Personal and Ubiquitous Computing (2018) 22:671–688
https://doi.org/10.1007/s00779-017-1108-z

ORIGINAL ARTICLE

Unobtrusive stress detection on the basis of smartphone usage data

Elena Vildjiounaite[1] · Johanna Kallio[1] · Vesa Kyllönen[1] · Mikko Nieminen[1] · Ilmari Määttänen[2] · Mikko Lindholm[1] · Jani Mäntyjärvi[1] · Georgy Gimel'farb[3]

Received: 27 June 2017 / Accepted: 13 December 2017 / Published online: 17 January 2018

Abstract

Stress has become an important health problem, but existing stress detectors are inconvenient in long-term real-life use because users either have to wear dedicated devices or expend notable interaction efforts in system adaptation to specifics of each person. Adaptation is necessary because individuals significantly differ in their perception of stress and stress responses, but typical adaptation employs supervised learning methods and hence requires fairly large sets of labelled data (i.e. information on whether each reporting period was stressful or not) from every user. To address these problems, we propose a novel unsupervised stress detector, based on using a smartphone as the only device and using discrete hidden Markov models (HMM) with maximum posterior marginal (MPM) decisions for analysis of phone data. Our detector requires neither additional hardware nor data labelling and hence is truly unobtrusive and suitable for lifelong use. Its accuracy was evaluated using two real-life datasets: in the first case, adaptation was based on very short (a few days) phone interaction histories of each individual, and in the second case—on longer histories. In these tests, the proposed HMM-MPM achieved 59 and 70% accuracies, respectively, which is comparable with results of fully supervised methods, reported by other works.

Keywords Mobile phone data analysis · Stress detection · Personalisation · Unsupervised learning · Hidden Markov models

1 Introduction

Stress is defined as an imbalance between external forces or loads and individual possibilities to cope with or resist those externals forces [18].Work-related stress is becoming a public health crisis [2], increasing sickness absences and presenteeism, and the total annual cost of work-related depression in Europe has been estimated to be 617 billion euros [20]. This problem calls for new solutions for automatic stress recognition from daily life data because methods, developed in laboratories, cannot be straightforwardly transferred to real-life settings.

In laboratories, stress detection is usually based on physiological data, collected during fairly short time periods (typically, not more than two hours). Electrocardiograph (ECG), skin conductance and respiration sensors are the most frequently used sensors, and the most popular stress detection feature is Heart Rate Variability (HRV), calculated from ECG data [1]. Unfortunately, physiological devices are sensitive to improper attachment. For example, although many lab studies successfully used chest belts with embedded ECG sensors, in a field study of Plarre et al. [24] 30% of data got lost or corrupted due to body jerks and loosening of chest belts with time. Rahman et al. [26] attempted to overcome this problem by monitoring data quality and asking the study participants to fix sensor attachment in cases of troubles, but nevertheless, 2–2.5 hours of ECG data were lost on average every day. Wrist bracelets may provide either heart rate sensors or GSR (galvanic skin response) sensors or both. Unfortunately, GSR sensor signal is also notably affected by physical activities and attachment problems [4]. Probably due to this reason Sano and Picard [28] did not observe a significant correlation between the reported high stress scores of the test subjects and GSR data from a wrist-worn sensor.

Although wrist bracelets were found to be more practical than chest belts [2, 12], they are nevertheless tiresome. According to [31], one third of wearable fitness tracker owners stop wearing them after about 6 months. As we are unaware of long-term studies with wearable physiological stress detectors,

✉ Elena Vildjiounaite
elena.vildjiounaite@vtt.fi

[1] VTT Technical Research Centre of Finland, Kaitoväylä 1, 90571 Oulu, Finland

[2] Medicum, 00014 University of Helsinki, Helsinki, Finland

[3] Department of Computer Science, University of Auckland, Auckland 1149, New Zealand

we do not know their abandonment rate, but a recent stress recognition review observed that "discomfort provoked by most of the physiological monitoring devices is still to overcome if they are going to be used in real-life activities" [1]. Hence, owners of such devices may also abandon them rather quickly or forget to charge, especially during busy times.

In addition, physiological parameters are affected by various daily life activities, e.g. eating, drinking, caffeine intake, conversation and motion [24]. To account for influence of physical activity on physiological parameters, Gjoreski et al. [12] combined activity recognition on the basis of accelerometer data with analysis of data from multiple physiological sensors (GSR, skin temperature, heart rate and blood volume pulse), but observed that success of this method strongly depends on the age, gender and physical fitness of the user.

Due to these difficulties with physiological sensors, many studies into the development of stress detection methods for long-term real-life use employed smartphones as the only data source [5, 10, 11, 13, 21, 31]. One work studied the feasibility of collecting physiological data in night time in addition to daytime mobile phone data, but wearing physiological sensors at night also appeared inconvenient: 12 subjects did not do it for more than one night [23]. Using mobile phones as the only data source seems to be easier for the users, but unfortunately, it does not ensure convenience of the whole system because existing data processing algorithms require each user to invest notable explicit interaction efforts into the collection of training data. Usually, stress labels (i.e. self-reports of test subjects regarding their stresses during the reporting period) are obtained upon periodic system prompts, and many studies observed that the test subjects find self-reporting tiring. For example, in [2] the ratio of provided self-reports to the number of system prompts ranged from 9 to 46% for different persons.

The reason for requiring each user to provide fairly large sets of labelled data is twofold: first, the majority of state-of-the-art stress detection algorithms are trained in a fully supervised way. Second, neither physiological nor behavioural changes due to stress are reflected in the same way in all individuals [1] and hence algorithms, trained using data of a target person only, usually achieve notably higher accuracies than algorithms trained using data of other persons. For example, in field studies into stress recognition on the basis of smartphones, general models (i.e. models, trained using data of many non-target individuals) achieved average accuracies 45% in [23] and 54% in [13], whereas accuracies of person-specific models (i.e. models, trained on data of each target individual separately) were higher by 10–20%. These accuracy gains, however, required each target individual to provide a fairly large number of stress labels: training dataset in [13] contained nearly 100 labelled instances per person; dataset in [23] was even larger. In real life, not every user would provide so large a number of self-reports to train fully supervised algorithms. Furthermore, training data cannot be collected "once and forever" as human behaviour and physiological condition evolve with time, and thus, stress detection models should be periodically updated to adapt to these changes.

Nevertheless not so many works studied ways to decrease the need for labelled training data. The majority of these works reduced the need for labelled data of the target individual by reusing labelled data of other persons; hence, they required obtaining large sets of labelled data of many subjects. Works, exploiting mobile phone data for stress detection, required obtaining also labelled data of a target individual in not-so-small quantities, i.e. each target individual had to provide about 60 self-reports [10, 21]. In addition, it was observed that success of reusing data of other individuals strongly depends on the degree of similarity between these individuals and the target person [21, 36].

To the best of our knowledge, only one work proposed stress detection methods, requiring no data labelling from the end users. Kusserow et al. [16] detected arousal as the temporal deviation of current HRV values from their average values. This system employs a chest belt with physiological and accelerometer sensors, which requires careful attachment and tightening, plus an additional accelerometer on a thigh to differentiate between physical activity-induced and arousal-induced deviations, and this setup is probably too complicated for a long-term real-life use.

Unlike the above works, our aim is truly unobtrusive stress detection both sensor-wise and algorithm-wise. Towards this goal, we use a mobile phone as the only sensing device and propose a novel unsupervised algorithm to learn person-specific stress recognition models on the basis of smartphone data. The feasibility of the proposed approach is confirmed by experiments with real-life data of 30 subjects, some of which used phones for work, and some—mainly for private purposes. To the best of our knowledge, this is the first study into unsupervised stress recognition using real-life mobile phone usage data.

2 Related work

To date, the majority of stress detectors are trained in a fully supervised way both in lab tests and field studies despite that obtaining stress labels in real life is more difficult than in a lab. Even studies, employing minimally invasive devices—mobile phones alone—employed supervised algorithms, most popular being SVM (Support Vector Machine) [11, 13], Naïve Bayes [10], and decision trees [5, 10, 11, 21, 31]. Studies using physiological sensors may overcome this problem by training stress recognition models on labelled lab data and then using these models in a field, but this approach requires test subjects to visit a lab and to spend there at least one hour [15, 24]. Furthermore, obtaining realistic mobile phone usage models in a lab would be more difficult than obtaining physiological models, if possible at all. Hence the majority of works, collecting field data, obtain labels via self-reports of the monitored subjects.

To date, the most common way to reduce the need for labelled data of each person is to exploit similarity between human beings. This can be done by clustering similar persons and training a separate model for each cluster. Then, a certain quantity of labelled or unlabelled data of each target person is used to determine the most appropriate cluster, and the corresponding cluster model is selected to detect stresses of the target person. Hernandez et al. [14] studied audio-based stress recognition in call centre and used unlabelled data to assign nine subjects to two clusters with K-means algorithm. Then, SVM model was trained using labelled data of subjects, similar to the target subject. This approach allowed to increase accuracy by 12% compared with the general model. Xu et al. [36] studied stress detection with neural networks using physiological data, collected in a lab; clustering of the test subjects was also based on their unlabelled data. Assigning of 44 subjects to two clusters allowed to increase accuracy (i.e. to reduce by 5% the mean difference between the predicted and actual stress indices), but the use of larger number of clusters decreased the accuracy compared with the general model.

Garcia-Ceja et al. [10] studied stress detection using acceleration data, collected by smartphones, and employed Naïve Bayes and Decision Tree classifiers. Clustering of test subjects also required labelled data, and in fairly large quantities, as self-reports for clustering were collected on workdays three times per day during approximately four weeks. The number of clusters was selected based on the quality of resulting clusters, and this approach allowed to increase recognition accuracy by 5–8% compared with general models.

