Towards Wearable Stress Measurement

by

Javier Hernandez Rivera

M. S., LaSalle Ramon Llull University (2007)
B. S., LaSalle Ramon Llull University (2004)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

© Massachusetts Institute of Technology 2015. All rights reserved.

Signature redacted

Author

Program in Media Arts and Sciences
August 7, 2015

Certified by

Rosalind W. Picard
Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by

Prof. Pattie Maes
Academic Head
Program in Media Arts and Sciences
Towards Wearable Stress Measurement
by
Javier Hernandez Rivera
Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on August 7, 2015, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences

Abstract

Chronic psychological stress carries a wide array of pathophysiological risks, including cardiovascular and cerebrovascular diseases, diabetes, and immune dysregulation. An important step in managing stress, before it becomes chronic, is recognizing precisely when and where it occurs. This thesis creates and evaluates new methods to improve the measurement of stress by leveraging state-of-the-art wearable devices.

The first part of the thesis systematically compares gathering self-reported stress levels with head and wrist-worn devices, and compares them to the traditional cellphone in the pocket. In particular, 15 participants were asked to carry these devices during five days of their regular work day and to self-report their emotional state several times a day with our custom experience sampling application. We found that both head and wrist-worn devices significantly outperformed the phone in terms of the amount of answered prompts and the speed to start answering. However, different factors such as interaction types, screen size, and familiarity with the devices affected users’ experience and responses.

The second part of the thesis develops novel methods to comfortably capture physiological signals associated with the stress response. In particular, 36 participants were asked to carry either a head-worn device, a smartwatch or a smartphone while performing different “still” body postures in a controlled laboratory study. Using the proposed methods, we demonstrated that wearable motion-sensitive sensors inside these devices can capture heart and breathing rates as accurately as FDA-cleared devices from traditional body locations. Furthermore, using the data collected from the 15 participants, we demonstrated that our methods can be opportunistically used in real-life when people are relatively “still.” In our study, for instance, the head-worn device provided accurate heart rate assessments around 20% of the work day.

Finally, the third part of the thesis uses supervised learning methods to automatically infer self-reported stress levels from different types of wearable data, including physiological, contextual and behavioral signals. While there is not a one-size-fits-all solution, we found that electrodermal activity, head motion and atmospheric pressure were the more relevant signals across the 15 participants. Furthermore, we characterized many of the challenges that plague the task of real-life stress recognition.

This work is supported by the National Science Foundation (NSF CCF-1029585), the Samsung Think Tank Team, Google, the Caja Madrid fellowship, and the MIT Media Lab Consortium.

Thesis Supervisor: Rosalind W. Picard
Title: Professor of Media Arts and Sciences
Towards Wearable Stress Measurement

by

Javier Hernandez Rivera

Signature redacted

Thesis Reader ________________  Rosalind W. Picard
Professor of Media Arts and Sciences
Director of the Affective Computing Research Group
Massachusetts Institute of Technology

Signature redacted

Thesis Reader ________________  Karen S. Quigley
Research Associate Professor of Psychology
Co-director of the Interdisciplinary Affective Science Laboratory
Northeastern University

Signature redacted

Thesis Reader ________________  Pattie Maes
Professor of Media Arts and Sciences
Director of the Fluid Interfaces Research Group
Massachusetts Institute of Technology
Acknowledgements

The work presented in this thesis is the result of collaborations with, and support from, a large number of people. While there is not space to list all their names, I would like to acknowledge some of the people who have contributed the most to the successful completion of this work.

First, I want to express my deepest gratitude to my adviser Professor Rosalind Picard whose unflagging curiosity and hard work have been a source of inspiration throughout these years. You have not only enabled and supported this thesis in all possible ways but have also taught me how to become a better researcher, collaborator, and human being. You will always be a role model to follow. I also want to acknowledge Professor Karen Quigley and Professor Pattie Maes who have provided invaluable insights and comments during the Ph.D. They have pushed me to consider all possible research angles and helped me become a more rigorous and thorough researcher. Thanks for patiently witnessing all my mistakes and sharing your precious time with me.

Throughout this thesis, I have been very fortunate to collaborate with many brilliant researchers (e.g., Daniel McDuff, Rob Morris, Pablo Paredes, Ivan Riobo, Yin Li), promising undergraduate students (e.g., Christian Infante, Will Drevo), dexterous staff members (e.g., Rashmi Karki, Kristina Bonikowski), and inspiring world experts in various disciplines (e.g., Matthew Goodwin, Rich Fletcher, Gregory Abowd, Mary Czerwinski, Asta Roseway). Among all the collaborators, I want to specially acknowledge Rob Morris who introduced me to the exciting and challenging topic of stress measurement, and Daniel McDuff with whom I have been able to explore many exciting research problems, especially in the context of this thesis. Your research has always been a source of inspiration to me.
The Ph.D. journey can feel long and tiresome at times and people around you can play a very important role throughout the process. During the years, I found a lot of comfort in members of the Affective Computing group (e.g., Akane Sano, Szymon Fedor, Yadid Ayzenberg, Karthik Dinakar, Ming-Zher Poh) who have shared similar journeys and challenges along the way. In addition, I was fortunate to find close friends (e.g., Maria Jose Nieves, Tomas Simon, Nuria Jane, Alberto Rodriguez, Judith Amores, Xavier Benavides, Oier Ariño, Hasier Larrea) that helped me forget about “stress” when I needed it the most.

