustrated at every sound presentation: during t_2 and t_4 they performed exactly the same driving task as t_1, with the exception that sounds alerts occurred every 20s regardless of driving performance. Thus, the TEPR reflects momentary stressful stimuli delivery, given that the exact time point of stimulus presentation is known. This technique is useful for basic research in controlled experiments, but it requires high control since every task-relevant event needs to be marked for offline analysis. In more naturalistic environments, tasks can have complex structures: stressful events might not be predictable and localized in time *a priori*. Moreover, the small magnitude of the TEPR (roughly 0.1mm as in Fig. 3-A) makes it particularly prone to confounding factors such as measurement noise and other sources of pupillary variation. For applied contexts, a valuable stress assessment method should be blind to both the temporal occurrence of stressors and the structure of the task.

The TEPR issues could be partially overcome using a more general-state measure, i.e. normalized average PD. This measure was influenced by the experimental manipulation as well. If we look at between-groups comparisons (i.e. *experimental* vs *control*), normalized average PD is definitely a powerful discriminator since it significantly increased at t_2, t_3, and t_4. Furthermore, it showed no between-groups difference at t_1, confirming its reliability. For within-groups analysis, however, discrimination becomes more challenging, as we are looking for subtle differences between stress induced by sounds and human observers (and combination of both), and we don' t know *a priori* which one is more stressful: within the *experimental* group, significant differences were found only between the non-stress- (t_1) and stress-trials (t_2, t_3), i.e. we could not discriminate between t_2 and t_3, t_2 and t_4, t_3 and t_4. Moreover, although average PD is higher at t_4 with respect to t_1 (see Fig. 7), the difference is not statistically significant: two factors could explain this. First, humans are likely influenced by habituation: at t_4, participants have already had some experience with both *sounds* (since t_2) and *observers* (since t_3), thus it is reasonable that they feel less stressed at t_4. Moreover, at t_4 they perform the LCT for the 4th time, which should also lower the stress induced by the LCT itself. We suggest that habituation (to both the LCT and the stressors) played a major role than the summation of two stressors (these two stressors are no more a novelty at t_4). Second, we are dealing with linear statistical models (ANOVA) and it should be clear that psychophysiological phenomena cannot be completely described in this domain.

With the aim of improving discrimination performance in a real-life oriented context, we devised an automated classifier. The results provided by the classifier are promising, yet we underline that they come from PD data recorded under controlled illumination. Designing such an autonomous stress measurement system, which relies solely on a short period of a time series (or a real-time signal), raises at least one question: what would happen if the level of noise increases, e.g. because of environmental effects? In the case of PD, we can regard environmental illumination as a major noise source. Since in our experiment illumination was controlled, we cannot answer this question with empirical evidence. Neural networks were essentially inspired from studies of brain structure (Widrow et al., 1973), and they are known to be remarkably tolerant to noise in input data. However, further research is needed to integrate - in our system - pupillary light reflex information, which could be estimated by measuring illumination and other factors (e.g. Watson & Yellot, 2012).

Concerning EDA measures, we obtained contrasting results: event-related concomitants of stressful sounds - i.e. SCRs - were found, and Factor Analysis confirmed their relatively stable response behavior (75.52% of explained variance with two factors). However, NS.EDR freq. - an indicator of adverse emotional states in HCI (Boucsein, 2012) - did not return the expected results, in that it did not increase with stress level. It is known that NS.EDR freq. calculation can provide different results depending on event-detection algorithms: overlapping SCRs are likely, especially in applied contexts such as the present experiment. We tried to overcome this problem by extracting an area measure of EDA, but the results remain unclear: it might be that the stress level in our experiment is too low - according to subjective measures - to be indexed by electrodermal measures. Another possible and more "technical" explanation could be the fact that we placed the electrodes on the participants' foreheads - because of experimental constraints (i.e. driving task) - instead of using palms and soles as recording sites: Dawson et al. (2000) suggested that emotion-evoked sweating is more evident in palmar and plantar zones because of higher sweat gland density ($600/cm^2$ for palms, $700/cm^2$ for soles, $181/cm^2$ for the forehead; see Sato et al., 1989). Finally, recent studies suggested that EDA signals have less discriminating power -

compared to PD signals - for stress classification (Ren et al., 2013; Zhai & Barreto, 2006).

