Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment

Kurt Plarre[⋆], Andrew Raij[□], Syed Monowar Hossain[⋆], Amin Ahsan Ali[⋆], Motohiro Nakajima[‡],
Mustafa al'Absi[‡], Emre Ertin[⋄], Thomas Kamarck[∨], Santosh Kumar[⋆],
Marcia Scott[△], Daniel Siewiorek[†], Asim Smailagic[†], Lorentz E. Wittmers, Jr.[‡]
University of Memphis,[⋆] University of South Florida,[□] University of Minnesota Medical School,[‡]
The Ohio State University,[⋄] Carnegie Mellon University,[†] University of Pittsburgh,[∨]
National Institute on Alcohol Abuse and Alcoholism[△]

ABSTRACT

Repeated exposures to psychological stress can lead to or worsen diseases of slow accumulation such as heart diseases and cancer. The main challenge in addressing the growing epidemic of stress is a lack of robust methods to measure a person's exposure to stress in the natural environment. Periodic self-reports collect only subjective aspects, often miss stress episodes, and impose significant burden on subjects. Physiological sensors provide objective and continuous measures of stress response, but exhibit wide between-person differences and are easily confounded by daily activities (e.g., speaking, physical movements, coffee intake, etc.).

In this paper, we propose, train, and test two models for continuous prediction of stress from physiological measurements captured by unobtrusive, wearable sensors. The first model is a *physiological classifier* that predicts whether changes in physiology represent stress. Since the effect of stress may persist in the mind longer than its acute effect on physiology, we propose a *perceived stress model* to predict *perception* of stress. It uses the output of the physiological classifier to model the accumulation and gradual decay of stress in the mind. To account for wide between-person differences, both models self-calibrate to each subject.

Both models were trained using data collected from 21 subjects in a lab study, where they were exposed to cognitive, physical, and social stressors representative of that experienced in the natural environment. Our *physiological classifier* achieves 90% accuracy and our *perceived stress model* achieves a median correlation of 0.72 with self-reported rating. We also evaluate the perceived stress model on data collected from 17 participants in a two-day field study, and find that the average rating of stress obtained from our model has a correlation of 0.71 with that obtained from periodic self-reports.

IPSN'11, April 12–14, 2011, Chicago, Illinois.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]

General Terms

Algorithms, Experimentation, Human Factors, Measurement

Keywords

Wearable sensors, physiological monitoring, stress inference

1. INTRODUCTION

In moderation, stress can be a positive force in everyday life. It can motivate action (e.g., when in danger), improve performance, and increase excitement [24, 44]. However, excessive, chronic, and repeated exposures to stress can lead to significant negative health consequences [38, 29]. Excessive stress can lead to headaches, trouble sleeping, and fatigue [30, 11, 3]. In the longer term, stress can be associated with risk for several chronic diseases including cardiovascular diseases [37, 42]. Animal and human studies have shown that stress can also play a role in psychological or behavioral problems, such as depression, addiction, rage, and anxiety [22, 2, 13, 14]. The main challenge in addressing the negative consequences of stress is a lack of robust methods that can continuously measure a person's exposure to stress in the natural environment.

In behavioral science, periodic self-reports are commonly used to measure perceived stress in natural environments. Self-reports allow collection of instantaneous measurements of perceived stress, often multiple times per day to reach a desired sampling of stress. However, self-reports only capture subjective aspects of stress, may miss stress episodes, and impose significant burden on the subject.

Since self-reports only capture perception of stress, they do not provide a proximal measure of the physical health consequences of stress, such as cardiovascular wear and tear, ulcer, and cancer. In addition, the episodic nature of stress means that discrete self-reports can miss stress episodes. To ensure capture of stress episodes, a continuous measure of stress is needed. Finally, the active participation required to provide self-reports means that self-reports are burdensome and obtrusive. To provide a self-report, a person must be willing to have their daily life interrupted to complete self-reports, sometimes as many as 20 times per day. A high subject burden may lead to compliance issues and may affect the quality of measures collected. Thus, a passive approach to measuring stress that requires little attention of the subject would be a significant advancement.

Physiological measurements could provide the basis for a continuous and passive approach to measuring stress. However, physiological measures present other challenges. First, the physiological sensors must be unobtrusive, wearable, and provide scientifically valid measurements in natural environments. Second, events that occur naturally in daily life, such as eating, drinking, caffeine intake, conversation, and physical activity, are confounders. They affect physiology and can even mask the physiological response to stress. Third, wide between-person differences in the physiological response to stress make it difficult to build a simple stress classifier that works on a large population. Fourth, building such a physiological classifier requires collecting ground truth in natural environments. However, the most viable approach to collecting ground truth in natural environments are periodic self-reports. Their subjectivity and discrete characteristics limit the quality and quantity of ground truth that can be collected in natural environments.

To our knowledge, the literature does not yet address all of these issues, nor does it provide a passive, scientifically valid, and continuous measurement of stress that works in natural environments. Several attempts at measuring stress or emotion physiologically exist in the literature [20, 31, 26, 41, 25, 27, 36, 9, 21], but these measurement tools are not suitable for use in natural environments. They have been exclusively applied and tested in controlled or semi-controlled environments, where it is easier to collect ground truth and control for physiological confounders. Some recent work has tried to address between-person differences in physiology using a personalized classifier of emotion for use in controlled environments [40, 25]. Most personalization schemes require capturing calibration data in controlled environments from the individual to whom the algorithm will be personalized. However, calibration stages in controlled environments are not scalable.

In this paper, we present two models that each allow continuous prediction of stress from physiological measurements captured in natural environments. The first model is a *physiological stress classifier* that predicts whether a one minute measurement corresponds to a physiological response to a stressor. As this model directly captures the notion of physiological stress, it is useful as a proximal measure of health outcomes that result from ”wear and tear,” such as heart diseases. The second model is a *perceived stress model* that predicts the stress rating a subject would provide during a particular minute. In other words, the perceived stress model predicts whether a person *feels* stressed during a particular minute. The output of the perceived stress model is useful as a proximal measure of psychological or behavioral outcomes associated with stress, such as depression. Finally, converting self-reported stress to a binary stress state is not an obvious process due to subjectivity and wide between-person differences. We develop and train a simple classifier to detect the stress state from self-reports.

All three models are trained and tested using data from a 21 person lab study where participants were carefully exposed to three diverse and validated stressors (public speaking, mental arithmetic, and cold pressor challenges) while physiological data and self-reports were collected. The physiological data was captured using a newly developed, wearable, unobtrusive sensor suite called AutoSense [1] that provides electrocardiography (ECG) using 2 electrodes, respiration using a respiratory inductive plethysmograph (RIP) and 3-axis acceleration, among several others.

The physiological stress classifier is modeled using a variety of ECG features that have been previously shown to respond to stress, such as heart rate and heart rate variability. We complement these with additional features from respiration such as inhalation duration, exhalation duration, minute ventilation, and stretch (i.e., difference between peak and valley of a respiratory cycle). After removing outliers from the feature set, we normalize the feature values to account for the baseline of each individual. This makes the model self-calibrating for each individual. We then compute various statistics over these features such as mean, median, and quartile deviation, making for a total of 35 features.

The perceived stress model maps the output of the physiological stress classifier to perceived stress. Physiological changes induced by stress may decay rapidly so the body can quickly regain its allostatic balance, but the perception of stress may take longer to decay after a stress event. In addition, repeated exposures to stress may cause them to accumulate in the mind, which may take even longer to fade away. Therefore, to model perceived stress, we propose an accumulation and decay model. The accumulation and decay parameters are personalized to each participant to account for between-person differences. The perceived stress model can also be regarded as an aggregation model that aggregates the output of the physiological classifier for a given window of say, 10 minutes, to obtain the effect of the past 10 minutes on the current stress state of the individual.