Maxhuni et al. [21] employed decision trees for stress detection on the basis of mobile phone usage data and compared several ways to obtain a model of a target person using scarce sets of his/her labelled data: (1) to train a model using only target person data; (2) to train a model using a mixture of the target person data and data of similar individuals; (3) to combine outputs of models of similar individuals. But again, “scarce sets” of labelled data in [21] contained self-reports, collected on workdays three times per day during four weeks. Nevertheless this not-so-negligible labelling effort appeared insufficient for a fully supervised training using a target person data only: on average, person-specific models achieved 62% accuracy in distinguishing between three stress levels (low, medium, high). Adding labelled data of other individuals allowed to increase accuracy by 1% in one case: when data of just one nearest neighbour was used in training, and these data were sampled according to its similarity to the target person. The use of all data of the nearest neighbour, as well as the use of greater number of nearest neighbours, resulted in decreased accuracy. Combining outputs of models of similar individuals allowed to increase an average accuracy by 10% compared with training of person-specific models on scarce datasets, but this result was obtained when outputs of models of the three most similar persons were combined via weighted sum, weights being dependent on distances from the target person. Using other numbers of similar persons resulted in lower accuracy, and voting-based combination of outputs of similar models resulted in accuracy as low as 50%.

Another fairly well-known way to reduce the need in the labelled data of a target person is to train a general model using all available data of other persons and then adapt a general model to a target person using a relatively small number of his/ her data samples [33]. To date, suitability of this approach to the stress detection task was studied only for audio and physiological data and only for the case of supervised training of general models. Hernandez et al. [14] proposed to incorporate into SVM class priors, reflecting individual tendency to report more or less stressful events. Shi et al. [29] instead trained SVM on data, normalised in a person-specific way. This normalisation was performed by subtracting from raw physiological data of each individual his/her average values in the neutral state.

To summarise, none of the proposed to date approaches to reduce the need in labelled data of a target person eliminated the need to obtain large sets of labelled data of many other persons, and none of the studies into stress detection on the basis of phone data eliminated the need to obtain labelled data of a target person in not-so-small quantities.

Unlike the above-listed works, Kusserow et al. [16] proposed an unsupervised method to detect arousal on the basis of physiological and acceleration data, but this method required a fairly complicated sensor setup: a chest belt with physiological and accelerometer sensors, which should be carefully attached and tightened, plus an additional accelerometer on a thigh. Arousal detection is based on calculation of so-called “additional heart rate” (AHR): current heart rate (HR) indicates arousal onset if the difference between the current HR and HR average over the last three minutes exceeds a certain threshold. Threshold is calculated dynamically in order to take into account possible influence of physical activities on heart rate increase. Therefore, threshold depends on physical activity score, calculated from accelerometer data. Accelerometer data were also used to train a Naïve Bayes classifier to recognise several activity primitives, such as standing, sitting, walking, and bending. Then, corresponding heart rates were clustered to identify which physical activities influence AHR least of all, and heart rate deviations, accompanied by physical activities from other clusters, were discarded.

Our algorithm also estimates deviations between current and usual mobile phone usage patterns, but building a model of "usual human behaviour" is a challenging task. Due to large variations of human activities, attempts to recognise short-term unusual behaviours are likely to fail: even if someone's phone usage behaviour is unusual during several hours in a row, it does not necessarily mean that he/ she is stressed; he/she may simply be doing other things. Hence, we detect stresses on a daily basis, and the most related to our work is the study into unsupervised detection of illnesses of the elderly on the basis of depth camera data: Vildjiounaite et al. [34] proposed to employ HMM (hidden Markov model) classifier to learn mappings between features, extracted from depth camera data, and activity levels of the monitored subjects. At the inference stage, the trained HMM recovered a sequence of hidden states according to the learned mappings, and a sequence of low activity levels was classified as an illness. In this work, we also employ HMM to learn mappings between different mobile phone usage features and hidden states, but we had to define phone usage-specific data features and HMM configuration.

3 Stress detection algorithm

Due to large varieties of human behaviours, it is very difficult to build a model encompassing every possible normal behaviour [7]. As this is the first study into unsupervised stress detection on the basis of mobile phone usage data, we build a model of normal behaviour for the whole day and distinguish between stressful and normal days. We do not attempt at differentiating between stress levels on a finer scale in this work because this is a challenging task even for fully supervised methods. For example, in [23] person-specific models, classifying each day into three stress levels on the basis of mobile phone usage data, achieved an average accuracy 55%, whereas in [21] person-specific models, trained using self-reports acquired three times per day during four weeks, classified each half-day into three stress levels with an average accuracy 62%.

Similar to [34], we model each day as a sequence of time windows. We employ overlapping time windows because they allow to detect unusual periods more accurately than consecutive windows [8].

3.1 Mobile phone usage data

Previous studies did not provide abandon information regarding interrelations between various phone features and stresses. Sano and Picard [28] reported that high stress correlated with a smaller mean of "screen on" times, especially in the evenings, and smaller standard deviation of "screen on" times in the evenings. Percentage and length of sent SMS were also smaller during high stress periods. Correlations between high stress and standard deviation of "screen on" times during daytime, as well as correlations between high stress and phone call features, appeared to be insignificant. Correlations with other stress levels were not reported. Bogomolov et al. [5], on the contrary, found that physical interaction (inferred by detecting the proximity of other phones via Bluetooth communications) played the most significant role, whereas predictive power of the SMS data needs further investigation. Call futures, such as the number of incoming and outgoing calls, were found to be more useful than SMS features, but no other details were provided [5]. Other studies [11, 13, 21] did not report any correlations. On the other hand, previous studies reported that extensive data collection quickly drains phone battery. For example, Muaremi et al. [23] collected microphone, acceleration, and GPS data in addition to fairly detailed application usage data, and in some cases, they had to disable data collection before 17.00 because of low battery, despite the battery being fully charged during the night.

As previous studies did not ascertain the importance of collecting detailed call and SMS data, we decided against doing it, moreover that collection of detailed call data can be perceived as privacy-threatening. Instead, we collected cumulative usage data for several application categories: communication, infotainment, entertainment, well-being, etc. Usage data were collected as follows: each time when an application of a certain category was started or moved to the foreground, this time served as a starting timestamp of using this category unless this new application replaced another application of the same category. End timestamps were obtained in the same way.

These data were then pre-processed: for each minute, we calculated whether a certain category was active and, if so, for how many seconds. Therefore, each day was represented by a matrix of 1440 rows (minutes) where each row contained usage values of the chosen application categories for this minute. The normalised category-wise application usage data served as input to the HMM classifier, presented below (we employed the conventional min-max normalisation). In addition, we collected location data because we expected that normal phone usage behaviour may differ in different places, for example, in home vs. at work.

3.2 Unsupervised hidden Markov model training

We model a day as a sequence of time windows and employ a discrete HMM to classify each window in this sequence into a pre-defined number of classes. Every hidden state in the HMM represents one of the output classes, and features of the mobile phone usage data, calculated within each time window, serve as observations.

Stress detection algorithm is trained as follows. First, a set of reference models of normal (typical) phone usage behaviour of each person is created in an unsupervised way. Reference models are built in context-dependent ways, i.e.

we create a separate reference model for each time window, thus providing for time-dependency of human activities. For example, if an individual is in a hurry to get to work in the mornings, he/she may use a mobile phone less actively than during daytime or in the evenings. Hence, it is infeasible to mix morning and evening data: both usual morning patterns and usual evening patterns may appear unusual in comparison with their combination. Furthermore, we create several location-specific reference models for the same time window if during this window the user stays in some place long enough, i.e. if training data contains sufficiently large number of samples, collected (not necessarily at once) within this time window in this place. In this work, we have chosen a fairly small time threshold for "long enough" data collection in one place because data were not abundant (see Section 4.2).

We do not assign semantic names to such "typical places"; instead, we find clusters of GPS coordinates in the training data and call these clusters "place 1", "place 2", etc. This way, we provide for location-dependency of human behaviour in addition to time-dependency. For example, reference models of someone's "window 2 & place 2" and "window 18 & place 2" may be both models of phone usage in a swimming pool (if this individual sometimes swims in the mornings and sometimes in the evenings) and thus may contain only zero values, suggesting that not using a phone at all in this context is quite a normal behaviour. A reference model of the same user for "window 18 & place 20" may reflect his/her phone usage patterns during dinners in a favourite restaurant, where he/she may use a phone more actively. In addition to location-specific reference models, for each time window we create a reference model from data samples, collected outside of typical places, i.e. not belonging to any location cluster.

Ideally, a reference model of normal behaviour should be built using only samples of non-stressed behaviour. In unsupervised learning this is not possible; hence, for creating each reference model, we use in exactly the same way all training samples, obtained within a certain time window either in a typical place or "elsewhere" (outside of typical places). A reference value $V_{\mathrm{RTP}i}$ of phone application category i for time window T and place P is calculated according to formula (1), where P can be either identifier of a typical place or "elsewhere" position; $V_{t:i,\ P}$ is normalised usage time of application category i, obtained at time moment $t \in \mathrm{T}$ in place P; and m is total number of samples, obtained in place P and time window T in all days in the training dataset.

$$V_{\mathrm{RTP}i} = \frac{1}{m}\sum_{t=1}^{m} V_{\mathrm{t:i,P}} \tag{1}$$

Then, for each data sample in time window T, we calculate the context-dependent deviation of this sample from the corresponding reference model. A deviation D_t of the sample t with n features from the reference behaviour is calculated in the following way: first, we check whether this sample belongs to any location cluster or not, then select a reference model of the corresponding time window T and place P and calculate the deviation according to formula (2):

$$D_t = \frac{1}{n}\sum_{i=1}^{n}\left(V_{\mathrm{t:i,P}} - V_{\mathrm{RTP}i}\right) \tag{2}$$

Then, we use deviations D_t, calculated for all samples of time window T, to calculate the following window features F_{T}:

- mean values of deviations D_t within the time window T, calculated according to formula (3), where K is number of minutes in a window T

$$F_{\mathrm{T}} = Mean_{\mathrm{T}} = \frac{1}{K}\sum_{t=0}^{K} D_t \tag{3}$$

- standard deviations of D_t within the time window T, calculated according to formula (4), where K is number of minutes in a window T, and $Mean_{\mathrm{T}}$ is calculated according to formula (3)

$$F_{\mathrm{T}} = SD_{\mathrm{T}} = \sqrt{\frac{1}{K}\sum_{t=0}^{K}\left(D_t - Mean_{\mathrm{T}}\right)^2} \tag{4}$$

Then, we discretise each value F_{T} by dividing an interval [− 1, 1] into several sub-intervals and using the sub-interval's number as an observation in the HMM. The experiments below employed the following sub-intervals: [− 1.0, − 0.6]; [− 0.6, − 0.3]; [− 0.3, 0]; [0, 0.3]; [0.3, 0.6]; [0.6, 1.0]. A sequence of the discretised deviations of day window features from the reference model is a sequence of the HMM observations for this day.