Subjective ratings showed moderate yet significant correlations with normalized average PD. Although this result is encouraging, we remark that relying only on ANS measures is not the key for automatic stress measurement. A widely accepted perspective states that emotions are organized according to two principal dimensions, i.e. arousal and valence (Mauss & Robinson, 2009). The former ranges from states of low activation (e.g. calm) to states of high activation (e.g. excited), while the latter counters positively-tuned states (e.g. happy) versus negatively-tuned ones (e.g. angry). While ANS measures are known to be reliable indexes of arousal, they don' t give us indications about valence (see for example Janisse, 1977). Thus, pupil diameter could increase because of positive stress (eustress) or negative stress, in the same way. In the present study, we examined the valence dimension by means of subjective ratings. However, this requires some active intervention by the user' s side, which is not suitable for an automatic stress measurement system. In this perspective, automatic valence indexes will be investigated in future research, with particular attention towards facial expressions: like for PD, these measures can be acquired unobtrusively by using video cameras.

REFERENCES

Bach, D.R., Friston, K.J., Dolan, R.J. (2010). Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations. *International Journal of Psychophysiology*, *76*, 52-55. doi: 10.1016/ijpsycho.2010.01.011

Barman, D., Chowdhury, N. (2012). A method for movie business prediction using back-propagation neural network. *International Journal of Information Technology and Computer Science*, *11*, 67-73. doi: 10.5815/ijitcs.2012.11.09

Beatty, J. (1982). Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin*, *91*, 276-292. doi: 10.1037/0033-2909.91.2.276

Beatty, J. (1986). The Pupillary System. In Coles, M.G.H., Donchin, E., Porges, S.W. (Eds.), *Psychophysiology: Systems, processes, and applications*, pp. 43-50. Guilford, New York, USA.

Beatty, J., & Lucero-Wagoner, B. (2000). The Pupillary System. In J.T. Cacioppo, L.G. Tassinary, & G.G. Berntson (Eds.), *Handbook of psychophysiology*, 2nd ed., pp. 142-162. Cambridge University Press, Cambridge, USA.

Benedek, M., Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*, 80-91. doi: 10.1016/j.jneumeth.2010.04.028

Benedetto, S., Pedrotti, M., Bridgeman, B. (2011). Microsaccades and exploratory saccades in a naturalistic environment. *Journal of Eye Movement Research*, *4* (*2*), 1-10.

Bergamin, O., & Kardon, R.H. (2003). Latency of the pupil light reflex: sample rate, stimulus intensity, and variation in normal subjects. *Investigative Ophthalmology & Visual Science*, *44* (*4*), 1546-1554. doi: 10.1167/iovs.02-0468

Bernick, N., & Oberlander, M. (1968). Effect of verbalization and two different modes of experiencing pupil size. *Perception & Psychophysics*, *3* (*5A*), 327-330. doi: 10.3758/BF03212478

Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). Springer, New York, USA.

Boucsein, W., Fowles, D.C., Grimnes, S., Ben-Shakhar, G., Roth, W.T., Dawson, M.E., Filion, D.L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, *49*, 1017-1034. doi: 10.1111/j.1469-8986.2012.01384.x

Bradley, M.M., Miccoli, L., Escrig, M.A., & Lang, P.J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*, 602-607. doi: 10.1111/j.1469-8986.2008.00654.x

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433-436. doi: 10.1163/156856897X00357

Cacioppo, J., & Tassinary, L.G. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, *45* (*1*), 16-28. doi: 10.1037/0003-066X.45.1.16

Czaja, S.J., Sharit, J. (1993). Stress reactions to computer-interactive tasks as a function of task structure and individual differences. *International Journal of Human-Computer Interaction*, *5*, (*1*), 1-22. doi: 10.1080/10447319309526053

Dawson, M.E., Schell, A.M., Filion, D.L. (2000). The electrodermal system. In J.T. Cacioppo, L.G. Tassinary, & G.G. Berntson (Eds.), *Handbook of psychophysiology*, 2nd ed., pp. 200-223. Cambridge University Press, Cambridge, USA.