In the lab, across 21 subjects, our physiological stress model obtains 90% accuracy, using only 13 (out of 35) features. Even when ECG or RIP are used in isolation (say, if only one of these sensors is functional), we obtain $> 85\%$ accuracy. The fact that we are able to obtain such high accuracy using RIP alone is a surprise given that the stress literature does not highlight respiration being as discriminatory of stress as ECG. In addition, the perceived stress model achieves a median correlation of 0.72 with self-reported ratings of stress provided by participants in the lab. The self-report classifier obtains an accuracy of 84%.

We also conducted an initial analysis of the generalizability of the lab-derived classifiers (physiological and perceived) to the field. Participants wore the sensor suite for two days (12-14 hours each day) as they went about their normal daily life. Throughout each day, approximately 25 self-reports of stress were collected. We applied the lab-derived perceived stress model to this field data after appropriate screening and cleaning. We found that the average rating produced by the perceived stress model has a correlation of 0.7 with the average rating of stress provided by each subject in the field. Figure 1 depicts how the three models are constructed and applied.

To our knowledge, this work provides the first classifier of stress that can be readily used in natural environments without pre-calibration. This innovation was made possible because of the development of the AutoSense wearable sensing suite which we could use to collect measurements both from a rigorous lab stress protocol (that has been repeatedly validated in behavioral science) and from the natural environment of the same individuals.

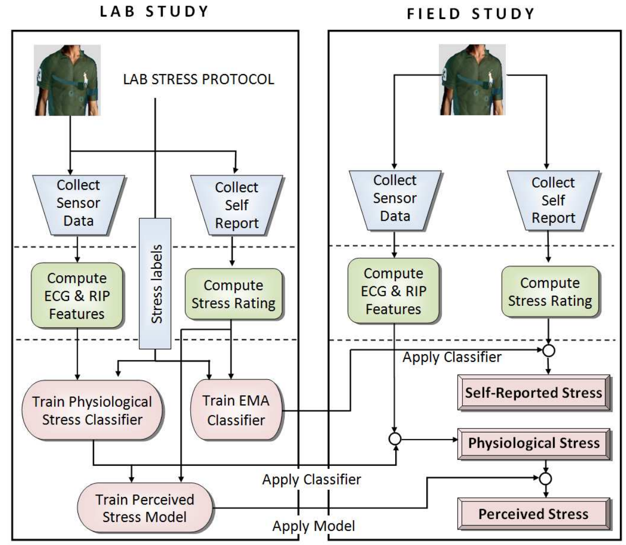

Figure 1: Three models - physiological stress classifier, perceived stress classifier and the self-report (or EMA) classifier - constructed and trained from data collected in the lab (left side of the figure). These classifiers are applied to the data collected in the field (right side of the figure) to obtain three distinct measures of stress.

Potential New Applications. In addition to enabling the self-monitoring of stress by individuals and the study of stress by behavioral scientists, development of a stress classifier opens the door for several new applications. First, real-time inferences of stress could be used to trigger **timely interventions** relevant to the user (e.g., the phone could play a "soothing" song) when a user's stress level is too high. Second, **reactivity to an intervention** - how it changes physiology and stress levels, could also be measured in real-time. This would enable personalized selection/evaluation of interventions in the field. Third, common, everyday **interruptions** - a significant source of stress - could be managed by the phone based on the user's current stress level. For example, a call from one's boss might be routed to voice mail if previous measurements indicate a call from the boss when at home leads to excessive stress [12]. Last, but not least, stress measurements could also be used as part of a system that extracts and uses subjective information about a person from her sensor data (subjective sensing [28]). For example, real-time measurements of stress could be linked to a subjective navigation system which selects a longer, but less stressful, route for driving to work.

Organization: Section 2 describes some related work. The lab and field study designs are presented in Section 3. Sections 4, 5 and 6 present the design, development, and evaluation of the physiological, perceived, and self-reported stress classifiers, respectively, on the lab data. Section 7 presents the results of applying our models to the field data. Section 8 concludes the paper and points out several opportunities for future work in this area.

2. RELATED WORK

William James raised provoking questions on the relation between physiology and psychology in 1890 [23]. John Cacioppo and others subsequently revitalized the interest in predicting psychological state from physiological measurements [10]. Over the past two decades several markers of stress have been identified that are activated by stress. These include heart rate, heart rate variability, respiratory sinus arrythmia (RSA), respiratory patterns, electrodermal response and blood pressure [26, 25, 21]. While it has been shown that these features do respond to stress, they may be activated by other demands on the body such as speaking, change in posture, physical activity, etc. Hence, using these measures to predict stress has been exceedingly difficult.

The first challenge is the availability of an unobtrusive, wearable sensor system that can collect measurements from multiple modalities and process them on the body. Leveraging recent developments in wearable sensing and smart phones, we have developed the AutoSense [1] sensor suite that collects ECG, respiration, activity, and other measurements and wirelessly transmits them to a smart phone.

The second challenge is to account for confounding factors that may overwhelm the changes in physiology caused by change in stress level. From 1996 onwards, Myrtek and colleagues attempted to predict changes in emotion from physiological measurements, but they did not find significant correlations between those exhibited by physiology and those collected in self-reports [31]. The main hypothesis for the lack of correlation was the presence of confounders. In even the most recent attempts in inferring emotion in the natural environment, only those measurements that were collected close to the markings provided by subjects were used, due to a lack of ground truth available for the rest of the data [20].

Most recent work has focused on inferring emotion from physiological measurements [26, 41, 25, 27, 36, 9]. In most of these protocols, only measurements collected when specific emotions are experienced by subjects (e.g., when seeing pictures/videos or listening to music) are used for classification. The classifiers developed in these studies can't be applied to infer emotion in the natural environment since it is not know how well these models can distinguish measurements when emotions are experienced from those when emotions are not strong, which was the challenge encountered in [31]. Furthermore, work on emotion classification cannot be directly applied to detecting stress, since each emotion classification targets a specific set of emotions and does not cover all negative emotions that may constitute stress.

There has been some work on detecting stress, most notably [21] in which four drivers wore physiological sensors and drove on highway and non-highway city streets. On average, driving on city streets was more stressful than highway driving, which was more stressful than being parked in a garage. The work showed that selected 5-minute segments of driving, regarded by human raters to correspond to low, medium, and high stress, could be classified with 97% accuracy from physiological measurements. However, since the labeling of 5-minute driving segments is based on human raters, and these segments may not constitute validated stressors, these results do not generalize to other real-life situations. In contrast, the stressors used in our work have been well-validated in various scientific studies[4, 5, 7, 6]. A limited number of subjects (i.e., 4) also limit the applicability of the model in [21] to a wider population. Finally, the model presented in [21] is not evaluated in unsupervised natural life conditions.

In a preliminary work, we trained a support vector machine to classify stress using a similar data set as used in this work [40]. However, several issues with this initial approach emerged after more extensive analysis. First, self-reported ratings of stress were used as ground truth to train the classifier. Self-reports are inherently subjective and sometimes inaccurate, and thus may not represent an ideal ground truth. Second, a single threshold was applied to self-report ratings from all subjects to classify them into stressed and not-stressed. Given that all participants in this work were exposed to the same lab stressors (which have repeatedly been shown to elicit stress in most subjects), lab stressors are used as ground truth in this work, instead of self-reports.

The third challenge in inferring stress from physiological measurements is accounting for wide between-person differences. It has been observed in several recent works that a personalized model produces better accuracy than a population-level model [25, 40]. Although personalized models produce better accuracy, they are not as practical since they require collecting training data on each subject to produce the personal classifier.

In summary, the work presented in this paper is, to the best of our knowledge, the first one to provide a population-level classifier that calibrates itself to each subject and provides 90% accuracy in predicting stress from physiological response under a variety of real-life stressors. In addition, this is the first work to provide a perceived stress model to map physiological response of stress to perceived stress. Finally, although self-report has been used extensively to collect subjective experience of stress, this is the first work that provides a classifier to classify self-reported ratings into stressed and non-stressed categories.