Finally, I want to express my most sincere appreciation for the people who have “always been there.” Those who loved me even during the most difficult and challenging times, and taught important life lessons that can only be learned over years of patience and perseverance. Without your love, I would have never been able to write these lines. These people are, of course, my loving parents (Tesi and Enrique), my siblings (Carlos, Enrique and Nuria), and last, but far from least, my life partner (Ethel). “Thanks” will never suffice.

“If you want to go fast, go alone. If you want to go far, go together”

African proverb
Contents

Abstract .............................................................................................................................. 3
Acknowledgements ........................................................................................................ 7
Contents .............................................................................................................................. 9
List of Figures .................................................................................................................. 15
List of Tables .................................................................................................................... 21
List of Abbreviations ..................................................................................................... 22
Introduction ..................................................................................................................... 23
  1.1 Motivation ............................................................................................................. 23
  1.2 Areas of Work ....................................................................................................... 25
  1.3 Thesis Aims ......................................................................................................... 26
  1.4 Dissertation Outline ............................................................................................ 27
Background Research ..................................................................................................... 31
  2.1 Stress Models ....................................................................................................... 31
  2.2 Fight or Flight Response ....................................................................................... 35
  2.3 Stress Measurement Approaches ......................................................................... 35
  2.4 Ambulatory Self-reports ...................................................................................... 38
  2.5 Context Gathering ............................................................................................... 40
  2.6 Comfortable Physiological Sensing .................................................................... 42
  2.7 Stress Recognition from Physiological Signals ................................................... 44
  2.8 Managing Stress with Technology ....................................................................... 45
  2.9 Conclusions ......................................................................................................... 51
Early Exploratory Work ................................................................................................. 53
  3.1 Stress Measurement ............................................................................................ 53
  3.1.1 Pressure-Sensitive Keyboard and Capacitive Mouse ....................................... 53
  3.1.2 Wearable Electrodermal Sensors ..................................................................... 55
3.2 Context Gathering ............................................................... 56
  3.2.1 Behavioral Activity .......................................................... 56
  3.2.2 Visual Context and Physiological Signals ....................... 58
3.3 Stress Management ............................................................. 60
  3.3.1 Smiles “in the Wild” ........................................................ 60
  3.3.2 Head-mounted Wearable Devices ........................................ 63
  3.3.3 Driving Scenario ............................................................. 65
3.4 Conclusions ........................................................................ 68

Real-life Workplace Study .......................................................... 71
  4.1 Wearable Devices ................................................................. 71
  4.2 Experimental Protocol ........................................................ 75
  4.3 Questions and Surveys .......................................................... 76
  4.4 Experience Sampling Questions ........................................... 79
  4.5 Ensuring Privacy of Data ...................................................... 80
  4.6 Recruitment and Compensation .......................................... 82
  4.7 Preliminary Data Overview .................................................. 82