Dennerlein, J., Becker, T., Johnson, P., Reynolds, C. and Picard, R. (2003). Frustrating computer users increases exposure to physical factors. In *Proceedings of the International Ergonomics Association*, pp. 24-29. Seoul, Korea.

Di Stasi, L., Catena, A., Cañas, J., Macknik, S.L., Martinez- Conde, S. (2013). Saccadic velocity as an arousal indexin naturalistic tasks. *Neuroscience and Biobehavioral Reviews*, *37* (*5*), 968-975. doi: 10.1016/j.neubiorev.2013.03.011

Ellis, C.J. (1981). The pupillary light reflex in normal subjects. *British Journal of Ophthalmology*, *65*, 754-759. doi: 10.1136/bjo.65.11.754

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861-874. doi: 10.1016/j.patrec.2005.10.010

Fujigaki, Y., Mori, Kazuko (1997). Longitudinal study of work stress among information system professionals. *International Journal of Human-Computer Interaction*, *9*, (*4*), 369-381. doi: 10.1207/s15327590ijhc0904_3

Gitelman, D. R. (2002). ILAB: A program for postexperimental eye movement analysis. *Behavior Research Methods, Instruments, & Computers*, *34*, 605–612. doi: 10.3758/BF03195488

Goldwater, B.C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, *77*, 340-355. doi: 10.1037/h0032456

Granholm, E., Steinhauer, S.R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *52*, 1-6. doi: 10.1016/j.ijpsycho.2003.12.001

Granholm, E., Verney, S.P. (2004). Pupillary responses and attentional allocation problems on the backward masking task in schizophrenia. *International Journal of Psychophysiology*, *52*, 37-51.

doi: 10.1016/j.ijpsycho.2003.12.004

Hamid, N.A., Nawi, N.M., Ghazali, R. (2011). The effect of adaptive gain and adaptive momentum in improving training time of gradient descent back propagation algorithm on classification problems. *International Journal on Advanced Science, Engineering and Information Technology*, *1* (*2*), 178-184.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2010). *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press.

ISO 26022, 2010. Road vehicles - Ergonomic aspects of transport information and control systems - Simulated lane change test to assess in-vehicle secondary task demand. ISO/TC 22/SC 13.

Jainta, S., Baccino, T. (2010). Analyzing the pupil response due to increased cognitive demand: An independent component analysis study. *International Journal of Psychophysiology*, *77*, 1-7. doi: 10.1016/j.ijpsycho.2010.03.008

Janisse, M.P. (1977) Pupillometry: The Psychology of the Pupillary Response. Hemisphere, Washington DC, USA.

Kleiner, M., Brainard, D.H., Pelli, D.G. (2007). What' s new in Psychtoolbox-3? Perception 36 ECPV Abstract Supplement.

Klingner, J. (2010). Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pp. 275-282. ACM, New York, USA.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 Symposium on Eye-Tracking Research & Applications*, pp. 69-72. ACM, New York, USA.

Kuchinke, L., Vo, M.L.H., Hofmann, M., Jacobs, A.M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*, 132-140. doi: 10.1016/j.ijpsycho.2007.04.004

Kuhlmann, J., & Böttcher, M. (1999). *Pupillography: Principles, methods and applications.* W. Zuckschwerdt Verlag, München, Germany.

Lanting, P., Bos, J.E., Aartsen, J., Schuman, L., Reichert-Thoen, J., Heimans, J.J. (1990). Assessment of pupillary light reflex latency and darkness adapted pupil size in control subjects and in diabetic patients with and without cardiovascular autonomic neuropathy. *Journal of Neurology, Neurosurgery, and Psychiatry*, *53*, 912-914. doi: 10.1136/jnnp.53.10.912

Lew, R., Dyre, B.P., Werner, S., Wotring, B., Tran, T. (2008). Exploring the potential of Short-Time Fourier Transforms for analyzing skin conductance and pupillometry in real-time applications. In *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*, pp. 1536-1540. Human Factors and Ergonomics Society, Santa Monica, USA.