3. DATA COLLECTION USER STUDY

We conducted a two-phase user study to collect training and test data for the three models of stress. In the first phase, physiological and self-report measures were collected from 21 participants while they were subjected to known, validated stressors in a lab setting. The controlled exposure to stressors in the lab provided the training data needed to develop the stress models. In the second phase, physiological and self-report measures were again collected from the same participants, but in their natural environment over two separate days. This section describes the study in more detail, including the population from which the participants were selected, the measures collected from them, and the procedure they followed in the lab and field to collect the data needed to train and test the models.

3.1 Participants

Participants were recruited via flyers posted in the Duluth campus of University of Minnesota. Potential participants completed a preliminary phone screening interview to confirm initial eligibility. The interview included questions concerning current or recent history of medical or psychiatric disorders, medication intake, and health related behavior (e.g., smoking, drinking). Those who met the initial screening requirements were invited to an on-site screening. In the on-site screening, participants were asked if they had any history of a major illness or psychiatric disorder, weighed within ±30% of Metropolitan Life Insurance norms, consumed two or less alcoholic drinks a day, and did not routinely use prescription medications (except contraceptives). Participants read and signed a consent form approved by the Institutional Review Board and completed the laboratory portion of the study. Participants received monetary compensation for their participation. Twenty-one college students (mean age ± SD: 20.6 ± 1.9) were recruited. Half of them were women and the majority were Caucasian (96%).

3.2 Measures

Sensory Measures. The AutoSense wearable sensor suite (shown in Figure 2) was used to monitor cardiovascular, respiratory, and thermoregulatory systems, systems known to respond to stress and other psychologically and physically demanding conditions. Six sensors were used: 1) an electrocardiograph (ECG) attached to the body with two electrodes to measure electrical output of the heart, 2) the ECG electrodes were used to measure skin conductance, 3) a skin temperature thermistor attached to the skin mid thorax, 4) an ambient temperature sensor, 5) a three-axis accelerometer, and 6) a respiratory inductive plethysmograph (RIP) band to measure relative lung volume at the rib cage. The sensors were integrated onto two wireless motes[1]. One mote was dedicated to the RIP sensor and the second mote hosted all other sensors. Each mote is 2.5 square-inches and powered by rechargeable 750 mAh batteries. The lifetime for streaming raw data on a wireless channel is up to 72 hours for moderate datarate (60 samples/node/sec). The system also uses an 802.15.4-to-Bluetooth bridge that sends the data received from sensors to a mobile phone via Bluetooth. More details on AutoSense is available at [1].

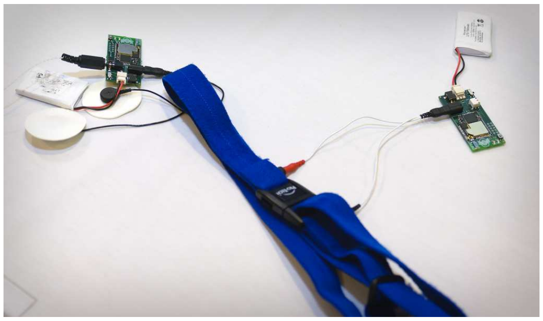

Figure 2: The AutoSense wearable sensor system includes six sensing modalities including ECG, respiratory inductive plethysmograph (RIP), and three-axis accelerometer.

Self-Report. In both the lab and field studies, participants completed questions describing their subjective stress state on a mobile phone. Responses to the questions were synchronized to the physiological data. Five questions were used: 1) Cheerful?, 2) Happy?, 3) Angry/Frustrated?, 4) Nervous/Stressed?, and 5) Sad?. Each item was answered on a four-point scale: 0 (NO), 1 (no), 2 (yes), and 3 (YES).

3.3 Lab Procedure

To prepare for the laboratory session, participants were asked to wear or bring a comfortable fitting shirt and not wear metal objects or accessories on the session day (in the

[1] A mote is a computing platform which contains a microcontroller, wireless radio and some sensors

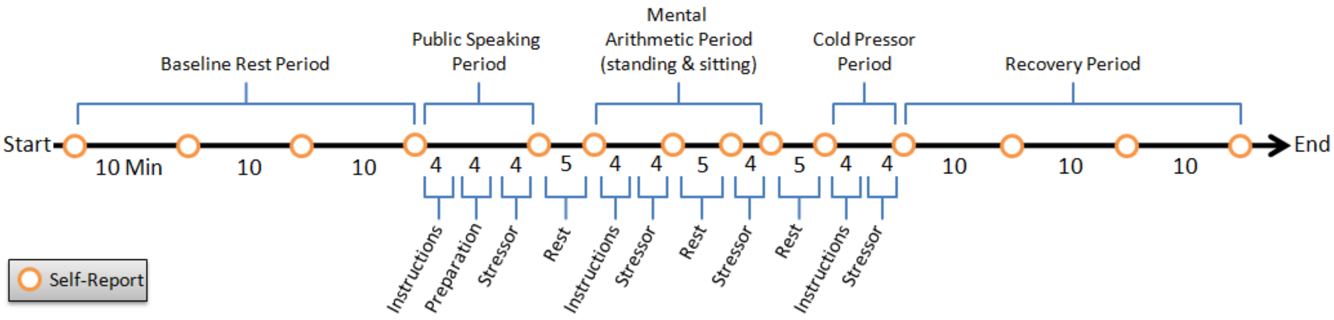

Figure 3: The Lab Study Procedure

unlikely event that they could interfere with the sensors). As several factors can affect physiological signals besides stress, additional requirements were added to the study to control for these factors. Participants were instructed to avoid caffeine, tobacco, and strenuous exercise at least 4 hrs before the beginning of the session, alcohol 24 hrs prior to the session, and pain medication (including over the counter drugs such as aspirin) for at least 72 hours before the session. If a participant did not meet the criteria, he or she was rescheduled.

Figure 3 summarizes the lab protocol. The lab session lasted approximately two hours. Participants began the session by reading and signed the Informed Consent Document. If consent was given, the participant was fitted with the sensors. The sensors continuously collected physiological signals throughout the remainder of the session. Next, the participant had a 30 minute baseline rest period. During the rest period, the participant relaxed on a recliner and watched neutral nature programming. Following the rest period, the participant was exposed to three stressors — public speaking, mental arithmetic, and cold pressor (presented in this order). These stressors were chosen because they are representative of the social, cognitive, and physical stressors experienced in natural everyday life and are known to induce stress in most people [4, 5, 7, 6].

Public Speaking: During the public speaking stressor, the participant was asked to deliver a 4 minute speech which was preceded by 4 minutes of silent preparation. To increase the social stress inherent in the task, the participant was told that his or her speech would be videotaped and subsequently evaluated by staff members at a later time.

Mental Arithmetic (sitting and standing): This stressor required the participant to continuously add the digits of a three-digit number and add the sum to the original number. As physiology is affected by posture, the arithmetic task was presented in two segments, seated and standing. This allowed training the stress model to be resistant to changes in posture. The order of the seated and standing tasks were counterbalanced to control for a potential learning effect.

Cold Pressor: After the recovery from the mental arithmetic, the participant was asked to insert his or her dominant hand in ice-cold water up to their wrist. The session lasted 90 seconds, unless the participant decided to pull his/her hand out earlier.

After each stressor, the participant was given a five minute break before beginning the next stressor. The rest periods allowed the participant's physiological response to the stressor to partially subside before exposure to the next one. After the last stressor, the participant underwent a 30 minute recovery phase similar to the baseline rest period experienced earlier in the session. The participant was then scheduled for subsequent field sessions.

Self-Report Schedule: Fourteen self-reports were administered in total throughout the lab session. The self-reports were scheduled to capture both subjective stress and nonstress states. The first self-report was completed at the beginning of the lab session, immediately after putting on the sensors. Next, self-reports were administered at 10 and 20 minutes into the 30 minute baseline rest period. Self-reports were also administered before and after each stressor. Finally, an additional three self-reports were administered every ten minutes during the 30 minute recovery period.