The HMM of a day has finite sets of hidden states S = {1,…, N} (output classes) and observations X = {1,…, J}. The proposed HMM is illustrated in Fig. 1. For example, Fig. 1 shows that cumulative usage time of applications of category 1 was equal to five seconds during the first minute of window 1 and equal to zero during the second and third minutes; cumulative usage time of applications of category 2 was equal to nine seconds during the third minute of window 1; two first samples of window 1 were obtained in place P1, while the third sample was obtained in place P2, etc.

Let $S_D = \{s_T : T = 1, \ldots, D\}$ be a sequence of the hidden states and $X_D = \{x_T : T = 1, \ldots, D\}$ be a corresponding sequence of obtained observations for D time windows of a day. The HMM assumes that every observation x_T at time T

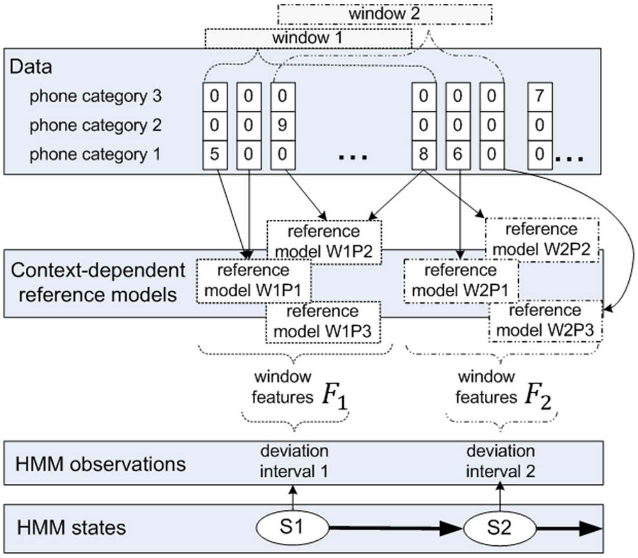

Fig. 1 HMM model and input features

depends, in a probabilistic sense, only on a hidden state s_T and the latter (excluding s_1) depends in turn only on the previous state s_{T-1}. Let $\alpha = [\alpha(s|v) : s, v \in S]$; $\beta = [\beta(x|s) : x \in X, s \in S]$, and $\pi = [\pi(s) : s \in S]$ denote conditional probabilities of transitions between the discrete states; conditional probabilities of observations, given the state, and unconditional probabilities of the initial state, respectively. Then, the HMM is characterised by the joint probability of the sequences of evolving states and observations [27]:

$$\mathrm{p}(S_D, X_D) = \pi(s_1)\beta(s_1|x_1)\prod_{T=2}^{D}\alpha(s_T|s_{T-1})\beta(x_T|s_T) \tag{5}$$

Given a sequence of unlabelled observations X_T, HMM is trained, i.e. its parameters (α, β, π) are learned in a fully unsupervised mode by applying the conventional Baum-Welch forward-backward algorithm [27]. In other words, we do not use any labels in training, and mapping of deviations from usual behaviour into output classes does not require defining any thresholds: it is learned from training data.

3.3 Inference with hidden Markov models

The trained HMM models process each test day as follows: first, we obtain a sequence of observations, i.e. a sequence of discretised window features F_T. Then a sequence of the hidden states can be obtained by the Bayesian maximum a posteriori (MAP) decision rule using the well-known Viterbi dynamic programming algorithm [27]. The MAP rule minimising the error probability assumes that the cost of errors for a given sequence of observations is just the same, irrespectively of their number (i.e. a single erroneous state is as bad as any other number of errors). An alternative way is to account for all the individual errors and minimise their expected number. In this case, the hidden states are to be recovered with the Bayesian MPM (maximum posterior marginal) rule that selects for each time moment the hidden state with the maximum posterior marginal probability. In this work, we employ HMM with MPM decision rule because it outperformed conventional HMM with MAP in two different studies comparing these two approaches: (1) detection of emotions of the show audience [32] and (2) detection of illnesses of the elderly [34].

The posterior marginals $\{p_T(s|X_D) : s \in \mathbf{S}; T = 1, \ldots, D\}$ are calculated for each hidden state s of time window T by the forward and backward message propagation [27]:

$$p_T(s|X_D) = \frac{\mu_T(s|x_1,\ldots,x_T)m_T(s|x_D,\ldots,x_T)}{\sum_{v\in\mathbf{S}}\mu_T(v|x_1,\ldots,x_T)m_T(v|x_D,\ldots,x_T)} \tag{6}$$

where $\mu_T(s|x_1, \ldots, x_T)$ and $m_T(s|x_D, \ldots, x_T)$ denote the forward and backward message, respectively, for the state s at each window T. These messages are computed successively from the beginning and the end of the observed sequence X_D: $\mu_1(s; x_1) = \pi(s)\beta(x_1|s)$;

$$\mu_T(s; x_1, \ldots x_T) = \sum_{v=1}^{D}\mu_{T-1}(v; x_1, \ldots x_{T-1})\alpha(s|v)\beta(x_T|s) \tag{7}$$

and $m_D(s; x_D) = 1$; $m_T(s|x_D, \ldots, x_T) = \sum_{v=1}^{D} m_{T+1}(v|x_D, \ldots, x_{T+1})\,\alpha(v|s)\beta(x_{T+1.})$

A "stress score" of the day can be calculated as a conventional likelihood of generating a given sequence of observations by the learned HMM (the lower the likelihood, the less normal the sequence), but this estimation takes into account the order of recovered hidden states. Hence, if in the training data for example a sequence S1-S1-S2-S2 occurs more frequently than a sequence S1-S2-S1-S2, the latter will be estimated as "less normal" than the former and the HMM will therefore lack robustness to small deviations from usual time routines, such as a shift of usual activities by an hour or two. As many human beings do not have strict schedules, this approach will result in a too inflexible model. Therefore, we calculate a stress score of each day by assigning numerical scores A_T to the recovered hidden states of time windows T. For HMM with three hidden states, we assign numerical scores according to formula (8):

$$A_T = \begin{cases} 1, & \mathit{if\ time\ window\ T\ is\ classified\ as}\ S_1 \\ 0, & \mathit{if\ time\ window\ T\ is\ classified\ as}\ S_2 \\ -1, & \mathit{if\ time\ window\ T\ is\ classified\ as}\ S_3 \end{cases} \tag{8}$$

Then the day score A is calculated according to formula (9):

$$A = \frac{1}{D}\sum_{T=1}^{D} A_T \tag{9}$$

For example, the stress score of a sequence "S_2, S_2, S_3, S_1" is

$$A = \frac{0+0-1+1}{4} = 0.$$

Calculation of day scores according to formulas (8) and (9) relies on the observations of Sano and Picard [28] that during stressful days mean of phone usage times may decrease (when features F_T are calculated according to formula (3)) or SD of phone usage times may decrease (when features F_T are calculated according to formula (4)). The decrease in standard deviation may be caused by a decrease in phone usage times (e.g. if a person does not use a phone at all during two hours, the standard deviation of usage times per minute, calculated over these two hours, will be equal to zero). Change in standard deviation may also indicate more or less even distribution of usage times per minute and hence more or less hectic behaviour. Therefore, we calculated another measure of irregularity of phone usage times by assigning numerical scores $A_{T+1,\,T}$ to transitions between recovered hidden state T + 1 and the previous state T as follows:

$$A_{T+1,T} = abs(k-j) \;\; if \;\; window\ \mathrm{T} + 1 \; is \; classified \; \mathrm{as} \;\; S_k \;\; and \;\; window\ \mathrm{T} \; is \; classified \; \mathrm{as} \;\; S_j \tag{10}$$

Then, a day score was calculated according to formula (11):

$$A = \frac{1}{D-1} \sum_{T=1}^{D-1} A_{T+1,T} \tag{11}$$

In this case, the stress score of a sequence "S_2, S_2, S_3, S_1" is

$$A = \frac{(2-2)+(3-2)+(3-1)}{3} = 1.$$

Classification of a day as "stressful" vs. "normal" requires then a threshold TA—a borderline between scores of these two classes. We experimented with two types of stress thresholds:

$$TA_1 = Mean_A + \frac{SD_A}{w} \tag{12}$$

$$TA_2 = Mean_A - \frac{SD_A}{w} \tag{13}$$

where $Mean_A$ and SD_A are mean and standard deviation of days' scores A, respectively, and w is a parameter, specifying to which degree a day score should be an outlier to denote stress.

Use of stress threshold TA_1 (formula (12)) implies that we expect scores of stressful days to be higher than scores of normal days: a day is classified as "stressful" if its score A exceeds the threshold TA_1; otherwise it is classified as "normal". Use of stress threshold TA_2 (formula (13)) implies that we expect scores of stressful days to be lower than scores of normal days: a day is classified as "stressful" if its score A falls below threshold TA_2; otherwise it is classified as "normal."