Loewenfeld, I., Lowenstein, O. (1993). *The Pupil: Anatomy, physiology, and clinical applications* (Vol. I). Iowa State University press, Ames, USA.

Lüdtke, H., Wilhelm, B., Adler, M., Schaeffel, F., Wilhelm, H. (1998). Mathematical procedures in data recording and processing of pupillary fatigue waves. *Vision Research*, *38*, 2889-2896. doi: 10.1016/S0042-6989(98)00081-9

Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet respresentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11* (*7*), 674-693. doi: 10.1109/34.192463

Mauss, I.B., Robinson, M.D. (2009). Measures of emotion : a review. *Cognition and Emotion*, *23* (*2*), 209-237. doi: 10.1080/02699930802204677

Marshall, S.P. (2000). Method and apparatus for eye tracking and monitoring pupil *dilation* to evaluate cognitive activity. US Patent No. 6,090,051.

Marshall, S.P. (2002). The Index of Cognitive Activity: Measuring Cognitive Workload. In *Proceedings of the 7th Conference on Human Factors and Power Plants*, pp. 7.5-7.9. IEEE Computer Society.

Minin, L., Benedetto, S., Pedrotti, M., Re, A., Montanari, R. (2011). Measuring the effects of visual demand on lateral deviation: A comparison among driver' s performance indicators. *Applied Ergonomics*, *43* (*3*), 486-492. doi: 10.1016/j.apergo.2011.08.001

Minu, K.K., Lineesh, M.C., Jessy John, C. (2010). Wavelet neural networks for nonlinear time series analysis. *Applied Mathematical Sciences*, *4* (*50*), 2485-2495.

Mitsopoulos-Rubens, E., Trotter, M.J., Lenné, M.G. (2011). Effects on driving performance of interacting with an in-vehicle music player : A comparison of three interface layout concepts for information presentation. *Applied Ergonomics*, *42* (*4*), 583-591. doi: 10.1016/j.apergo.2010.08.017

Moré, J.J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. *Numerical Analysis*, *630*, 105-116.

Nakayama, M. (2006). Influence of blink on pupillary indices. *Biomedical Circuits and Systems Conference*, *2006*, pp. 29-32. IEEE Conference Publications. doi: 10.1109/BIOCAS.2006.4600300

Nakayama, M., & Shimizu, Y. (2002). An estimation model of pupil size for 'Blink artifact' and its applications. In Verleysen, M. (Ed.) *10th European Symposium on Artificial Neural Networks*, pp. 251-256. D-side publications, Bruges, Belgium.

Nakayama, M., & Shimizu, Y. (2004). Frequency analysis of task evoked pupillary responses and eye-movement. In Duchowsky, A.T., Vertegaal, R. (Eds.) *Eye-Tracking Research & Applications Symposium 2004*, pp. 71-76. ACM Press, New York, USA.

Nakayama, M., Yamamoto, K., Kobayashi, F. (2012). Estimation of sleepiness using pupillary response and its frequency components. *International Journal of Bioinformatics Research and Applications*, *8* (*5/6*), 342-365. doi: 10.1504/IJBRA.2012.049621

Nhan, B.R., Chau, T. (2010). Classifying affective states using thermal infrared imaging of the human face. *IEEE Transactions on Biomedical Engineering*, *57* (*4*), 979-987. doi: 10.1109/TBME.2009.2035926

Palinko O., Kun A.L. (2011). Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. In *Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pp. 329-336. Public Policy Center, University of Iowa, USA.

Palinko O., Kun A.L. (2012). Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. In *Proceedings of the 2012 Symposium on Eye-Tracking Research & Applications*, pp. 413-416. ACM, New York, USA.

Palinko O., Kun A.L., Shyrokov A. & Heeman P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-*

Tracking Research & Applications, pp. 141-144. ACM, New York, USA.