3.4 Field Procedure

The same participants returned to the lab on two separate days for field study. They were outfitted with the sensors and given a mobile phone to carry with them for the day as they went about their normal life. The mobile phone stored physiological samples captured from the wireless sensors. In addition, the phone periodically prompted the participant to complete self-report questionnaires (approximately 25 per day). We use the field data to test our model of psychological stress derived from the lab data.

4. MODELING PHYSIOLOGICAL STRESS FROM SENSOR MEASUREMENTS

We train and test a physiological classifier using features derived from one-minute measurements from ECG and RIP, both of which were sampled at 64 HZ. We achieve 90% classification accuracy on lab data using both RIP and ECG features together. Remarkably, we attain only marginally lower accuracies using just one or the other modality. Previous work has highlighted the discriminatory power of ECG for stress, but to our knowledge, this work is the first to show that RIP features alone can be used to classify stress accurately. Thus, either modality could be used if the other is not available. The respiration band is less burdensome than ECG electrodes and thus provides the best combination of wearability and accuracy needed to deploy stress classification in the field. In addition, respiration measurements can be used to infer other human states such as conversation [35, 34]. We also found that classification accuracies improve when a simple within-person normalization is applied to the features. The remainder of this section describes how features are selected and computed, and the classification results in more detail.

4.1 Feature Selection and Computation

For selecting features, we considered features usually reported in the literature as distinguishing for stress. In addition, for respiration, which is not as extensively investigated, we worked with physiologists in identifying new features. We selected those that were found to be distinguishing either visually or when used in a classifier. We now present details of the features that were selected. We note that although we collected skin conductance and temperature measurements in addition to ECG and RIP, we do not use these measurements in our analysis. Analyzing the effects of adding these measures is a subject of future work.

4.1.1 ECG Features

For ECG, all features are derived from windows of RR intervals — the time duration between successive R peaks in an ECG signal. RR intervals correspond to the duration between heart beats (see Figure 4), and thus can be used to calculate several statistical features describing the behaviors of the heart. We derive four additional features from RR intervals for each minute of data — the ratio between low and high-frequency components of heart beat and heart beat frequency in 3 bands (low, medium, and high).

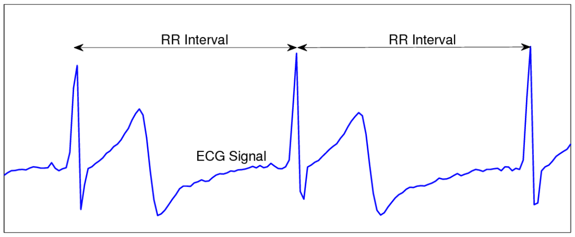

Figure 4: An RR interval is the duration between two successive R peaks in the ECG signal.

Preprocessing for Training: Several preprocessing steps are taken to prepare the data for training. As ECG responds very quickly to stressors (Figure 5), careful synchronization of physiological measurements to the timing of stressors is done before labeling windows. The key issue here is to ensure that data lost in wireless transmission do not affect synchronization. RR intervals are then computed using the Tompkin's algorithm [32]. Next, RR intervals that are found to be more than 2 standard deviations (SD) away from the mean of each minute, are flagged as outliers and ignored. The RR intervals of each subject have been shown to be normally distributed [8] and hence we adopt this method of filtering outliers. For increased robustness, we use quartile deviation (QD) in place of standard deviation (SD). For normal distribution, $2SD = 3.32QD$, and hence we use $3.32QD$ for identifying outliers in each minute. In addition, we ignore one minute of data immediately following each self-report. The latter is done to remove the effect of interruption induced by self-report prompts, which acutely affects physiology as seen in Figure 5. After outliers are removed, the base features are normalized to account for between-person differences before computing any statistical features. After the features are computed, each one-minute window is labeled with the ground truth. Windows within the period of the lab stressors are labeled as stressed, and all other windows are labeled as not stressed.

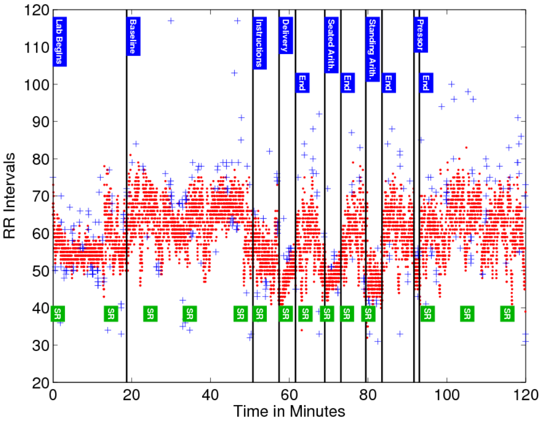

Figure 5: RR intervals (red dots) obtained from a participant during the lab session. Outliers (blue crosses) are removed from the dataset before computing statistical features, as is each minute immediately following a self-report (SR moments are marked in green). Physiology changes immediately after self-reports due to interruption from self-report prompt.

4.1.2 Respiration Features

Computation of the respiration features involves the identification of each respiration cycle, which is composed of an inhalation and an exhalation period. Thus a respiration cycle starts from a valley that corresponds to the start of an inhalation phase and ends at another valley that marks the start of the next inhalation (see Figure 6).

We investigated 7 different features computed from the respiration signal (see Figure 6 for illustration). **Inhalation Duration** corresponds to the time elapsed from a valley of a signal to the next peak, which denotes the maximum expansion of the chest in the respiration cycle. **Exhalation Duration** corresponds to the time duration between the peak and the next valley. **Respiration Duration** is the sum of the inhalation and exhalation duration, i.e. the duration of a breath. **IE Ratio** is the ratio of inhalation duration to exhalation duration. **Stretch** is the difference between the amplitude of the peak and the minimum amplitude the signal attains within a respiration cycle (see Figure 6). **Minute Ventilation/Minute Volume** is the volume of air inhaled (inhaled minute volume) or exhaled (exhaled minute volume) from a person's lungs in one minute. We use the inhalation minute volume as a feature. We estimate it by computing the area under the curve of the inspiration phase of a respiratory cycle. **Breath Rate** is simply the number of breath cycles per minute.

Respiratory sinus arrhythmia (RSA) is another feature sometimes used in emotion classification (e.g. [41]). It is a multimodal feature derived from both ECG and respiration that describes the variability in RR intervals due to respiration; inspiration shortens RR intervals and expiration grows RR intervals. It is computed by subtracting the shortest RR interval from the longest RR interval within each respiratory cycle, using the peak-valley method [17].

Preprocessing for Training: As a preprocessing step, the incoming respiration signal is segmented into segments of uniform duration. We choose the duration of 1 minute for this purpose and discard segments that are missing more than 15% of their samples due to packet loss. We adapt the peak-valley detection algorithms presented in [43, 45] to identify the peak and valley in each respiratory cycle and mark the start and end of a respiratory cycle. In order to remove spurious peaks, we set a threshold for peaks. From experimentation, we find that setting this threshold to the 75th percentile of the signal amplitudes for each window works well. We also require the duration between two successive peaks be at least 1.5 seconds. Through visual inspection, the performance of the peak-valley detection algorithm is found to be satisfactory.

For a one-minute window, we obtain one measure for breath rate and minute ventilation. For all the other features the number of measures obtained correspond to the number of breath cycles found in each window. To account for between subject differences, we normalize each feature by accounting for the mean value of the corresponding feature for each subject.