Wearable Experience Sampling ................................................... 85
  5.1 Considered Locations .......................................................... 85
  5.2 Experience Sampling Tool ................................................... 87
    5.2.1 Implementation .............................................................. 87
    5.2.2 Interaction ................................................................. 88
    5.2.3 Prompting Criteria ....................................................... 90
  5.3 Results ................................................................................ 92
    5.3.1 Response Rates ........................................................... 92
    5.3.2 Response Times ........................................................... 94
    5.3.3 Response Distribution .................................................. 96
    5.3.4 Usability of Devices ..................................................... 98
    5.3.5 Visual Context ............................................................ 101
  5.4 Discussion ......................................................................... 102
  5.5 Conclusions ...................................................................... 105
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Discussion</td>
<td>153</td>
</tr>
<tr>
<td>7.4</td>
<td>Conclusions</td>
<td>154</td>
</tr>
<tr>
<td>8.1</td>
<td>Real-life Stress Measurement Challenges</td>
<td>155</td>
</tr>
<tr>
<td>8.2</td>
<td>Surveys</td>
<td>155</td>
</tr>
<tr>
<td>8.3</td>
<td>Self-reports</td>
<td>157</td>
</tr>
<tr>
<td>8.4</td>
<td>Data Preparation</td>
<td>159</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Wearable Devices</td>
<td>161</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Physiology Preprocessing</td>
<td>161</td>
</tr>
<tr>
<td>8.4.3</td>
<td>Feature Extraction</td>
<td>163</td>
</tr>
<tr>
<td>8.4.4</td>
<td>Feature Normalization</td>
<td>164</td>
</tr>
<tr>
<td>8.4.5</td>
<td>Refining Stress Ratings and Outlier Correction</td>
<td>166</td>
</tr>
<tr>
<td>8.5</td>
<td>Experimental Setting</td>
<td>166</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Class Labels</td>
<td>167</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Classification</td>
<td>167</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Missing Data</td>
<td>168</td>
</tr>
<tr>
<td>8.5.4</td>
<td>Performance</td>
<td>168</td>
</tr>
<tr>
<td>8.6</td>
<td>Results</td>
<td>170</td>
</tr>
<tr>
<td>8.7</td>
<td>Discussion</td>
<td>173</td>
</tr>
<tr>
<td>8.8</td>
<td>Conclusions</td>
<td>176</td>
</tr>
<tr>
<td>9.1</td>
<td>Discussion</td>
<td>181</td>
</tr>
<tr>
<td>9.2</td>
<td>Summary of Contributions</td>
<td>185</td>
</tr>
<tr>
<td>9.3</td>
<td>Limitations and Future Work</td>
<td>186</td>
</tr>
<tr>
<td>9.4</td>
<td>Concluding Remarks</td>
<td>190</td>
</tr>
<tr>
<td>A.</td>
<td>Tutorial Provided to Participants</td>
<td>191</td>
</tr>
<tr>
<td>B.</td>
<td>Surveys of the Main Study</td>
<td>195</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Pressure-sensitive keyboard (left), and capacitive mouse (right) ..................54
Figure 2: Example of data from one participant that contains calls (dots), stress ratings
(darker areas represent more stressful calls), and break times (squares)..............55
Figure 3: Interface of the multimodal annotation tool...................................................57
Figure 4: A system for emotional reflection........................................................................58
Figure 5: Schematics of the wearable sensor and the digital mirror interface. The data is
transferred between the two systems via the cloud .............................................. 58
Figure 6: (a) Timeline and (b) mosaic visualizations linking physiological and visual
context........................................................................................................................59
Figure 7: Campus map with different locations color coded in terms of smiles (top) and
smile values per location (bottom) .............................................................................61
Figure 8: Example of live-feed as people interacted with one of the installations. Green
blobs indicate faces detected as smiling, while yellow blobs indicate non-smiling
faces ..........................................................................................................................................................61
Figure 9: (Left) students printed several smiling faces and hung them to increase the
smile-barometer, and (right) the software detected the face of a dog ......................62
Figure 10: Average and standard deviations of intensity of smiles (left) and average
number of interactions with the system (right) during the week ...............................63
Figure 11: Google Glass and other wearable devices can be combined to enhance the
management of emotions when most needed .............................................................64
Figure 12: Prototype of a Remote Controller car that reflects the physiological readings
of the driver..................................................................................................................68
Figure 13: Real-time visualization of physiological data on the dashboard of a vehicle..68
Figure 14: Stress-free-routes to prevent future stress.........................................................68
Figure 15: Notification displayed on the Google Glass, the Gear Live and the Galaxy S4
when battery levels were under 10% ...........................................................................76
Figure 16: Summary of surveys and questions for the different phases of the study ......79
Figure 17: We developed a novel ESM application that can be deployed on smartphones, wrist and head-worn devices. The application presents several prompts containing 2D-Grid and 5-point Likert scale questions.

Figure 18: Prompts could be postponed for 5 minutes by pressing the volume down button of the phone (left), the side button of the watch (middle), and by swiping two fingers down on the side of the glasses (right).

Figure 19: Example of daily scheduled prompts for the three devices of one of the participants.

Figure 20: (Left) Average number of prompts received for each participant across devices, and (right) percentage of these prompts that were answered. Green lines indicate standard error.

Figure 21: (Left) Average total response time for the prompts with each device, (middle) average time between the triggering of the prompt and the first user interaction, and (right) average time between the first user interaction and the submission of the final answer.

Figure 22: (Left) Average total response time for the prompts with each device when considering the first two days versus the last two days of the study.

Figure 23: Average range of responses for the 2D-Grid (left) and 5-Likert scale questions (right) across participants for each of the devices.

Figure 24: Boxplots for each of the responses across devices. The boxplot shows the maximum, minimum, median and 75th percentiles.

Figure 25: Distribution of responses for several qualitative aspects of the phone (top row), Watch (middle row) and Glass (bottom row). Red lines indicate mean responses and green-slashed lines indicate standard error.

Figure 26: Distribution of responses for several qualitative aspects of the wearable camera. Red lines indicate mean responses and green-slashed lines indicate standard error.

Figure 27: Head-mounted wearable device Google Glass and some of the sensors.