4 Experiments

4.1 Data collection

4.1.1 First dataset

Data were collected by the Institute of Behavioural Sciences at the University of Helsinki (Finland) and VTT Technical Research Centre of Finland in their research project, aiming at studying how mobile phones can be used for stress detection and promoting a healthy lifestyle. The Ethical Review Board in Humanities and Social and Behavioural Sciences of the University of Helsinki has reviewed and approved this study. Fifty-six participants (9 males and 47 females) were recruited on the basis of their answers to the web questionnaire; these were the subjects who volunteered to participate in the data collection and who met the following selection criteria: good health, non-smoking, interest in technology, willingness to use mobile applications and possession of an Android smart phone. Exclusion criteria included acute mental health conditions, serious diseases, absence of an Android smart phone, and attendance to other stress/well-being studies. The participants signed an approval letter prior to the study. The majority of the subjects were monitored for four days, but a few subjects were monitored for five days, and a few subjects were monitored for three days.

Mobile phone data included two parts: (1) an Android smartphone application, collecting phone usage data as described above and (2) self-reporting. In this data collection, six application categories were chosen: social (Skype, social networks, etc.), entertainment (games, music, etc.), infotainment (news, books, etc.), business (calendar, editing, etc.), well-being (weight watching, exercise monitoring, etc.), and any other interaction with a phone. Data logs only contained information whether a user interacted with an application of certain type or not during each minute; contents of the web pages or keystrokes were not logged.

Self-reporting was prompted by an Android notification every 45 min during daytime from 9 am to 9 pm. It was up to the user when or if at all she/he provided a self-report. Regarding stress, the users had to answer whether stress had occurred during the current reporting period and, if yes, evaluate on a 7-level Likert scale the following statements: (1) "The situation was very stressful" (1 = This is not at all true, 7 = This is completely true) and (2) "I could control the situation well" (1 = This is not at all true, 7 = This is completely true). In the experiments, presented below, we considered a day as "stressful" if either medium or high stress (4–7 on the Likert scale) was reported during this day; otherwise this day was considered normal. All subjects, who reported that stress had occurred, also reported that they coped with it quite well (their answers ranged from three to seven). Unfortunately, only half of the subjects provided enough self-reports to hope that their stresses were not missing. Thus, this data collection once again confirmed that practical stress detection systems should not rely on the availability of labels.

The resulting dataset contained nearly 100 days of phone usage logs of 28 subjects (4 males and 24 females); 39% of days were labelled as "stressful". These subjects aged from 20 to 47 years old (mean 25.5 years, standard deviation 6), and their occupations varied a lot: among them were sellers, teachers, a cleaner, a researcher, a fitness trainer, a cafeteria worker, a secretary, and a sign language interpreter.

4.1.2 Second dataset

Data were collected by VTT Technical Research Centre of Finland after modifying smartphone application, used in the first data collection. First, we split former "social" category into two categories: "communication" (e.g. calls and SMS) and "social" (use of social networks) because we wanted to improve distinguishing between private and business-related activities: for many occupations, business-related calls dominate over private calls, whereas accessing social networks for professional reasons via mobile phones is not so common. We also added "shopping", "travel" and "utility" (changing phone settings, updates, etc.) categories. Thus, here, we used ten application categories instead of six; various phone applications were assigned to the chosen categories by querying Google Play Store. Again, data logs only contained durations of interacting with different application types.

Second, we only asked test subjects to label the whole day as "stressful" vs. "normal" because we wanted to collect longer phone usage histories without annoying the subjects. This way we collected data of two persons, not participating in the first data collection: one female and one male, aged from 35 to 43 years old (both senior researchers). One of them used phone for business reasons very often, another one mainly used phone for private reasons. This dataset contained nearly 200 days, 35% of days were labelled as "stressful".

4.2 Experimental protocol

In the experiments, presented below, we used the following variants of the proposed HMM-MPM stress detectors:

- ***HMM-Mean***: window features F_{T} are calculated according to formula (3), and day score A_{Mean} is calculated according to formulas (8) and (9)
- ***HMM-SD***: window features F_{T} are calculated according to formula (4), and day score A_{SD} is calculated according to formulas (8) and (9)
- ***HMM-Mean-scatter***: window features F_{T} are calculated according to formula (3), and day score $A_{Mean-scatter}$ is calculated according to formulas (10) and (11)
- ***HMM-SD-scatter***: window features F_{T} are calculated according to formula (4), and day score $A_{SD-scatter}$ is calculated according to formulas (10) and (11)
- ***Fusion of selected variants with the "sum" rule***, such as

  HMM-SD + ***HMM-SD-scatter***: day score $A_{SD\,\&\,SD-scatter} = A_{SD} + A_{SD-scatter}$

  HMM-SD + ***HMM-Mean-scatter***: day score $A_{SD\,\&\,Mean-scatter} = A_{SD} + A_{Mean-scatter}$

  HMM-SD + ***HMM-SD-scatter*** + ***HMM-Mean-scatter***: day score $A_{SD\,\&\,SD-scatter\,\&\,Mean-scatter} = A_{SD} + A_{SD-scatter} + A_{Mean-scatter}$

  HMM-SD + ***HMM-Mean*** + ***HMM-Mean-scatter***: day score $A_{SD\,\&\,Mean\,\&\,Mean-scatter} = A_{SD} + A_{mean} + A_{Mean-scatter}$

For combining individual stress detectors we employed "sum" fusion rule because it does not require defining any additional parameters. For combinations of classifiers with "sum" fusion rule decision thresholds were also calculated using "sum" rule, for example:

$TA_{1\ for\ \sum_{i=1}^{N} HMM_i} = \sum_{i=1}^{N} Mean_{A_i} + \frac{\sum_{i=1}^{N} SD_{A_i}}{w}$, where HMM_i is an individual stress detector, and A_i - its output.

Accuracies of individual classifiers and of their combinations were evaluated according to two different variants of "leave one day out" protocol:

- semi-personal: reference models are created in person-specific ways, HMM models are trained using data of all other subjects in addition to the training data of the target individual
- person-specific: reference models are created in person-specific ways, HMM models are trained using data of the target individual only

These two variants were chosen because the first dataset contained at maximum five days of data per subject, which is too small data size for training person-specific HMM models. Therefore "leave one day out" protocol for the first dataset was only realised as follows: for each target individual we excluded one day from the data and used his/her remaining days for creating his/her reference models. Then, we used these remaining days of the target individual and data of all other subjects to train HMM models and used these models to calculate stress scores of the test day. Then the procedure was repeated for all other days of this person and for all other persons.

The second dataset contained enough data to train personal-specific models, and for this dataset we compared results of semi-personal and person-specific training. In the latter protocol reference models and HMM models were built using all data of this person except for one day, then the resulting models calculated stress scores for this day, and the procedure was repeated for all other days, acquired for this person. Obtaining semi-personal models, however, was done slightly differently from the tests with the first dataset; namely, in this case reference models were not built from all available training data of the target subject. Instead, for creating reference models we used 10 randomly selected days of each subject in order to study whether it is really necessary to wait until long interaction histories are acquired, or the system can start detecting stresses soon after installation.

In all experiments, building reference models and training HMM models were fully unsupervised. The whole system, however, requires specifying certain hyper-parameters. The issue of their optimal choice requires further study; in this work hyper-parameters were chosen based on common sense and availability of training data. We have chosen time window length equal to three hours and window shift equal to one hour because human beings do not follow strict schedules and for example times of leaving a workplace, as well as times of going to sleep, may easily vary by one-three hours. We have chosen 300 min of data, collected in one place, to be "long enough" time to build a place-specific reference model because for the chosen time window of three hours this threshold would allow to build such models very quickly and to refine them after more visits to these places: for example, time-specific and workplace-specific models could be created after just couple of days of data collection in a workplace, a model of normal behaviour in a favourite restaurant could be created after about five lunches there, and so on.

We also needed to specify parameter w in formulas (12) and (13) to calculate stress thresholds. As the share of stressful days in our data exceeded one third, we did not expect scores of stressful days to be notable outliers and hence experimented with $w = 2$; $w = 3$; $w = 4$, and $w = 6$.

4.3 Experimental results

4.3.1 Accuracies of individual detectors

Stress scores of detectors ***HMM-Mean*** and ***HMM-SD*** were calculated according to formulas (8) and (9), whereas stress scores of detectors ***HMM-Mean-scatter*** and ***HMM-SD-scatter*** were calculated according to formulas (10) and (11) (see Section 3.3). Hidden states in HMM were named in the following way: state S_1 was the state with the highest probabilities of observing large positive deviations of current features from reference models; state S_3 was the state with the highest probabilities of observing large negative deviations of current features from reference models, and state S_2 was the remaining state. In other words, in ***HMM-Mean*** state S_1 represents the case when phone usage times exceed average times, while state S_3 represents the case when phone usage times are lower than the average. In ***HMM-SD*** state S_1 represents the case when standard deviations (SD) of phone usage times exceed average SD, while state S_3 represents the case when SD of phone usage times are lower than average SD. For both types of models state S_2 represents the case when phone usage features are fairly close to their average values. Consequently, ***HMM-Mean*** assigns high scores to the days when phone applications were used more actively than normally, whereas high day scores of variation-based detectors ***HMM-SD, HMM-Mean-scatter*** and ***HMM-SD-scatter*** mean that during this day variations in application usage times were greater than normally.

Therefore, first of all, we studied whether different individual stress detectors assign higher or lower score to stressful days in comparison with normal days. Table 1 presents accuracies of individual stress detectors with weight $w = 2$: with this weight more prominent outliers are classified as stresses than with other weights. Accuracies were calculated as follows:

$TrueStress = \frac{NstressOK}{Nstress}$; $TrueNormal = \frac{NnormalOK}{Nnormal}$; $TotalAccuracy = \frac{NnormalOK+NstressOK}{Nnormal+Nstress}$; where

NstressOK is the number of correctly classified stressful days; *NnormalOK* is the number of correctly classified normal days, *Nstress* is the number of stressful days in the dataset, and *Nnormal* is the number of normal days.