Partala, T., Surakka, V. (2003). Pupil size variation as an indicator of affective processing. *International Journal of Human-Computer Studies*, *59*, 185-198. doi: 10.1016/S1071-5819(03)00017-X

Pedrotti, M., Lei, S., Dzaack, J., Rötting, M. (2011). A data-driven algorithm for offline pupil signal preprocessing and eyeblink detection in low-speed eye-tracking protocols. *Behavior Research Methods*, *43* (*2*), 372-383. doi: 10.3758/s13428-010-0055-7

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442. doi: 10.1163/156856897X00366

Pinzon-Morales, R.D., Hirata, Y. (2012). Customization of wavelet function for pupil fluctuation analysis to evaluate levels of sleepiness. In *Proceedings of the 11th International Conference on Telecommunications and Informatics, Proceedings of the 11th International Conference on Signal Processing*, pp.115-120. World Scientific and Engineering Academy and Society, Stevens Point, USA.

Ren, P., Barreto, A., Gao, Y., Adjouadi, M. (2013). Affective assessment by digital processing of the pupil diameter. *IEEE Transactions on Affective Computing*, *4* (*1*), 2-14. doi: 10.1109/T-AFFC.2012.25

Sato, K., Kang, W.H., Saga, K., Sato, K.T. (1989). Biology of sweat glands and their disorders. I. Normal sweat gland function. *Journal of the American Academy of Dermatology*, *20* (*4*), 537-563. doi : 10.1016/S0190-9622(89)70063-3

Sauter, S.L. (1991). Job stress and human-computer interaction. *International Journal of Human-Computer Interaction*, *4* (*3*), 3-4. doi: 10.1080/10447319109526018

Schweitzer, M.B., & Paulhan, I. (1990). Manuel pour l'inventaire d'anxiété Trait-Etat (Forme Y). Laboratoire de Psychologie de la Santé, Université de Bordeaux II. Bordeaux, France.

Shastri, D., Merla, A., Tsiamyrtzis, P., Pavlidis, I. (2009). Imaging facial signs of neurophysiological responses. *IEEE Transactions on Biomedical Engineering*, *56* (*2*), 477-484. doi: 10.1109/TBME.2008.2003265

Shi, B., Moloney, K.P., Pan, Y., Leonard, V.K., Vidakovic, B., Jacko, J.A., Sainfort, F. (2012). Wavelet classification of high frequency pupillary responses. *Journal of Statistical Computation and Simulation*, *76* (*5*), 431-446. doi: 10.1080/10629360500107873

Spielberger, C.D., Gorsuch, R.L., Lushene, R., Vagg, P.R., & Jacobs, G.A. (1983). Manual for the State-Trait Anxiety Inventory. Consulting Psychologists Press, Palo Alto, USA.

Tournois, J., Mesnil, F., & Kop, J.L. (2000). Autotricherie et hétérotricherie: Un instrument de mesure de la désirabilité sociale / Self-deception and other-deception : A social desirability measuring tool. *Revue Européenne de Psychologie Appliquée / European Review of Applied Psychology*, *50* (*1*), 219-232.

Van den Broek, E.L., Janssen, J.H., Westerink, J.H.D.M. (2009). Guidelines for Affective Signal Processing (ASP): From lab to life. In *Proceedings of the IEEE 3rd international conference on affective computing and intelligent interaction*, ACII, Vol. 1, pp. 704–709. IEEE Press, Amsterdam, The Netherlands.

Verney, S.P., Granholm, E., Marshall, S.P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology*, *52*, 23-36. doi: 10.1016/j.ijpsycho.2003.12.003

Vo, M.L.H., Jacobs, A.M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, *45*, 130-140. doi: 10.1111/j.1469-8986.2007.00606.x

Watson, A.B., Yellot, J.I. (2012). A unified formula for light-adapted pupil size. *Journal of Vision*, *12* (*10*): 12, 1-16. doi: 10.1167/12.10.12

Widrow, B., Gupta, N.K., Maitra, S. (1973). Punish/reward: learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-3* (*5*), 455-465. doi: 10.1109/TSMC.1973.4309272

Wyatt, H.J. (2010). The human pupil and the use of video-based eyetrackers. *Vision Research*, *50*, 1982-1988. doi: 10.1016/j.visres.2010.07.008

Zhai, J., Barreto, A. (2006). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *Engineering in Medicine and Biology Society*, *2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 1355-1358. doi: 10.1109/IEMBS.2006.259421