Figure 28: Example of an estimated pulse wave from Glass gyroscope data (red) and the ground truth blood volume pulse signal (blue). Bottom graphs show the Fourier
Spectrum of each signal. (FFT: Fourier Spectrum, GYR: Gyroscope, BVP: Blood Volume Pulse, HR: Heart Rate, bpms: beats per minute) ......................................114

Figure 29: Example of an estimated respiratory wave from Glass accelerometer data (red) and the ground truth respiration signal (blue). Bottom graphs show the Fourier Spectrum of each signal. (FFT: Fourier Spectrum, ACL: accelerometer, RESP: Respiration from chest band, RR: breathing rate, bpms: breaths per minute).........115

Figure 30: Bland-Altman plots for heart (top) and breathing rates (bottom) using gyroscope (left), accelerometer (center), and camera (right). Each graph shows the agreement of 648 pairs of measurements. Data from different participants are represented with dots of different colors. Mean error is depicted with slashed red and 95% limits are depicted with dashed green lines. (HR: Heart Rate, RR: Breathing Rate, GYR: Gyroscope, ACL: Accelerometer, RGB: Camera).............................. 118

Figure 31: Mean absolute error for heart (beats per minute) and breathing rates (breaths per minute) when considering different window sizes ranging from 5 to 60 seconds. (ACL: Accelerometer, GYR: Gyroscope, RGB: Camera)...................................... 121

Figure 32: We present a novel approach for measuring cardiac and respiratory parameters from wrist motions using a smartwatch (Samsung Galaxy Gear), even when the wrist is not held against the chest ................................................................. 123

Figure 33: Setup for the data collection during the sleep experiment. ECG electrodes were attached to the skin near the heart and on the abdomen. The Galaxy Gear was worn on the left wrist................................................................. 124

Figure 34: Example of pulse and respiratory wave estimation from both accelerometer and gyroscope readings on the wrist of a participant while standing up..............126

Figure 35: Bland-Altman plots for heart rate (left) and breathing rate (right) using the best combination of motion sensors for each case. The graphs show the agreement of 594 pairs (left) and 648 pairs (right) of measurements. Data from different participants are represented with dots of different colors. Mean error is depicted with dashed red and 95% limits are depicted with dashed green lines. (HR: Heart Rate, BR: Breathing Rate)........................................................................................................ 129

Figure 36: Mean absolute error for heart rate (top) and breathing rate estimation (bottom) when considering different sampling rates, ranging from 5 to 50 Hz ...................... 131
Figure 37: Representative example of heart rate estimation with motion-based sensors on the wrist (dashed-green) and with chest ECG electrodes (blue) during one night of sleep. Raw accelerometer (middle) and gyroscope (bottom) readings. Red areas in the top graph indicate regions with “excessive” accelerometer motion.

Figure 38: (Left) Percentage of excluded segments for different artifact thresholds. (Right) Mean absolute error (beats per minute) for different combinations of sensors.

Figure 39: Motion sensors of a smartphone can be used to recover heart and breathing rates of the users during stationary positions and activities such as listening on the phone.

Figure 40: Location of the smartphone during each part of the experimental protocol. Participants remained still while standing up, sitting down, and lying down, which were repeated before and after exercising, followed by watching a video, listening on the phone, and browsing the Internet while sitting down.

Figure 41: Bland-Altman plots for heart (left) and breathing rates (right) of the conditions that yielded the best mean absolute error when the phone was inside the pocket (top), inside the bag (middle), and on the hand (bottom). Mean error is depicted with slashed red and 95% limits are depicted with slashed green lines. (HR: Heart Rate in beats per minute, BR: Breathing Rate in breaths per minute, Accel: Accelerometer).

Figure 42: Distribution of motion levels observed in the laboratory experiments for the pocket (top), wrist (middle), and head (bottom) accelerometer sensors in which only ballistocardiographic signals were observed (N: 648 per histogram).

Figure 43: Average mean absolute error (blue line) and standard error (green lines) across all participants (dot black lines) for the pocket (left), wrist (center), and head (right) locations. The total number of considered segments for these graphs were 227K, 205K, and 203K, respectively.

Figure 44: Distribution of motion levels from motion sensors in the pocket (left), wrist (center), and head (right) when considering the whole dataset. The total number of considered hours were 488, 445 and 441, respectively.
Figure 45: Temporal distribution of potential assessment from motion sensors in the pocket (left), wrist (center), and head (right) locations, when considering motion levels that yielded a mean absolute error of $\leq 5$ beats per minute. Each row represents a different participant, grey rectangles indicate different days, blue areas indicate aggregated distribution, and grey dots indicate separate heart rate assessments. The results were generated from 1374 hours of accelerometer data.

Figure 46: Distribution of stress levels by each of the self-reported affective states (left) and work conditions (right) during the study. Top graphs show the original stress self-reports, and bottom graphs show the derived stress self-reports. Each dot represents a separate self-report and its color indicates the stress level, ranging from dark red indicating the highest stress level (5) to dark blue indicating the lowest stress level (1). Black circles indicate the average response. ($N = 512$ for each of the graphs).

Figure 47: Average self-reported stress levels for different days of the week (Monday to Friday from left to right, respectively).

Figure 48: Example of Glass accelerometer data and estimated activity level.

Figure 49: Summary of features extracted from the wearable signals.