Table 1 shows that in the experiments with the first dataset, accuracies of all individual detectors were rather low. This is an expected result because to date behaviour-based stress detection has been successful only when person-specific models were trained; use of other models resulted in fairly low accuracies even when such models were trained in a fully supervised way [13, 23]. In the experiments with the second dataset semi-personal models achieved higher accuracies because refining categorisations of phone applications and availability of longer interaction histories improved reference models of normal behaviour. Person-specific models achieved the highest accuracies; hence, not only models of normal behaviour should depend on a person, but also mappings of deviations from normal behaviour into HMM states (for example, for some individuals only notable deviations from normal behaviour denote stress, whereas for other subjects also medium deviations denote stress).

As explained in the beginning of this section, ***HMM-SD*** assigns higher scores to the days on which standard deviations of application usage times per minute exceed their average

Table 1 Results of different HMM-MPM variants, achieved for $w = 2$ and for different thresholds

Stress detector	Threshold	First dataset semi-personal			Second dataset semi-personal			Second dataset person-specific		
		Total accuracy	True stress	True normal	Total accuracy	True stress	True normal	Total accuracy	True stress	True normal
HMM-Mean	TA_1	0.52	0.38	0.66	0.57	0.22	0.74	0.58	0.24	0.77
	TA_2	0.52	0.24	0.75	0.65	0.44	0.75	0.66	0.43	0.79
HMM-SD	TA_1	0.56	0.30	0.77	0.61	0.54	0.65	0.71	0.47	0.83
	TA_2	0.54	0.49	0.59	0.61	0.29	0.78	0.51	0.18	0.68
HMM-Mean-scatter	TA_1	0.58	0.27	0.84	0.65	0.44	0.75	0.68	0.44	0.81
	TA_2	0.56	0.22	0.84	0.55	0.20	0.73	0.56	0.35	0.67
HMM-SD-scatter	TA_1	0.54	0.41	0.64	0.65	0.35	0.80	0.65	0.32	0.82
	TA_2	0.49	0.24	0.70	0.61	0.52	0.65	0.56	0.21	0.77

values. ***HMM-Mean-scatter*** assigns higher scores to the days with notable differences between phone usage times in adjacent time windows, and ***HMM-SD-scatter*** assigns higher scores to the days with notable differences between SD of phone usage times in adjacent time windows. Table 1 shows that on both datasets these detectors achieved higher accuracies with the threshold TA_1 than with the threshold TA_2. This result suggests that on average human beings tend to use phone in less regular fashion during stressful days than during normal days. ***HMM-Mean***, on the other hand, assigns higher scores to the days on which phone applications were used longer than usually. On the first dataset ***HMM-Mean*** achieved equal total accuracies with both thresholds, but true stresses were detected more accurately with the threshold TA_1; hence, on average subjects in the first data collection used phone more during stressful days than during normal days. Subjects in the second dataset used phone less on stressful days, but not-so-big differences between accuracies of ***HMM-Mean*** on the second dataset, achieved with the two thresholds, suggest that change in phone usage times is less reliable stress indicator than increase in variations of phone usage times.

Semi-personal models did not achieve higher accuracies with other weights than that reported in Table 1, and we are not presenting them here in details. Table 2 presents results of the individual detectors, achieved by person-specific models for the second dataset with other weights (best results are highlighted). Results in Table 2 are presented only for the thresholds, most appropriate for each stress detector. Results of variation-based detectors are presented only for the threshold TA_1, because their accuracies with the threshold TA_2 were notably lower: for example, total accuracy of ***HMM-SD*** with the threshold TA_2 and weights $w = 4$ and $w = 3$ was 0.47, which is much lower than ***HMM-SD*** accuracies with the threshold TA_1 and weights $w = 4$ and $w = 3$ (0.70 and 0.72 respectively). On the contrary, results of ***HMM-Mean*** detector are presented for the threshold TA_2 because ***HMM-Mean*** accuracies with the threshold TA_1 were lower, although the differences between the accuracies with these two thresholds were not as notable as in case of variation-based detectors. For example, total accuracy of ***HMM-Mean*** with the threshold

Table 2 Results of different HMM-MPM variants, achieved by person-specific models for the second dataset

Stress detector	Threshold	Weight	Total accuracy	True stresses	True normal
HMM-Mean	TA_2	6	0.59	0.57	0.60
		4	0.60	0.54	0.64
		3	0.62	0.47	0.70
HMM-SD	TA_1	6	0.64	0.65	0.64
		4	***0.70***	***0.59***	***0.75***
		3	***0.72***	***0.54***	***0.81***
HMM-Mean-scatter	TA_1	6	0.62	0.54	0.67
		4	0.64	0.51	0.71
		3	0.65	0.51	0.73
HMM-SD-scatter	TA_1	6	0.55	0.57	0.53
		4	0.62	0.57	0.64
		3	0.62	0.57	0.64

TA_1 and weights $w = 4$ and $w = 3$ was 0.56, while ***HMM-Mean*** accuracy with the threshold TA_2 and the same weights was equal to 0.60 and 0.62 respectively.

Table 2 shows that when data size is large enough for training person-specific models, ***HMM-SD*** learns to distinguish between stressful and normal days with fairly high accuracy 70%.

One of the test subjects in the second data collection kindly provided additional information regarding labelled days: several days were described as "very stressful", several days of an illness and conference trip were also marked, and for some days this subject stated that only part of a day was stressful. "Part of a day was stressful" description, however, does not mean that stress level was low; it only means that the stress did not last the whole day. Figure 2 displays scores of four different person-specific stress detectors for these days.

Figure 2 shows that none of the individual detectors can reliably differentiate between "part of day stressful", "whole day stressful", and "very stressful" descriptions, and that days, described as "part of day stressful", often received higher scores than "very stressful" days. For variation-based detectors (***HMM-SD, HMM-SD-scatter*** and ***HMM-Mean-scatter***) it is natural to assign the highest scores to cases when stress did not last the whole day because of the differences between application usage during relatively normal and during stressful periods. Scores of a time-based detector, ***HMM-Mean***, depend on application usage times, and subject's descriptions explain that she did not use her phone actively during "very stressful days": on day 19 she had face-to-face meetings, on days 26 and 45 she edited documents on her computer before deadlines, and on day 29 she also worked on a computer plus had a party in the evening. As the phone was not used almost at all on day 29, all detectors assigned low scores to this day. Unfortunately, inactive phone usage could not be chosen as an indicator of stress because it may occur also during normal days (and indeed occurred in our data). Three other "very stressful" days received fairly high scores from ***HMM-SD***, however, which is a promising result.

The first day of an illness, day 3, received fairly high scores from all variation-based detectors, whereas the second day received fairly low scores. This result is logical because during the first day the subject worked, and it is not easy to work while feeling sick. Second day of an illness the subject spent at home, so this day was fairly relaxing. Disagreement between scores of different classifiers, calculated for the third and fourth days of an illness, is also logical: during these days the subject worked again, and she did not describe these days as stressful, but they were not really normal either.

High disagreement between scores of different classifiers regarding days of conference trip (days 50–54) is due to insufficient data quantity for new locations. Because of this, location-specific reference models for conference area were created only by the end of the conference, and phone usage patterns in a conference were mainly compared with the reference model of "elsewhere" location, i.e. an average of patterns in all locations outside of typical places. A more accurate comparison could be achieved by creating a separate reference model of "travelling" behaviour, which we plan to do in future. Even without this model ***HMM-SD*** and ***HMM-SD-scatter*** assigned the highest scores to the most stressful conference day—day 53, when the subject had a conference presentation. Fairly high scores for the day 54 are also justified by a long delay of a plane during return trip - although the subject did not label this delay as stress, it was nevertheless unpleasant and tiring.

The majority of stresses of this subject were due to work-related issues, but stress on day 59 was caused by a message about health problem at friend's family, i.e. this stress was

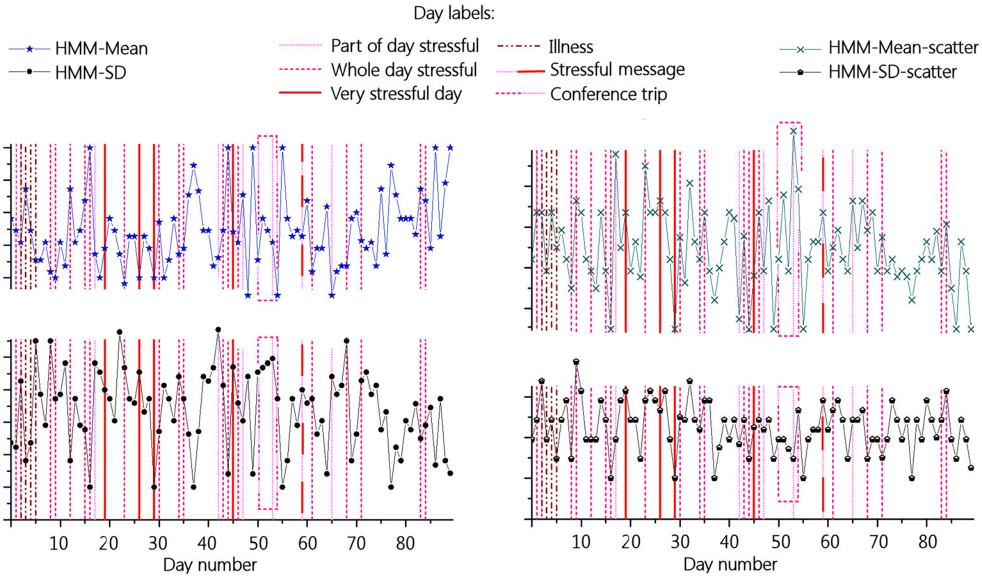

Fig. 2 Stress scores of different HMM-MPM variants, calculated by person-specific models of one subject

personal. Variation-based classifiers assigned fairly high scores to this day, suggesting that both personal and work-related stresses display themselves in hectic behaviour.