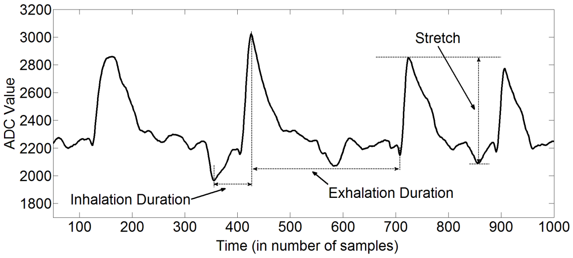

Figure 6: Three base features are computed from a respiration signal — inspiration duration, expiration duration, and stretch. They are shown for respiration measurements collected during speaking.

4.1.3 Statistics over Features

To reduce the effect of noise and outliers (e.g. spikes in the respiration signal due to movement) we compute four statistics on those features for which we have multiple values in each minute (e.g., RR intervals, stretch, RSA, etc.). We compute the mean, median, quartile deviation, and 80th percentile of the normalized features. The 80th percentile attempts to capture close to the highest value of the corresponding feature in each minute, while discounting extreme values. Overall, we compute a total of 35 features that are used to train the classifiers.

4.2 Classifiers and Evaluation Metric

Selected features and ground truth are used in WEKA [19] to train the classifiers. We trained three types of classifiers, a J48 decision tree, a J48 decision tree with adaptive boosting (AdaBoost), and a support vector machine (SVM). We chose the J48 decision tree because it is simple to implement and requires minimal computational resources compared to other classifiers [33], making it appealing for use on a smart phone. Adaptive boosting is a generalized technique for improving the performance of classifiers that does not require significant additional computational resources for classification [16]. SVMs are known to perform well for high dimensional feature space because they find the maximum separation between classes in a feature space [39]. We use 10-fold cross validation to obtain the performance measures of all three classifier types[2].

Classifier performance is measured using accuracy, kappa, confusion matrices, as well as receiver operating characteristic (ROC). Accuracy is defined as the number of correctly classified windows divided by the total number of windows. Kappa measures the correlation between predictions and ground truth, taking into account the probability that agreement comes from chance. The confusion matrix contains the number of true positives and true negatives on the diagonal and the false positives and false negatives off the diagonal. ROC is a graphical plot of the sensitivity, or true positive rate vs. false positive rate for a binary classifier system as its discrimination threshold is varied.

4.3 Classification Results on Lab Data

We now present the results of applying and evaluating our physiological classifier on the lab data. The number of valid datapoints available from the lab study was 929 minutes (271 classified as stress, and 658 as baseline). To avoid problems with unequal sample sizes, the sample sizes were equalized before training the classifiers, by selecting a random subsample of 600 minutes (271 stress and 329 baseline). Tables 1 and 2 show the performance of several classifiers trained on all 35 features. In the following, we describe the classification accuracy when ECG and RIP are used in isolation, impact of normalization on accuracy, classification accuracy for individual stress tasks, and accuracy when using only a select subset of features.

Classification Accuracy when Using ECG or RIP in Isolation: Figures 7 and 8 show the performance of classifiers when trained only on ECG or RIP features. Surprisingly, both features were highly discriminatory of stress. Training with ECG features alone produced a SVM classifier with 86% accuracy. Training only with respiration features led to 87% accuracy.

Effect of Normalization: Normalization generally improves the accuracy of classification. Tables 1 and 2 show the performance of several classifiers trained on all 35 features, with and without normalization. We obtain 90% accuracy using a J48 decision tree with Adaboost trained on normalized features (Table 1). Without normalization, the decision tree's accuracy decreases to 88%. Normalization decreases the number of false negatives (instances of stressed misclassified as not stressed) and false positives (instances of not stressed misclassified as stressed). Normalization also improves the performance of other classifiers (see Table 2). For normalization, we only account for the mean. Accounting for the standard deviation did not lead to better accuracy.

Classification Accuracy for Individual Stress Tasks: We found that it is possible to achieve accuracies of 95% or greater if the goal is to classify stress for specific stressors rather than a wide variety of stressors. Figures 7 and 8 show how the classifier performs if trained and tested on data from individual stressors. Using the speaking stressor data, accuracy as high as 99% can be obtained using only respiration features. This is likely because of the unique signature of speaking in respiration patterns [34] (i.e., speaking can

[2] Dividing the data into training (66% of the data) and testing data, we obtain 92% accuracy using 13 selected features.

	Accuracy	Kappa	Confusion Matrix
Not Normalized	88.00%	0.7574	NS 295 34 S 38 233
Normalized	90.17%	0.8010	NS 303 26 S 33 238

Table 1: Performance of a J48 Decision Tree with Adaboost trained on all features with and without normalization. In the confusion matrices, NS means not stressed and S means stressed.

	J48 Decision Tree	J48 with Adaboost	SVM
Not Normalized	82.33%	88.00%	88.17%
Normalized	87.67%	90.17%	89.17%

Table 2: Accuracies of three different classifiers trained with normalized and unnormalized features.

be detected with high accuracy and public speaking causes stress). In [34], we find that several of the features used to discriminate stress here are also discriminatory of speaking. In addition, classification rates of >95% can be obtained for mental arithmetic while standing using both RIP and ECG features.

Classification Accuracy when Using Only Selected Features: Figures 7 and 8 shows the accuracy of a classifier trained using the 13 most distinguishing features chosen by a correlation-based feature selection algorithm [18]. The selected RIP features were minute ventilation, mean and median of inspiration duration, quartile deviation of respiration duration, median of IE ratio, and median and 80th percentile of stretch. For ECG, the selected features were Heart Rate Power in Bands 1 and 3, and mean, median and 80th percentile of RR intervals. In addition, the 80th percentile of RSA was also selected. Using a smaller set of features does not reduce performance significantly. In fact, sometimes accuracy improves (see Figures 7 and 8). Therefore we use the smaller feature set for implementation on the mobile phone for real-time detection of stress level in the field.

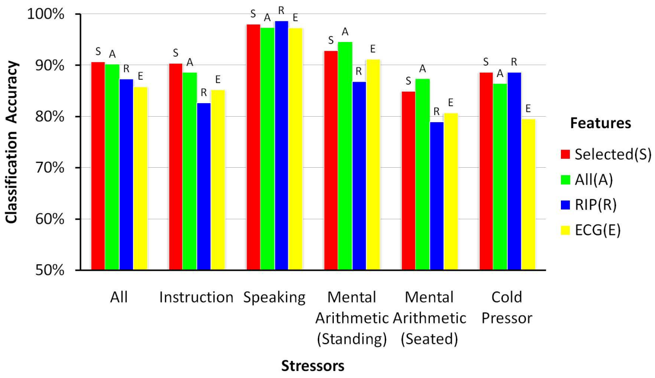

Figure 7: Best possible accuracies of classifiers trained on features computed over all and individual stressors. Accuracies are shown for using ECG features alone, RIP features alone, both ECG and RIP together, and using a smaller set of features selected by a feature selection algorithm.

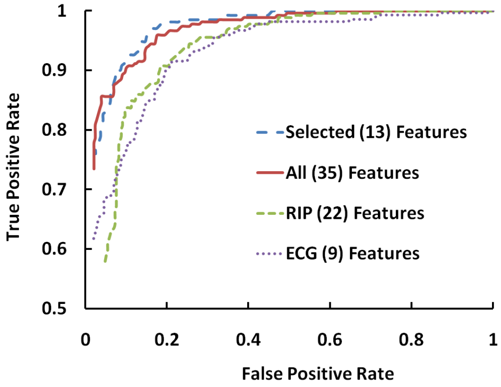

Figure 8: ROC curves for the best classifiers when using all 35 features, 13 selected features, only the 9 ECG features, and only the 22 RIP features. The curves show that performance of classifiers using all features and a subset of features are comparable with Area Under ROC Curve (AUC) being 0.964 and 0.967 respectively (with 1 being the best possible value). Also, RIP features have similar discriminatory power as ECG features (AUC for ROC curves being 0.924 and 0.926 respectively).