Figure 50: Average percentage of missing samples and standard errors (green bar) across all participants for each of the signals.

Figure 51: Average classification scores (%) and standard errors (green bar) across all the participants for each of the signals considered individually. (HR: Heart Rate, HRV: Heart Rate Variability, BR: Breathing Rate, TEMP: Skin Temperature, Hum.: Humidity, Temp.: Ambient Temperature, Press.: Atmospheric Pressure).

Figure 52: Representative examples of EDA responses from two different participants and their derived self-reported stress levels.

Figure 53: Disclosure signage used to inform other people about the cameras and the study.
List of Tables

Table 1: Main stress measurement methods and some of their main advantages and challenges .................................................................................................................. 38
Table 2: List of wearable devices each participant carried during the study .......... 74
Table 3: Questions asked for each of the prompts and definitions of the used concepts ..81
Table 4: Summary of main interaction differences across devices ......................... 90
Table 5: Heart rate estimation with Glass ....................................................................... 117
Table 6: Breathing rate estimation with Glass.................................................................117
Table 7: Mean absolute error of heart rate with Glass (N = 216 per column) .......... 118
Table 8: Mean absolute error of breathing rate with Glass (N = 216 per column) ...... 118
Table 9: Heart rate estimation with Gear .........................................................................127
Table 10: Breathing rate estimation with Gear................................................................ 127
Table 11: Mean absolute error of heart rate with Gear and E3 (N = 216 per column) ... 127
Table 12: Mean absolute error of breathing rate with Gear (N = 216 per column) ........ 127
Table 13: Heart rate estimation from Gear and E3 during sleep (N = 95,510) ............ 132
Table 14: Heart rate estimation with the Phone .............................................................. 139
Table 15: Breathing rate estimation with the Phone......................................................... 139
Table 16: Summary of laboratory findings for heart rate estimations ......................... 149
Table 17: Overview of the different signals and features used in the context of stress recognition (R: Reported)........................................................................................................ 163
Table 18: Peak detection parameters used for each of the wearable signals with the MATLAB FINDPEAK function .............................................................. 165
Table 19: Missing data (%) for each of the participants and wearable signals. Values over 25% are highlighted in red bold................................................................. 171
Table 20: Classification scores (%) for each of the participants and wearable signals. Scores over 65% are highlighted in green bold. (HR: Heart Rate, HRV: Heart Rate Variability, BR: Breathing Rate, TEMP: Skin Temperature, Hum.: Humidity, Temp.: Ambient Temperature, Press.: Atmospheric Pressure) ......................... 171
Table 21: Regression coefficients when deriving new self-report stress levels........... 205
List of Abbreviations

ACL: Accelerometer
BCG: Ballistocardiography
BR: Breathing Rate
BVP: Blood Volume Pulse
ECG: Electrocardiography
EDA: Electrodermal Activity, Galvanic Skin Response
ESM: Experience Sampling Method, Ecological Momentary Assessment
FFT: Fast Fourier Transform
Gear: wrist-worn wearable device, smartwatch, Samsung Gear Live
Glass: head-mounted wearable device, Google Glass
GYR: Gyroscope
HR: Heart Rate
HRV: Heart Rate Variability
Hum: Humidity
Narrative: wearable camera, Narrative Clip
Phone: smartphone, Samsung Galaxy 4
PNS: Parasympathetic Nervous System
PPG: Photoplethysmography
Press: Atmospheric Pressure
RESP: Respiration
RGB: Camera
SNS: Sympathetic Nervous System
SVMs: Support Vector Machines
TEMP: Skin Temperature
Temp: Ambient Temperature
Watch: wrist-worn wearable device, smartwatch, Samsung Gear Live
Chapter 1

Introduction

1.1 Motivation
Do you remember the last time you felt genuinely stressed at work? Maybe you had a pressing deadline and very little time to write a report or perhaps you received an unpleasant e-mail you had to reply to. Although you might not have been completely aware about feeling stressed, your body might have been experiencing a sequence of physiological changes that may have induced pupil dilation, deeper breathing, intensified beating of the heart, and increased muscle tension, among many other changes. This chain of physiological changes and their associated behavioral effects, plays a very important role in our daily functioning. Stress not only regulates important processes such as attention and memory acquisition but also helps us tune the body to face daily challenges and threats. However, repeated triggering of this response during daily activity can result in chronic stress, contributing to a wide array of pathophysiological risks, including cardiovascular disease, cerebrovascular disease, diabetes, and immune dysregulation (Cacioppo et al., 2000; McEwen & Seeman, 2003; Armaiz-Pena et al., 2009).