4.3.2 Accuracies of combined detectors

Table 3 and Table 4 present results of the selected combined detectors, achieved for the first and second datasets respectively (other possible combinations did not result in higher accuracies than that presented in these tables). Both tables present accuracies, achieved with the threshold TA_1 because accuracies with the threshold TA_2 were lower.

Table 3 and Table 4 show that combinations of detectors, relying on variations of phone usage times, achieved on average higher accuracies than combinations, involving ***HMM-Mean***, except for ***HMM-Mean-scatter + HMM-SD-scatter*** combination. Figure 2 suggests these two detectors work in similar ways instead of complementing each other, so their fusion is not likely to achieve as good results as a fusion of complementary modalities. Interestingly, combinations, involving ***HMM-Mean***, achieved higher accuracies with the threshold TA_1 than with the threshold TA_2; hence, these combinations classified as stressful days when phone usage was more active than normally. This can happen for example when test subjects use their phones for work purposes and get stressed by increase in work-related communications, or when their personal communications interfere with their work-related calls. Table 3 and Table 4 show also that scores of the stresses are not prominent outliers: in all cases the best accuracies were achieved with the weights $w = 4$ and $w = 3$, whereas $w = 2$ resulted in too high threshold value.

Table 4 shows that person-specific models outperformed semi-personal models also in case of combined detectors, but the best combined detector (***HMM-SD + HMM-SD-scatter + HMM-Mean-scatter***) nevertheless achieved reasonably high accuracy 68–70%, which means that stress detection system does not necessarily require long background data collection; semi-personal models can be trained and used for stress detection after just 10 days of data collection, and the system can switch to using more accurate person-specific models after collecting more data.

Figure 3 displays the stress scores of the combined detector that worked best of all for both datasets (***HMM-SD + HMM-SD-scatter + HMM-Mean-scatter***), calculated for the same subject and the same days with detailed descriptions as in

Table 3 Results of the selected combined HMM-MPM variants, achieved for the first dataset

Stress detector	Weight	Total accuracy	True stresses	True normal
HMM-SD + HMM-SD-scatter	6	0.53	0.54	0.52
	4	0.53	0.48	0.57
	3	0.54	0.46	0.61
	2	0.54	0.35	0.70
HMM-SD + HMM-Mean-scatter	6	***0.60***	***0.51***	***0.68***
	4	***0.62***	***0.51***	***0.70***
	3	0.59	0.46	0.70
	2	0.56	0.35	0.73
HMM-Mean-scatter + HMM-SD-scatter	6	0.57	0.38	0.73
	4	0.59	0.38	0.77
	3	0.59	0.35	0.80
	2	0.59	0.32	0.82
HMM-Mean + HMM-Mean-scatter	6	0.57	0.62	0.52
	4	0.57	0.59	0.55
	3	0.52	0.43	0.59
	2	0.58	0.30	0.82
HMM-SD + HMM-SD-scatter + HMM-Mean-scatter	6	***0.59***	***0.57***	***0.61***
	4	***0.59***	***0.54***	***0.64***
	3	0.58	0.46	0.68
	2	0.53	0.27	0.75
HMM-SD + HMM-Mean + HMM-Mean-scatter	6	0.53	0.57	0.5
	4	0.58	0.57	0.54
	3	***0.59***	***0.54***	***0.64***
	2	0.56	0.24	0.82

Table 4 Results of the selected combined HMM-MPM variants, achieved for the second dataset

Stress detector	Weight	Semi-personal models			Person-specific models		
		Total accuracy	True stresses	True normal	Total accuracy	True stresses	True normal
HMM-SD + HMM-SD-scatter	6	0.62	0.21	0.82	0.66	0.72	0.64
	4	0.64	0.19	0.87	***0.71***	***0.63***	***0.74***
	3	0.67	0.17	0.92	***0.73***	***0.60***	***0.80***
	2	0.68	0.14	0.94	0.74	0.51	0.85
HMM-SD + HMM-Mean-scatter	6	0.60	0.48	0.66	0.67	0.66	0.67
	4	0.62	0.44	0.70	***0.68***	***0.56***	***0.74***
	3	0.68	0.43	0.80	0.68	0.51	0.77
	2	0.69	0.33	0.87	0.69	0.40	0.84
HMM-Mean-scatter + HMM-SD-scatter	6	0.61	0.59	0.62	0.60	0.49	0.67
	4	0.64	0.51	0.71	0.61	0.49	0.68
	3	0.66	0.51	0.74	0.63	0.46	0.73
	2	0.64	0.32	0.80	0.66	0.43	0.79
HMM-Mean + HMM-Mean-scatter	6	0.61	0.43	0.70	0.63	0.47	0.71
	4	0.65	0.30	0.83	0.67	0.41	0.81
	3	0.69	0.21	0.93	0.68	0.29	0.88
	2	0.67	0.10	0.96	0.68	0.19	0.93
HMM-SD + HMM-SD-scatter + HMM-Mean-scatter	6	0.63	0.51	0.70	0.67	0.68	0.67
	4	***0.68***	***0.49***	***0.78***	***0.70***	***0.63***	***0.74***
	3	***0.70***	***0.40***	***0.85***	***0.71***	***0.56***	***0.79***
	2	0.68	0.19	0.92	0.71	0.46	0.84
HMM-SD + HMM-Mean + HMM-Mean-scatter	6	0.63	0.43	0.74	0.71	0.54	0.79
	4	0.68	0.38	0.82	***0.74***	***0.49***	***0.87***
	3	0.70	0.24	0.93	0.70	0.37	0.87
	2	0.66	0.06	0.97	0.69	0.15	0.98

Fig. 2. In addition, Fig. 3 displays cumulative phone usage times per day, i.e. sums of usage times of all applications.

Figure 3 shows that phone usage times vary a lot and do not well correlate with stresses. This can be explained by variety of work-related and personal activities of this subject: during some days she had a lot of work-related communications; on some other days she mainly worked at her computer; on some other days she worked at home in the daytime and/or in the evenings; on some days she travelled or attended parties. Nevertheless the combined detector succeeded to recognise three out of four very stressful workdays, displayed in Fig. 3, and a day of a stress, caused by personal reasons. Scores of these “very stressful” days were not the highest among all scores, but the highest scores were also assigned to the days

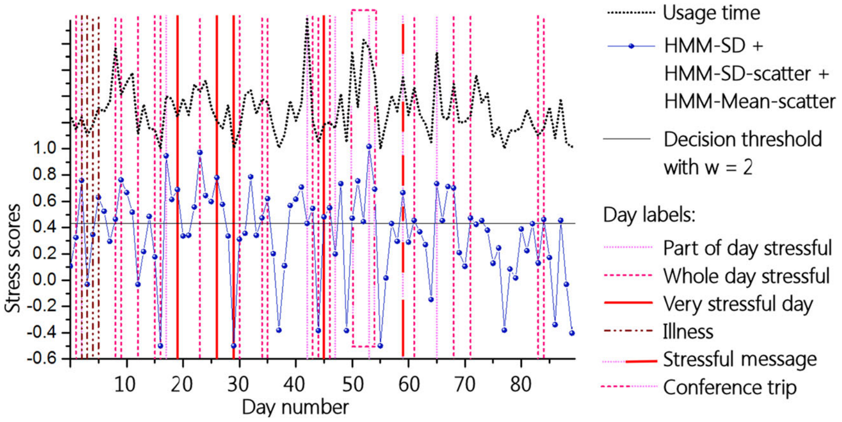

Fig. 3 Stress scores of the best combined detector and cumulative usage times of phone applications, calculated for the test person who provided detailed day labels

labelled as "stressful". These days were described by the subject as follows: day 17—as "busy, three face-to-face meetings", day 23—as "I worked long to get the document ready before my holidays", and day 53 was a day of public presentation in a conference.

The combined detector assigned fairly high scores also to a few normal days. According to the descriptions of the test subject, nothing special happened on day 41, so its high score remains a mystery. Other days were described as follows: on day 32 one usual activity was cancelled, and the day did not follow usual routines. It did not cause stress, but the day was not perfectly normal either. On day 39 the subject "felt a bit tired in the morning", probably due to poor night sleep; and on day 40 the subject had several meetings, which implies certain tension.

Very low scores, assigned to three stressful days, are due to very low phone usage time on these days: on days 16 and 29 the subject mainly worked on a computer, and on day 29 she also attended a party; on day 44 the subject spent a lot of time driving and carrying things.

5 Discussion

5.1 Comparison with results of other works

Direct comparison of our results with that of other works is not possible because all works experimented with different datasets and evaluated their approaches in different ways. For example, Sano and Picard [28] monitored each subject for five days, but did not report accuracies of stress detection on daily basis; instead, they classified each person into one of two classes: "has reported high stress" vs. "has not reported high stress". Subjects' classification accuracy, achieved using the best mobile phone data feature (SD of percentage of "screen on" times between 6 and 9 pm), was 75% despite fully supervised training. Bogomolov et al. [5] evaluated stresses on a daily basis using two-level scale (i.e. whether a day was stressful or not). They achieved 72% accuracy using fully supervised training on a fairly large dataset: data collection lasted eight weeks. Sysoev et al. [31] also evaluated stresses using a two-level scale and achieved an average accuracy 73% using different fully supervised methods, but dataset size is not clearly described in the paper.

Our unsupervised method achieved about 60% accuracy on a dataset containing only three-five days of data per subject and 68% accuracy in the case when 10 days of the target person data were used for training semi-personal models. Using longer phone usage histories for training person-specific models allowed to achieve 70–73% accuracies in different system configurations. We believe that collecting even longer histories would allow to refine place-specific models of normal behaviour and further increase accuracy. It would be also beneficial to learn activity-specific models, e.g. separate models for travelling, office work, working remotely from home and normal home usage, which can be achieved by monitoring of specific work-related applications, such as for example VPN client or work email. Due to unsupervised learning, this kind of model refining requires just time, but no efforts from end users.