5. PERCEIVED STRESS MODEL

The perceived stress model maps physiological stress to perceived stress. The perceived stress model considers the value of perceived stress as hidden states in a Hidden Markov Model (HMM) that transitions among "stressed" and "non-stressed" states at each minute, treating the output of the physiological classifier as observables. We then estimate the probability of the current minute being stressed as a linear function of the observation from the physiological classifier for the current minute and the probability of the previous minute being stressed. Each time the physiological classifier marks a minute of data as "stressed," the perceived stress score of the participant is increased by an accumulation factor. With each passing minute, this score decays at an individual specific exponential rate. To account for wide between-person differences, the model allows the accumulation and decay rates to be personalized to each subject using their self-report ratings. We describe the model below, and then evaluate its performance on the lab data.

5.1 Model Definition

Let $s_k \in \{0, 1\}$ denote the perceived stress at discrete time k (in our case, time is measured in minutes), with $s_k = 1$ denoting *the subject is perceiving stress*, and $s_k = 0$, *the subject is not perceiving stress*. Let $x(k) \in \{0, 1\}$ be the physiological stress, i.e., the output of the classification algorithm. Our goal is to estimate the rating of perceived stress, which we model as $\pi_k(i) = \Pr[s_k = i | x_0, \ldots, x_k]$, for $i \in \{0, 1\}$ (with appropriate scaling). To do so, we use a Hidden Markov Model (HMM), with hidden states s_k, observations x_k, transition and emission probabilities given by

$$a(i,j) = \Pr[s_k = i | s_{k-1} = j],$$
$$b(i,j) = \Pr[x_k = i | s_k = j],$$

and prior probability $a_0(i) = \Pr[s_0 = i]$, as the starting point.

The posterior distribution of s_k satisfies

$$\pi_k(i) = \frac{\sum_{j=0}^{1} a(i,j) b(x_k,i) \pi_{k-1}(j)}{\sum_{i,j=0}^{1} a(i,j) b(x_k,i) \pi_{k-1}(j)}. \qquad (1)$$

By defining $\pi_k = [\pi_k(0), \pi_k(1)]^T$, so both values of i are accounted for in one common expression, we can rewrite (1) as $\pi_k = B(x_k) A^T \pi_{k-1}$, where A is the transition matrix, and $B(x_k)$ is a diagonal matrix containing the emission probabilities. By observing that $\pi_k(0) = 1 - \pi_k(1)$ for all k, we can show that

$$\pi_k(1) = \frac{\alpha_k^1 \pi_{k-1}(1) + \beta_k^1}{\alpha_k^2 \pi_{k-1}(1) + \beta_k^2}, \qquad (2)$$

where α_k^1, α_k^2, β_k^1, and β_k^2 are functions of the transition and emission probabilities, and of the output of the physiological stress classifier, x_k. In order to avoid overfitting and simplify the model, we approximate (2) by a linear recursion. To do so, we observe that, if x_k is fixed (say, $x_k = j$, for all k and $j \in \{0,1\}$), (2) has a unique stable equilibrium point (as a consequence of the Perron-Frobenius theorem). Let $\bar{\pi}^j$ be these two points for $j \in \{0,1\}$. We can linearize (2) around the two equilibrium points to obtain

$$\hat{\pi}_k = \alpha^j (\hat{\pi}_{k-1} - \bar{\pi}^j) + \bar{\pi}^j,$$

where $\hat{\pi}_k$ approximates $\pi_k(1)$ and $j = x_k$. We know that both of these recursions are stable because (2) is stable. To further simplify the model, we reduce the number of parameters to be estimated to two. We identify $\bar{\pi}^0$ with 0 (by translating the coordinate system) to represent that if the physiological classifier produces a long string of 0's, then the stress rating is assumed to be 0. In addition, we set $\alpha^0 = \alpha^1 = \alpha$ (assuming a common slope for both linear approximations) and $(1 - \alpha)\bar{\pi}^1 = \beta$ to obtain the final model

$$\hat{\pi}_k = \alpha \hat{\pi}_{k-1} + \beta x_k.$$

The value of $\hat{\pi}_k$ is initialized to the average rating of stress reported in the first self-report considered in the analysis (we eliminated the self-reports before the baseline period). This average is over the ratings provided by the subject for all stress related questions (with appropriate reverse coding of positive questions).

The α parameter models the decay of perceived stress in a person's mind while the β parameter models the accumulation of perceived stress, due to repeated exposures to stressors. Trough this, we attempt to describe the fact that, if the stress classifier outputs "stress" right before a self-report, we expect that with high likelihood, the answers to that self-report to indicate stress, and similarly, if a strong stressor caused the classifier to indicate stress for several consecutive minutes, we expect a self-report to indicate stress, even several minutes after the stressor has ended.

For the lab data, we take the initial time, k_1, to be the time of the second self-report, and the initial condition $\hat{\pi}_{k_1}$ as the value of the perceived stress obtained from the self-report at that time. We discard the first self-report because it was taken before the baseline period, and factors such as physical activity before the start of the lab session might confound the physiological classifier.

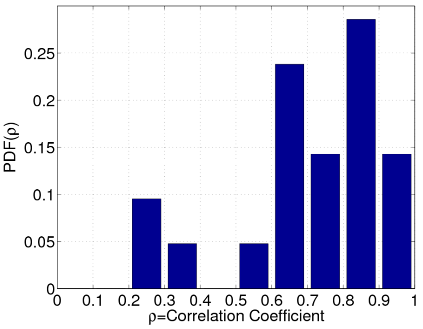

Figure 9: The probability density across 21 participants for ρ, the linear correlation coefficient between the output of the perceived stress model and self-reported rating of stress in the lab. The median correlation is 0.72.

We use the lab data to estimate the model parameters (α and β), using the least squares method. Let $\bar{k}_i$, $i = 1, \ldots, m$, with $\bar{k}_1 = k_1$, be the times at which self-reports were taken, and s_i^* the value of perceived stress, estimated from the responses. We define the cost function

$$J(\alpha, \beta) = \sum_{i=1}^{m} \left(\hat{\pi}_{\bar{k}_i} - s_i^*\right)^2.$$

The optimum values of α and β for each subject can be found using any suitable optimization method.

5.2 Evaluation of the model on the lab data

Unlike the physiological classifier which was trained and tested on labeled data, the perceived stress model is not a classifier. Rather, it aims to predict the self-reported rating of stress. Hence, to evaluate the model on the lab data, we compute the values of $\hat{\pi}_k$ at each time $\bar{k}_i$, for each subject, and then compare these values to those obtained from the self reports by computing their correlation coefficient. Figure 9 shows a histogram of the correlation coefficients obtained from comparing the perceived stress model, with the optimal values of α and β, and self-reports, for 21 participants. We find that the median correlation is 0.72, denoting reasonable agreement, given a large number of participants.

6. STRESS FROM SELF-REPORT

Self-reports provide subjective stress ratings on a scale of 0 (not stressed) to 3 (stressed) rather than a clear binary classification of stress. Converting self-reported stress to a binary stress state is not an obvious process. Participant self-reports are subjective and thus exhibit between-person differences. For example, in the lab, the mean self-reported stress rating after exposure to stressors was .97 (actually corresponding to not stressed) with a standard deviation of .48 (min: 0.17, max: 2.25). Even when accounting for each individual's baseline stress levels, the mean percent difference of self-reported stress from the baseline was 100% with a standard deviation of 60% (min: 0, max: 218%). These wide between-person differences must be accounted for before we can use field self-reports as ground truth.

We take a machine learning approach to this problem by training a self-reported stress classifier that takes as input an individual's subjective self-reports and classifies each into one of two states, not stressed or stressed. To train the classifier, we compute features over the lab self-reports. We use the lab data to train this classifier because we know ground truth - which self-reports were given after stressors (when the user was stressed) and which were not (when the user was not stressed).