While our daily lives contains a wide gamut of stressors (e.g., social, physical, environmental), this thesis focuses on advancing the measurement of negative stress experienced at the workplace. This type of stress, also known as occupational stress, can be defined as the harmful emotional and physical responses that occur when high demanding job conditions cannot be met by the resources of the worker. Indeed, the relationship between demands and resources plays a very important role in terms of stress
appraisal and health outcomes (more details in Chapter 2). Instead of focusing on long-term and chronic stress, this work studies short-term and acute stress changes that can be triggered in a matter of seconds. This type of stress is usually associated with feelings of frustration, anger and fear and can lead to dissatisfaction and lack of motivation in the long-term (Canadian Centre for Occupational Health and Safety'). Furthermore, high levels of stress can impair decision making, decrease productivity, and lead to high amounts of accidents and job absenteeism (The American Institute of Stress2). As a result, several surveys have similarly reported that stress can significantly increase medical and insurance costs (up to 50% based on Johnston et al., 2009). Moreover, it has been estimated that that the business costs associated with workplace stress are around $300 billion per year in the U.S. alone (Rosch, 2001). Despite the well-studied negative outcomes, workplace stress is still considered a necessary evil by many people as it helps us keep up with the pace of modern society.

An important step in managing stress, before it becomes chronic, is recognizing precisely when and where it occurs. Technologies that automatically recognize stress can be extremely powerful, both diagnostically and therapeutically. As a diagnostic tool, technologies such as these could help individuals and clinicians gain insight into the conditions that provoke maladaptive stress responses and study how they change over time. If a person could know, for instance, that during the last week s/he experienced more stress than usual, the person could gain more awareness and incorporate behavioral changes to reduce unnecessary stressors (e.g., increase breaks or exercise, change the type of work activity, or socialize or sleep more). As a therapeutic tool, these technologies could be used to automatically detect when stress-management interventions are needed, so that problems can be addressed as soon as they are detected, rather than hours or even days later. Moreover, existing technologies could also use this type of information to produce more human-like forms of interaction (Picard, 1997) and automatically tune their behavior to avoid creating additional stress and frustration (e.g., Moraveji & Soesanto, 2012). For instance, if a computer user is feeling stressed, the computer could change the intonation

---

1 http://www.ccohs.ca/oshanswers/psychosocial/stress.html
2 http://www.stress.org/workplace-stress
of notifications to be more empathetic or completely prevent non-urgent ones to avoid adding more cognitive load.

Researchers have studied a wide variety of approaches to measure stress, such as self-reports, hormone analysis, and the measurement of physiological signals. However, each of these approaches has its own set of limitations such as requiring the attention of the person, being obtrusive and/or capturing partial information about stress. Moreover, there is often great variability in how people perceive, experience and physiologically express stress, obstructing efforts to build a one-size-fits-all stress recognition system. Finally, these methods have been traditionally tested in controlled or semi-controlled settings where most of the real-life variables that introduce noise are controlled or eliminated. The research at the core of this thesis combines state-of-the-art wearable devices with machine learning methods to advance the measurement and analysis of stress at work. As wearable devices are designed to be in close contact with the user during daily life, we expect to have unique opportunities to develop less-intrusive and personalized methods that enhance the measurement of stress.

1.2 Areas of Work

For over a century, psychologists and behavioral scientists have thoroughly studied stress and other emotions in laboratory settings. However, these emotions may be partly contrived – elicited by controlled “lab” stimuli, which may not matter to the participants in the studies, at least not compared with real life stressful events. With the recent improvement of wearable biosensors, researchers have started to study more natural and spontaneous emotions and their role in real-life interactions, creating a new set of complex challenges that need to be addressed. This thesis explores how wearable devices and their different form-factors (e.g., body location, sensors) can help address a subset of these challenges and move us one step closer to real-life wearable stress measurement. In particular, the areas of focus are the following:

Self-report data collection. Self-reports are considered the gold standard measure of stressful experience. To effectively gather this type of information while minimally disrupting daily activity and incorporating cognitive and memory biases, this work creates
an Experience Sampling tool that can be deployed on different types of wearable devices. Furthermore, we use the application to systematically quantify how different wearable form-factors such as body location can play an important role in the reporting process.

**Physiological sensing.** The stress response is a well-studied set of physiological changes that can be monitored for the purpose of stress measurement. However, to accurately gather this type of information in the midst of daily activity, cumbersome electronics such as sticky electrodes on the chest are usually required. This thesis leverages some of the most recent commercially available wearable devices and pushes the boundaries of existing sensors to improve the measurement of physiological stress parameters.

**Stress recognition from wearable signals.** Many of the studies focusing on stress measurement mainly relied on a single sensor modality in controlled laboratory settings. This thesis considers several physiological, behavioral and contextual signals captured by different types of wearables, and systematically evaluates them in a real-life workplace scenario.

### 1.3 Thesis Aims

The specific research aims of the thesis are the following:

- Development of a multi-platform Experience Sampling tool that can work on different types of wearable devices (head-worn, wrist-worn and inside the pocket).
- Evaluation of how different wearable form-factors can impact the self-report of stress levels in real-life.
- Development of novel methods to unobtrusively measure cardiac and respiratory physiological parameters from wearable motion-based sensors.
- Evaluation of the validity of these novel methods in both controlled and uncontrolled scenarios.
- Assessment of generalization of the previous results in a longitudinal real-life workplace setting.
• Development of software algorithms to automatically quantify the quality of the physiological readings based on motion data and other physiological readings.
• Assessment of the recognition value of different types of wearable signals (physiological, behavioral and contextual) in a real-life workplace scenario.