Other works evaluated stresses on a three-level scale (low, medium and high stress) and two-three times per day, e.g. whether stresses occurred in the morning, afternoon or evening. In [21] collection of labelled data during four weeks three times per day resulted in 62% accuracy of person-specific models, trained using data of the target person only. The majority of tested approaches to reuse data of other individuals for detecting stresses of the target subject resulted in similar or lower accuracies. One method allowed to increase accuracy to 72%, but only when data of three most similar persons were reused; attempts to reuse data of other numbers of subjects resulted in lower accuracies. Other works, studied methods to reuse data of multiple subjects for decreasing the need for labelled data of the target person, also reported strong dependency of the accuracies on parameter choice, whereas in our study accuracies 57–59% on one dataset and 68–73% on another dataset were achieved for fairly broad range of thresholds, and threshold choice mainly affected trade-offs between rates of true positives and true negatives.

Regarding stress recognition on daily basis vs. two-three times per day, for personalised coaching to cope with stresses it may be more beneficial to detect exact time of stress and to associate it with a certain event, but this would probably require making physiological sensors more practical than they are nowadays. Regarding detection of stress occurrences vs. stress assessment on the three-level scale, various studies reported that the most dangerous kind of stress is chronic stress [19, 22, 30] and long-lasting stress of low intensity may have as bad an impact on health as a short stress of high intensity [17]. Therefore unsupervised stress detection on daily basis may suffice for long-term well-being monitoring, whereas supervised approaches may fail to obtain necessary training data. Human beings can be lazy even to upgrade computer security from time to time [35], and all the more so when data labels are requested regularly [2]. In addition, phone usage behaviour may change with time or after getting a new phone model or a new job, and these changes would require updating of stress detectors. Adaptation of supervised models would require the additional collection of labelled data in such cases, while unsupervised learning would not require any user efforts.

5.2 Specifics of real life

By definition, "stress is a state of mental tension and worry caused by problems in life, work, etc." [Merriam-Webster dictionary, http://www.learnersdictionary.com/definition/stress, last accessed 24.04.2017]. Therefore stress is not necessarily a short-term state: stress can be caused by tiredness or anxiety due to postponed tasks, lack of inspiration or a chain of events none of which would be stressful alone. The most typical assumption in stress research, however, is that stresses are fairly short-term conditions: that is why lab studies are performed by placing test subjects into difficult situations for a while and measuring their immediate reactions. Field studies often aim at detecting short-term arousal [16] or evaluating mental workload during short time intervals [9] too. Thanks to the detailed comments, provided by one of the test subjects, we observed that in real life tiredness, dissatisfaction and a chain of trivial episodes cause feeling of stress fairly often and hence realistic stress detectors cannot simply reuse methods, developed on the basis of lab data. For example, the subject explained certain "stressful" labels as follows:

- "*I tried to write, but I wasn't inspired.*"
- "*Nothing special happened but I didn't feel fully recovered from yesterday.*"
- "*This should not have been stressful day, but somehow I still felt stressed or not fully recovered.*"
- "*Because of many interruptions during the day, I just couldn't focus the things I wanted to.*"

In the experiments, reported in this paper, we observed that such stresses also cause deviations from normal behaviour, although not so prominent (for example, see days 35 and 43 in Fig. 3), and hence, they may be easier to recognise on daily basis than in shorter time periods.

Stress detection on the basis of mobile phone data does not require users to wear any additional sensors, and this is its great advantage for real life because existing wearable physiological sensors are sensitive to attachment problems, and physiological signals are affected by physical activities, food intake and hormonal changes [9, 24]. Stress detection on the basis of mobile phone data, however, is influenced by inconsistency of phone usage patterns: users may pay little attention to their phones when they are stressed, but may ignore them also when they are relaxed and lazy, and these two cases cannot be distinguished by analysing phone usage data because none of them presents data to analyse. Fig. 2 and Fig. 3, together with the detailed descriptions, written by one of the test subjects, illustrate this problem: several stressful days received as low scores as several normal days because of very little phone usage during these days.

Subjective perception of stress creates additional difficulties: for example, one subject labelled the day as normal, but added a comment: "*nothing special happened, but because of several meetings and emails in the morning I felt very (too) busy.*" Usually, a feeling of "being too busy" does not develop into a feeling of being stressed when problems are solved successfully, but whether such things can be distinguished on the basis of phone usage data or not requires additional long studies. In the present study our method classified this day as stressful when stress detection threshold was calculated with weights $w = 6$ and $w = 4$, and non-stressful with smaller weights, but we are unaware of any methods to distinguish between positive feelings of coping well with a difficult day and negative "being too busy" feelings even on the basis of physiological sensors: it seems that the majority of studies into stress detection were concerned only with recognition of negative stresses.

5.3 Specifics of unsupervised learning

In this study we used fairly generic phone usage features, moreover that not so many previous works reported how chosen features correlate with stresses. For example, in [21] the following features were chosen: number and duration of incoming/outgoing calls; duration of incoming/outgoing SMS; most common contact-SMS; number and duration of used applications (social, system). Some of these features are highly private, and their correlation with stresses is person-dependent, e.g. for somebody long duration of outgoing calls may denote his/her immersion into a stressing work problem, whereas for another person it may denote a relaxing chat with family members. In supervised training in [21] person-specific meanings of different features were learned from data labels, but unsupervised training cannot be controlled this way. Unsupervised training can be controlled only by pre-defining either "greater than" or "lower than" thresholds. Then the proposed method can for example learn that for some person call duration is a non-discriminative feature (if in his/her model probabilities of observing long calls are nearly equal in all states), and it can learn that for some person only very long calls denote stress, whereas for another person also a bit longer than average calls denote stress. However, the proposed method cannot learn that long calls denote the normal condition of one person and stress of another person because input data does not contain any helpful information. Therefore, we consider valuable our finding that stresses display themselves in the more hectic usage of phone applications than normal days.

In our study variation-based stress detectors worked for test subjects of many occupations; hence, hectic phone usage patterns seem to be a fairly generic stress indicator. To the best of our knowledge, none of the previous studies into stress detection reported the value of hectic phone usage patterns for stress recognition. On the contrary, Sano and Picard [28] reported that high stress correlated with a smaller standard deviation of "screen on" times in the evenings. Correlations between high stress and standard deviation of "screen on" times during daytime, as well as correlations between other stress levels and SD appeared to be insignificant in [28]. As Sano and Picard [28] also reported that high stress correlated with a smaller mean of "screen on" times in the evenings, it may be so that the decrease in SD was caused by the decrease in phone usage times. Unlike [28], we did not use variations of usage times of different applications directly; instead, we used deviations of normalised usage times from their typical values, and we observed an increase in SD of these deviations in cases of both high and medium stresses.

5.4 Future work

In future we plan to study which additional data could help to distinguish between stress and laziness and between positive and negative perceptions of "being too busy". For example, accelerometer data may be indicative of stress [10], and location data may be used in more ways than we employed in this work, e.g. mobility radius [28] was found indicative of depression. It may be worth also to combine phone usage data with data from environmental sensors, such as motion sensors or keyboard/mouse usage data.

We also plan to collect long-term data of more individuals and to study individual differences and different types of stressors. As various works observed, people are very heterogeneous and their stress-proneness differs. Self-assessment is regarded as a good measure of stress [3], but self-assessment requires the subjects to constantly pay attention to their conditions, which may increase their mental workloads and add up to their stresses. Unobtrusive sensor-based stress detection does not require any efforts from the subjects and hence may facilitate automatic recognition of individual stressors and personalised tutoring to help cope with problems. For example, it may help to understand whether everyday stressors of a certain individual are cognitive-emotional (e.g. rumination, hurry), social (e.g. interpersonal conflicts, financial problems, high work demands), physiological (e.g. dieting, insufficient sleep, exercise) or environmental (e.g. noise, pollution), and provide appropriate support. Individual's responses to stress depend also on personality and learned coping styles [6, 25], but in any case, the most harmful case is the chronic stress [19, 22, 30] that occurs when stressors are persistent and long-lasting. In future we plan to complement stress data with data regarding potential daily life stressors. We also plan to attempt at recognising individual daily stressors and developing methods to increase individual's awareness of how they are exposed to stress, i.e. help people reflecting their stresses and activities affecting them. With a further research on a system integrating personalised objective stress detection and coping solutions, we could help users to reduce subjective stress-related problems and to better prevent a chronic stress.

6 Conclusion

This work proposed unsupervised method for detecting stresses on the basis of mobile phone usage data, collected in fairly privacy-preserving way: when application of a certain category is started/stopped or moved to a foreground/background, we only store activation of a category; application name, phone number, typed/retrieved texts, etc. are not stored. Therefore the proposed method is truly unobtrusive both from the point of view of employed sensors (does not require any additional devices) and from the point of view of data processing methods (requires neither data labelling nor storage of private data), whereas previous studies, to the best of our knowledge, required users either to wear dedicated stress recognition devices and properly attach them, or to provide large sets of labelled data for system training, or both. As human beings usually are not keen to provide sizeable amounts of training data, supervised learning does not seem practical, moreover that human behaviour evolves over time and thus stress detectors cannot be trained once and forever. Unsupervised training, on the contrary, allows for lifelong learning, e.g. automatic updates of trained models when the users get new phones or when their personal lives or work tasks change.