Two types of features were computed, momentary features and aggregate features. Momentary features are features associated with a single self-report from a single individual (e.g., mean, standard deviation, etc. of responses to questions associated with stress). Aggregate features are features computed across all self-reports provided by an individual (e.g., mean, standard deviation, etc. of momentary stress across self-reports from the individual). Thus, aggregate features provide a sense of how a particular individual answers self-reports. In addition to commonly used statistical features (mean, standard deviation, minimum, maximum, interquartile range, etc.), z-scores and histogram bin counts over momentary and aggregate features were computed. The histogram features provide characteristics of the distribution of the features. The z-scores provide the distance of a feature from its mean in terms of standard deviations (computed by subtracting the mean from the feature and dividing it by the standard deviation).

Feeding these features to WEKA, we found a J48 decision tree over these features correctly classified 84% (258/307) of self-reports from the lab using just one feature, the z-score of momentary self-reported stress. This feature is sufficient because it incorporates both global and local characteristics of self-reports in a single measure. If the z-score for the self-report is greater than .6, the self-report is classified as stressed. Intuitively, this means that if self-reported stress is greater than 60% of the deviation from the participant's global mean, the participant was stressed when completing the self-report.

7. APPLYING STRESS MODELS TO THE FIELD DATA

In this section, we apply the three stress models (physiological, perceived, and self-report) to data collected on two separate days in the natural environment. We first screen and clean the data to obtain valid minutes of measurements. Next, we evaluate the correlation between the stress rating obtained from the perceived stress model and that from self-reports. Finally, we use the three stress models to determine the percent of time that participants were found to be stressed in the field from each model.

7.1 Screening & Cleaning of Field Data

The data collected in the field are subject to several sources of noise, confounds, and losses. In addition to removal of outliers, several minutes of data had to be removed as described here. First, all minutes of data corresponding to the time when a self-report was completed are removed, as self-report prompts affect physiological signals even in the lab (see Section 4.1.1 and Figure 5). Second, all minutes of data that occur concurrently with significant motion (as detected by the accelerometer) are removed, as motion and physical activity overwhelm the physiological response to stress. Third, we remove two minutes following physical activity, since we find that physiology returns to baseline within two minutes after activity[3]. Figure 10 shows that a significant portion of the field data must be removed due to physical activity. Fourth, all minutes of ECG data that have less than 30 valid R-R intervals or less than 66% of RIP samples are removed since features can not be computed reliably for these minutes.

From a total of 422 hours of data collected in the field, 37% had to be removed due to confounding from physical activity. An additional 29.45% of data were removed due to poor quality or losses in the wireless transmission. The stress models are applied to the remaining 33.55% of data (i.e., 142 hours) of valid data. Finally, out of 21 subjects, 4 subjects are eliminated from the analysis because of missing sensor data (ECG or RIP), excessive noise, and missing self-reports, leaving 17 subjects for the field evaluation.

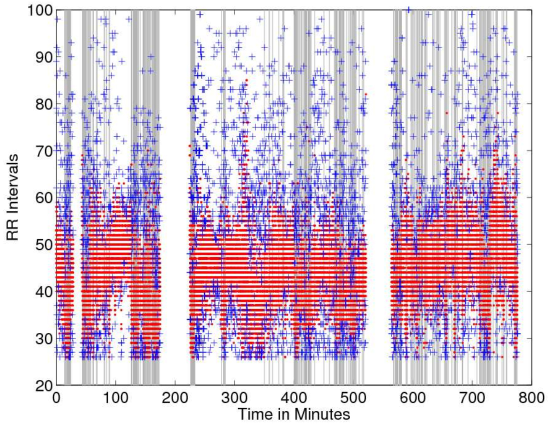

Figure 10: RR intervals (red asterisks) measured in the field from a specific participant. A significant portion of the day cannot be classified due to physical activity (vertical gray lines), outliers (blue crosses), or data loss from system malfunction (white areas).

7.2 Evaluation of Perceived Stress Model on Field Data

For evaluating the perceived stress model on lab data, we evaluated how the perceived stress model matched each instance of self-report for individual subjects (see Figure 9). For evaluating the perceived stress model on the field data we can not use the same approach. This is due to excessive loss in the data collected in the field; in fact, the average length of consecutive valid data is $<$ 4 minutes. Hence, the model does not get a sufficient number of valid minutes to match each individual self-report rating collected in the field. For the same reason, we can not personalize the perceived stress model using the self-report rating collected in the field. Therefore, we use α and β determined for each subject from their lab data. We expect the model to get better when it is calibrated from self-reports collected in the field since there may exist lab-to-field variability in providing self-report ratings.

[3] If subjects undergo intense physical activity, then it may take longer than 2 minutes for the physiology to return to baseline [15].

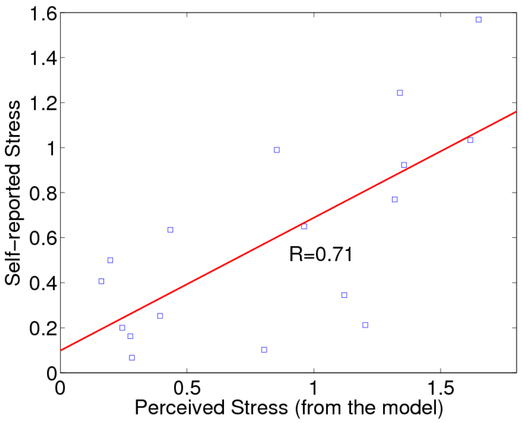

Figure 11: Agreement between self-report rating of stress and that predicted by the perceived stress model. The horizontal axis represents the value of the perceived stress, averaged over both days, and the vertical axis is the averaged self-report rating. Each of the 17 subjects have one data point in this graph.

For evaluation, we compare the average rating of stress provided by each subject over two days in the field and compare it to the average rating produced by the perceived stress model on the same subject. Figure 11 shows that we obtain a correlation of 0.71 between self-reported stress and the output of the perceived stress model.

7.3 Three Measures of Stress in the Field

The three models of stress presented in this paper — physiological (Section 4), perceived (Section 5), and self-report (Section 6) — provide three distinct measures of stress. We present a summary measure from each model, namely, percentage of time that the subjects were found to be stressed in the field across both days. We apply the self-report classifier on the z-score over each momentary measure of self-report to classify it into a stress state. To classify each minute of perceived stress rating, we first apply the correlation coefficient (from Figure 11) to scale it, then compute z-score for each minute, and then apply the self-report classifier to obtain a stress state for each minute.

The physiological classifier shows that the subjects were physiologically stressed 35.14% of the time, which excludes physiological activation due to motion or activity. The perceived stress model, which is a more robust and aggregate measure of stress, shows that subjects were stressed 26.61% of the day. Finally, self-report shows that subjects reported themselves to be stressed 28.08% of the time.

8. CONCLUSIONS AND FUTURE WORK

In this work, we proposed, developed, and evaluated the first continuous classifier of perceived stress that can be readily used in natural environments without pre-calibration. This innovation was made possible because of the development of a novel wearable sensing suite which we used to collect measurements from a rigorous lab stress protocol that has been repeatedly validated in behavioral science.

Through several innovations, our approach both improves the accuracy and simplifies the adoption of stress inferencing in natural environments. First, we correct for between-person differences using a self-calibrating normalization and a population-level model, removing the need for pre-calibration in a controlled setting. Second, we found that respiration features were highly discriminatory of physiological stress, allowing the use of a single, unobtrusive respiration band to capture stress. Third, we developed a new model that maps physiological stress to perceived stress. It is the first model to incorporate the prolonged psychological effect of stressors on the individual. The output of this model was correlated with stressors in the lab and had good concordance with self-report ratings collected in the field.