1.4 Dissertation Outline

The outline of the remainder of the thesis is as follows:

Chapter 2 overviews relevant research in the context of stress measurement and some of its main challenges. In particular, the chapter describes: 1) some of the commonly used workplace stress models, 2) physiological changes associated with the stress response, 3) several approaches to measuring stress, 4) current approaches for ambulatory self-report collection, 5) methods to measure daily context, 6) different approaches to comfortably measure physiological signals, 7) previous studies considering physiological signals to automatically infer stress levels, and 8) relevant work on stress management using technology.

Chapter 3 provides a brief overview of early exploratory research in stress measurement, context gathering, and stress management that have helped shape and motivate this work. In particular, we describe how computer peripherals (e.g., pressure sensitive keyboard and capacitive mouse), and wearable electrodermal sensors can be used to capture different aspects of stress in controlled and real-life settings, respectively. Then, we describe a mobile application for quick annotation of behavioral activity and a system that enables the capture and visualization of visual context and physiological information during daily life. Finally, we overview different approaches to help manage stress with technology by 1) measuring and promoting smiles, 2) exploring the unique form-factors of head-mounted wearable devices, and 3) designing context-specific stress management interventions in a driving scenario.

Chapter 4 provides the details of a real-life workplace data collection which enabled the study of the different areas of work of this thesis. In particular, the chapter reviews the
following: 1) selection of wearable devices, 2) experimental protocol of the study, 3) surveys and experience sampling questions, 4) privacy issues that needed to be considered for participants during our study, and 5) recruitment of participants and their compensation. Finally, the chapter provides a preliminary overview of the collected data.

**Chapter 5** analyzes the data described in Chapter 4 to assess the pros and cons of using different form-factors of wearable devices in the context of ambulatory self-report measurements. To do so, we designed and developed a custom Experience Sampling application and used it to deliver prompts on head and wrist-worn devices, and compared them with the traditional phone in the pocket. The chapter includes a quantitative and qualitative analysis of relevant aspects during the reporting process.

**Chapter 6** describes novel methods to extract cardiac and respiratory parameters from wearable motion sensors in mostly controlled laboratory studies. In particular, we systematically compare the performance of these methods considering different types of sensors (e.g., accelerometers, gyroscopes) and their combinations, when placed on three different body locations (e.g., head, wrist, pocket). Moreover, we also explore some of the unique advantages of each of the considered locations. For instance, we used a head-mounted camera to also capture physiological signals, and the wrist sensor to comfortably monitor physiological parameters while sleeping at home.

**Chapter 7** leverages some of the methods proposed in Chapter 6 to estimate physiological parameters in the real-life workplace dataset described in Chapter 4. In particular, we assess the performance of the heart rate estimation methods when using wearable motion sensors on the head, wrist and pocket, and quantify the distribution of accurate assessments in our study.

**Chapter 8** develops supervised machine learning methods to apply toward the automatic inference of self-reported stress levels from the wearable data captured during our real-life workplace study (Chapter 4). In particular, we start by studying the distribution of self-reported stress levels as well as their relationship with other affective and work-related
questions. Then, we develop different methods to extract and process different types of wearable signals (physiological, behavioral and contextual), and use them in the context of automated stress recognition.

Lastly, Chapter 9 concludes the dissertation by discussing some of the main findings and their limitations, as well as identifying areas of potential future work for each of the different topic areas of this thesis (real-life self-report data collection, comfortable physiological measurement, and stress recognition from wearable data). Finally, we provide some concluding remarks.
Chapter 2

Background Research

This chapter provides an overview of relevant research in the context of stress measurement and some of its main challenges. In particular, the chapter describes: 1) some of the frequently used workplace stress models, 2) physiological changes associated with the stress response, 3) several approaches to measure stress, 4) current approaches for ambulatory self-report data collection, 5) methods to capture daily context, 6) different approaches to comfortably measure physiological signals, 7) previous studies using physiological signals to automatically infer stress levels, and 8) relevant work on stress management using technology.

2.1 Stress Models

One of the main challenges when defining and measuring stress is the great variability in how people perceive and experience stress. Writing a thesis, for instance, may be very relaxing and enriching for some people but may be very stressful for others. Moreover, this same activity may elicit different levels of stress in the same person from one day to another, perhaps depending on factors such as the amount of sleep s/he had during the previous day (Pärkkä, et al., 2009), whether s/he exercises regularly (McEwen & Seeman, 2003), and the amount and type of social interactions s/he had during the day (Mark et al., 2014). All of these factors complicate the task of modeling stress.