The proposed method is based on the finding that stresses display themselves in more hectic phone usage patterns than normal days. Unsupervised stress recognition is performed by discrete HMM with MPM decisions. The proposed HMM models each day as a sequence of time windows and learns which kind of phone usage behaviour is normal for each time window in different places, thus providing for possible time-dependency and location-dependency of user behaviour. In the experiments with the dataset, collected in the course of normal lives of test subjects and containing from three to five days of data per person, the proposed method achieved about 60% accuracy, which is a good result considering that data size was too small to learn person-specific models, and non-personalised behavioural models did not achieve high stress detection accuracies even in the works employing fully supervised training. In the experiments with another real-life dataset, which size allowed learning person-specific models, the proposed method achieved about 70% accuracy, which is similar to the accuracies of fully supervised methods, reported by other works. Furthermore, reasonable stress detection accuracy 68% was achieved when just 10 days of the data of the target subject were used for training his/her semi-personal models, which means that stress recognition application should not necessarily require long background data collection. This capability may facilitate acceptance of the proposed system by not-so-patient end users.

In this study the proposed algorithm successfully recognised different types of stressful days: days when stresses were caused by fairly short-term events (e.g. work-related presentations and meetings, as well as a personal message about health problem at friend's family) and days when stresses were due to long working hours in order to meet a deadline, or due to tiredness and long-lasting anxiety. The latter stress type cannot be easily induced in lab studies and hence lab studies did not report whether their algorithms are capable of detecting such stresses. As test subjects rarely provide as many stress labels as they are requested to provide, not to speak about detailed labels, to the best of our knowledge, field studies did not discuss detection of such stresses either. In data, collected in this study, such stresses occurred fairly often; hence, they may notably contribute to stress-related illnesses and therefore deserve more attention.

The main drawback of the proposed approach is its inability to detect anything when users ignore their phones, but smartphones become increasingly more "natural parts" of our lives: for example, stressed individuals may use them to discuss their problems in social networks, or may need to make work-related phone calls. Furthermore, this problem is intrinsic to all sensor types: physiological devices work only when users put them on, and environmental sensors work only when users are present nearby. Therefore, a more reliable solution could be to combine diverse sensors in an unobtrusive way, and we consider results, achieved using mobile phone data alone, as a first step for encouraging further studies into unobtrusive stress detection.

References

1. Alberdi A, Aztiria A, Basarab A (2016) Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. J Biomed Inform 59:49–75. https://doi.org/10.1016/j.jbi.2015.11.007
2. Adams, P., Rabbi, M., Rahman, T., Matthews, M., Voida, A., Gay, G., Choudhury, T., Voida, S. (2014) Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild, 8th International Conference on Pervasive Computing Technologies for Healthcaren.d. pp. 72–79
3. Andreou E, Alexopoulos EC, Lionis C, Varvogli L, Gnardellis C, Chrousos GP, Darviri C (2011) Perceived stress scale: reliability and validity study in Greece. Int J Environ Res Public Health 8(8):3287–3298. https://doi.org/10.3390/ijerph8083287
4. Bakker J, Holenderski L, Kocielnik R, Pechenizkiy M, Sidorova N (2012) Stess@ work: from measuring stress to its understanding, prediction and handling with personalized coaching. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, pp 673–678
5. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland A (2014) Pervasive stress recognition for sustainable living. In: Pervasive computing and communications workshops (PERCOM workshops), 2014 I.E. international conference on, pp 345–350
6. Bolger N, Zuckerman A (1995) A framework for studying personality in the stress process. J Pers Soc Psychol 69(5):890–902. https://doi.org/10.1037/0022-3514.69.5.890
7. Chandola, V., Banerjee, A., Kumar, V., Anomaly detection: a survey, ACM computing surveys 41, 3, article 15 (2009), 58 pages
8. Chandola V, Banerjee A, Kumar V (May 2012) Anomaly detection for discrete sequences: a survey, knowledge and data engineering. IEEE Transactions on 24(5):823–839
9. Cinaz B, Arnrich B, La Marca R, Tröster G (2013) Monitoring of mental workload levels during an everyday life office-work scenario. Pers Ubiquit Comput 17(2):229–239. https://doi.org/10.1007/s00779-011-0466-1
10. Garcia-Ceja, E., Osmani, V., Mayora, O., Automatic stress detection in working environments from smartphones' accelerometer data: a first step, IEEE journal of biomedical and health informatics 2016
11. Gjoreski M, Gjoreski H, Lutrek M, Gams M (2015) Automatic detection of perceived stress in campus students using smartphones. In: Intelligent environments (IE), 2015 international conference on, pp 132–135
12. Gjoreski, M., Gjoreski, H., Lutrek, M., Gams, M. (2016) Continuous stress detection using a wrist device: in laboratory and real life, Ubicomp 2016 Adjunct, pp. 1185–1193
13. Ferdous, R., Osmani, V., Mayora, O., Smartphone app usage as a predictor of perceived stress levels at workplace, 9th international conference on pervasive computing technologies for healthcare 2015
14. Hernandez, J., Morris, R.R., Picard, R.W. (2011) Call center stress recognition with person-specific models, In Proceedings of the 4th international conference on affective computing and intelligent interaction, pp. 125–134
15. Hovsepian, K., al'Absi M, Ertin, E., Kamarck, T., Nakajima, M., Kumar, S (2015) cStress: towards a gold standard for continuous stress assessment in the mobile environment, ACM International Joint Conference on Pervasive and Ubiquitous Computing
16. Kusserow M, Amft O, Troster G (2013) Modeling arousal phases in daily living using wearable sensors, in affective computing. IEEE Transactions on 4(1):93–105
17. Lamb S, Kwok KCS (2016) A longitudinal investigation of work environment stressors on the performance and wellbeing of office workers. Appl Ergon 52:104–111. https://doi.org/10.1016/j.apergo.2015.07.010
18. Lazarus RS (1993) From psychological stress to the emotions: a history of changing outlooks. Annu Rev Psychol 44(1):1–21. https://doi.org/10.1146/annurev.ps.44.020193.000245
19. Lucini D, Di Fede G, Parati G, Pagani M (2005) Impact of chronic psychosocial stress on autonomic cardiovascular regulation in otherwise healthy subjects. Hypertension 46(5):1201–1206. https://doi.org/10.1161/01.HYP.0000185147.32385.4b
20. Matrix (2013) Economic analysis of workplace mental health promotion and mental disorder prevention programmes and of their potential contribution to EU health, social and economic policy objectives, Executive Agency for Health and Consumers, Available at: http://ec.europa.eu/health/mental_health/docs/matrix_economic_analysis_mh_promotion_en.pdf
21. Maxhuni A, Hernandez-Leal P, Sucar LE, Osmani V, Morales EF, Mayora O (2016) Stress modelling and prediction in presence of scarce data. J Biomed Inform 63:344–356. https://doi.org/10.1016/j.jbi.2016.08.023
22. McEwen, B. S. (2012). Brain on stress: how the social environment gets under the skin. Proceedings of the National Academy of Sciences, 2012, 109, Supplement 2: 17180–17185
23. Muaremi A, Arnrich B, Tröster G (2013) Towards measuring stress with smartphones and wearable devices during workday and sleep. BioNanoScience 3(2):172–183. https://doi.org/10.1007/s12668-013-0089-2
24. Plarre, K., Raij, A., Hossain, S.M., Ali, A. A., Nakajima, M., al'Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., Smailagic, A., Wittmers, L.E. (2011) Continuous inference of psychological stress from sensory measurements collected in the natural environment, in Information processing in sensor networks (IPSN), 10th international conference on pp. 97–108
25. Puttonen et al (2005) Cloninger's temperament dimensions and affective responses to different challenges. Comprehensive Psychiatry 46(2):128–134

26. Rahman, Md.M, Bari, R., Ali, A.A., Sharmin, M., Raij, A., Hovsepian, K., Hossain, S.M., Ertin, E., Kennedy, A., Epstein, D.H., Preston, K.L., Jobes, M., Beck, J.G., Kedia, S., Ward, K.D, al'Absi, M., Kumar, S (2014) Are we there yet?: feasibility of continuous stress assessment via wireless physiological sensors, In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics pp. 479–488
27. Rabiner LR (1986) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
28. Sano A, Picard RW (2013) Stress recognition using wearable sensors and mobile phones. In: Affective computing and intelligent interaction (ACII), 2013 Humaine association conference on, pp 671–676
29. Shi, Y., Nguyen, M.H., Blitz, P., French, B., Frisk, S., Torre, F., Smailagic, A., Siewiorek, D., al'Absi, M., Kamarck, T., Kumar, S. (2010) Personalized stress detection from physiological measurements, In Proceedings of the 2nd International Symposium on Quality of Life Technology
30. Siegrist J (2008) Chronic psychosocial stress at work and risk of depression: evidence from prospective studies. Eur Arch Psychiatry Clin Neurosci 258(5):115–119. https://doi.org/10.1007/s00406-008-5024-0
31. Sysoev M, Andrej Kos A, Matevz Pogacnik M (2015) Noninvasive stress recognition considering the current activity. Pers Ubiquit Comput 19(7):1045–1052. https://doi.org/10.1007/s00779-015-0885-5
32. Vildjiounaite E, Kyllönen V, Mäkelä S-M, Vuorinen O, Keränen T, Peltola J, Gimel'farb G (2012) Semi-supervised context adaptation: case study of audience excitement recognition. Multimedia Syst 18(3):231–250
33. Vildjiounaite, E., Gimel'farb, G., Kyllönen, V., Peltola, J. 2015 Lightweight Adaptation of Classifiers to Users and Contexts: Trends of the Emerging Domain, The Scientific World Journal Article 434826, 29 p
34. Vildjiounaite E, Mäkelä S-M, Keränen T, Kyllönen V, Huotari V, Järvinen S, Gimel'farb G (July 2017) Unsupervised illness recognition via in-home monitoring by depth cameras, pervasive and mobile computing, volume 38. Part 1:166–187
35. Safeguards in a World of Ambient Intelligence, Eds. Wright et al., Springer, 2008
36. Xu Q, Nwe TL, Guan C (Jan. 2015) Cluster-based analysis for personalized stress evaluation using physiological signals. In: Biomedical and health informatics, IEEE journal of, vol.19, no.1, pp 275–281