Although this work represents an important step forward, there is still significant work to be done to build a robust, highly accurate (99+%) classifier of perceived stress. First, improvements in wearable sensors are needed to limit the amount of data lost or corrupted. Second, new methods must be developed to control for the effect of physical activity on physiological signals. Otherwise, significant portions of daily life (37% in our field study) can not be classified. Third, this work only used two sensing modalities, ECG and respiration. The introduction of features from other modalities, e.g., skin response, pulse transit time, oxygen level in the blood, body temperature, etc. can further improve the accuracy of our models. Fourth, the models proposed in this work provide a binary classification of stress. However, people experience stress at multiple strengths. A more realistic model would incorporate multiple levels or a continuous measure of stress. Fifth, the perceived stress model proposed in this paper is only a first step in this direction. More powerful models can be investigated that can more accurately capture the accumulation, decay, and superposition of multiple overlapping stressors. Finally, the time or amount of data needed to self-calibrate both the physiological and perceived stress models in the field can be investigated.

9. ACKNOWLEDGMENTS

We thank Fei Sha for many invaluable conversations on machine learning and personzalized modeling of stress. We also thank Md Mahbubur Rahman for his help with computing respiration features.

10. REFERENCES

[1] Autosense: A wireless sensor system to quantify personal exposures to psychosocial stress and addictive substances in natural environments. http://sites.google.com/site/autosenseproject.

[2] M. al'Absi. *Stress and addiction: Biological and psychological mechanisms*. Academic Press/Elsevier, 2007.

[3] M. al'Absi and D. Arnett. Adrenocortical responses to psychological stress and risk for hypertension. *Biomedecine & Pharmacotherapy*, 54(5):234–244, 2000.

[4] M. al'Absi, S. Bongard, T. Buchanan, G. Pincomb, J. Licinio, and W. Lovallo. Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors. *Psychophysiology*, 34(3):266–275, 1997.

[5] M. al'Absi, T. Buchanan, and W. Lovallo. Pain perception and cardiovascular responses in men with positive parental history for hypertension. *Psychophysiology*, 33(6):655–661, 1996.

[6] M. al'Absi, D. Hatsukami, and G. Davis. Attenuated adrenocorticotropic responses to psychological stress are associated with early smoking relapse. *Psychopharmacology*, 181(1):107–117, 2005.

[7] M. al'Absi, K. Petersen, and L. Wittmers. Adrenocortical and hemodynamic predictors of pain perception in men and women. *Pain*, 96(1-2):197–204, 2002.

[8] G. Berntson, K. Quigley, J. Jang, and S. Boysen. An approach to artifact identification: Application to heart period data. *Psychophysiology*, 27(5):586–598, 1990.

[9] E. Broek, V. Lisỳ, J. Janssen, J. Westerink, M. Schut, and K. Tuinenbreijer. Affective Man-Machine Interface: Unveiling human emotions through biosignals. *Biomedical Engineering Systems and Technologies*, 52:21–47, 2010.

[10] J. Cacioppo and L. Tassinary. Inferring psychological significance from physiological signals. *American Psychologist*, 45(1):16–28, 1990.

[11] G. Chrousos and P. Gold. The concepts of stress and stress system disorders: overview of physical and behavioral homeostasis. *Jama*, 267(9):1244, 1992.

[12] M. Danninger, T. Kluge, and R. Stiefelhagen. MyConnector: analysis of context cues to predict human availability for communication. In *Proceedings of the 8th international conference on Multimodal interfaces*, page 19. ACM, 2006.

[13] M. Enoch. Pharmacogenomics of alcohol response and addiction. *American Journal of Pharmacogenomics*, 3(4):217–232, 2003.

[14] M. Enoch. Genetic and environmental influences on the development of alcoholism. *Ann NY Acad Sci*, 1094:193–201, 2007.

[15] M. Esco, M. Olson, H. Williford, D. Blessing, D. Shannon, and P. Grandjean. The relationship between resting heart rate variability and heart rate recovery. *Clinical Autonomic Research*, 20(1):33–38, 2010.

[16] Y. Freund and R. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[17] P. Grossman, J. Beek, and C. Wientjes. A comparison of three quantification methods for estimation of respiratory sinus arrhythmia. *Psychophysiology*, 27(6):702–714, 1990.

[18] M. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[20] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris. Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life. *Pervasive Computing*, pages 156–173, 2010.

[21] J. Healey and R. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2):156–166, 2005.

[22] J. Henry. Stress, neuroendocrine patterns, and emotional response. *Stressors and the adjustment disorders*, pages 477–496, 1990.

[23] W. James. The principles of psychology (Vols. 1 & 2). *New York: Holt*, 1890.

[24] A. Johnson and E. Anderson. Stress and arousal. *Principles of psychophysiology: Physical, social and inferential elements*, pages 216–252, 1990.

[25] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on pattern analysis and machine intelligence*, pages 2067–2083, 2008.

[26] S. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010.

[27] S. Kreibig, F. Wilhelm, W. Roth, and J. Gross. Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films. *Psychophysiology*, 44(5):787–806, 2007.

[28] J. L. Subjective Sensing: Intentional Awareness for Personalized Services. In *NSF Workshop on Future Directions in Networked Sensing Systems: Fundamentals and Applications*, 2009.

[29] B. McEwen. Protection and damage from acute and chronic stress. *Ann NY Acad Sci*, 1032:1–7, 2004.

[30] B. McEwen and E. Stellar. Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine*, 153(18):2093, 1993.

[31] M. Myrtek and G. Brugner. Perception of emotions in everyday life: studies with patients and normals. *Biological psychology*, 42(1-2):147–164, 1996.

[32] J. Pan and W. Tompkins. A real-time QRS detection algorithm. *Biomedical Engineering, IEEE Transactions on*, BME-32(3):230–236, 2007.

[33] J. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[34] M. Rahman, A. Ali, A. Raij, E. Ertin, M. al'Absi, and S. Kumar. Online Detection of Speaking from Respiratory Measurement Collected in the Natural Environment. In *Proceedings of the 10th International Conference on Information Processing in Sensor Networks (IPSN) (to apprear)*. ACM, 2011.

[35] A. Raij, A. Ghosh, S. Kumar, and M. Srivastava. Privacy Risks Emerging from the Adoption of Inoccuous Wearable Sensors in the Mobile Environment. In *Proceedings of the 29th ACM Conference in Human Factors in Computing Systems (CHI) (to apprear)*. ACM, 2011.

[36] P. Rainville, A. Bechara, N. Naqvi, and A. Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International journal of psychophysiology*, 61(1):5–18, 2006.

[37] R. Rosmond and P. Bjorntorp. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. *Metabolism*, 47(10):1187–1193, 1998.

[38] R. Rosmond, M. Dallman, and P. Bjorntorp. Stress-related cortisol secretion in men: relationships with abdominal obesity and endocrine, metabolic and hemodynamic abnormalities. *Journal of Clinical Endocrinology & Metabolism*, 83(6):1853, 1998.

[39] B. Schölkopf and A. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002.

[40] Y. Shi, M. Nguyen, P. Blitz, B. French, S. Fisk, F. De la Torre, A. Smailagic, D. Siewiorek, M. al'Absi, E. Ertin, T. Kamarck, and S. Kumar. Personalized stress detection from physiological measurements. *International Symposium on Quality of Life Technology*, 2010.

[41] C. Stephens, I. Christie, and B. Friedman. Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. *Biological psychology*, 84(3):463–473, 2010.

[42] A. Steptoe, G. Fieldman, O. Evans, and L. Perry. Cardiovascular risk and responsivity to mental stress: the influence of age, gender and risk factors. *European Journal of Cardiovascular Prevention & Rehabilitation*, 3(1):83, 1996.

[43] B. Todd and D. Andrews. The Identification of Peaks in Physiological Signals. *Computers and biomedical research*, 32(4):322–335, 1999.

[44] H. Ursin and R. Murison. Classification and description of stress. *Neuroendocrinology and psychiatric disorder*, pages 123–132, 1984.

[45] A. Wilson, C. Franks, and I. Freeston. Algorithms for the detection of breaths from respiratory waveform recordings of infants. *Medical and Biological Engineering and Computing*, 20(3):286–292, 1982.