Over the years, researchers have devised different models to explain how, when and why stress is experienced at work. A model postulated by French (1973) is the Person-Environment Fit (PEF). This model focuses on the relationship of the person with the work
environment and how well the person’s needs fit with what the person receives. In particular, negative physical and psychological stress (a.k.a., job strain) will appear whenever one of the following incongruent situations appears: the person’s abilities do not match the demands of the job, or the person’s goals or aspirations do not match the resources offered by the work environment. Although this theory tackles the key aspect of considering individual characteristics (e.g., goals, abilities), it is only focused on subjective perceptions; in practice, it may be relatively difficult to quantify the level of fitness between the person and the environment. Furthermore, the theory may be too limited to reflect the nature of all real-life situations. For instance, a person’s goal may be to socialize with many people, but supplying that need may lead to an unproductive work-environment that elicits more stress in the long-term.

An alternative model that has been widely studied in the literature is the Demand-Control model (DC) devised by Karasek & Theorell (1990). This model postulates that the amount of job strain people experience will be determined by whether or not they have control over the demands of the work. Control (also called decision latitude) reflects the amount of decision authority (e.g., control over work situation) and skill discretion (e.g., capacity of using learned skills). While the model differentiates between low/high control and low/high demand conditions, the combination of low control and high demands is thought to be the condition that elicits the most psychological strain. This definition of job strain has been strongly associated with increased levels of epinephrine and cortisol in the blood stream and increased blood pressure which, in turn, contribute to a large number of adverse health consequences (e.g., stroke, hypertensive heart disease, arteriosclerotic cerebrovascular disease, depression, fatigue). While the PEF model focuses on individual perceptions, the Demand-Control model focuses on the structural features of individual’s interactions with the work environment (i.e., control, demands). Few years later, Johnson & Hall (1988) extended the DC model by incorporating social support as another important dimension as there is substantial research suggesting that it may buffer the impact of stressors. Although this model has been widely used, there are some major limitations. For instance, the model does not account for individual differences in susceptibility of stressors (e.g., same level of control and demands for different people may elicit different levels of job strain). Also, the model assumes that high control is always desirable but there are
situations where less control may be less stressful (e.g., when a person has a low perception of self-efficacy). Overall, the model may be adequate to gain an initial idea of how healthy a workplace or how stressed a population may be at a macro level, but may not be adequate to capture the stress process for all individuals and job situations.

A few years later, Siegrist (1996) devised the Effort-Reward Imbalance (ERI) model. Shifting the attention from control to reward, this theory postulates that job strain will appear if there is a high level of effort but a low level of reward. This incongruence violates core expectations about reciprocity and appropriate exchange in social situations. Rewards can come in many forms such as money, esteem, and job status control (e.g., job insecurity, promotion prospects). Efforts, on the other hand, can be either extrinsic (similar to job demands in the DC model) or intrinsic (e.g., high need for control, over commitment). It is the second type of effort that allows this model to capture some of the individual differences associated with stress. The long-term combination of high-effort and low-reward has been linked to increased systolic blood pressure and has been shown to be associated with elevated risks for heart disease and increased chances of suffering emotional exhaustion (Siegrist, 1996).

While the previous models describe a relevant subset of stress factors (e.g., level of fitness between person and environment, control, demands, rewards), they cannot easily incorporate new factors or dynamically adjust the importance of existing factors. To partially address these limitations Bakker, Demerouti et al. (2001 & 2007) proposed the Job Demands-Resources (JD-R) model which aggregates relevant factors into two categories: job resources and job demands. In particular, they define demands as “those physical, psychological, social or organizational aspects of the job that require sustained physical and/or psychological (cognitive and emotional) effort or skills and are therefore associated with certain physiological and/or psychological costs.” Examples of high demands include but are not limited to high work pressure, unfavorable physical environment, and emotionally demanding interactions with other people. Resources are defined as “those physical, psychological, social, or organizational aspects of the job that are either/or: functional in achieving work goals, reduce job demands and the associated physiological and psychological costs, stimulate personal growth, learning, and development.” Examples of different forms of resources include but are not limited to
organizational support, performance feedback, adequate materials (e.g., money), job autonomy, and positive organizational climate. Considering the previous two categories, the authors state that motivation is positively correlated with the amount of resources, and that the highest level of job strain will occur when there are high demands and low resources (i.e., low motivation).

Finally, while not specifically a workplace stress model, the Biopsychosocial model of challenge and threat (BPS) by Blascovich & Tomaka (1996, see also Blascovich & Mendes, 2000) takes the JD-R one step further and states that during a stressful scenario the ratio between resources and demands can elicit two very differentiated stress responses. In particular, when there are sufficient resources to meet the demands, an individual experiences the situation as challenging, and when there are insufficient resources to meet the demands, an individual experiences the situation as threatening. These two situations have been shown to lead to well differentiated physiological responses. On the one hand, challenging responses are associated with increased cardiac sympathetic activity (e.g., increase of heart rate, shorter pre-ejection period) and decreased or unchanged total peripheral resistance, leading to greater blood flow to the periphery of the body facilitating energy mobilization and performance (similar to performing aerobic exercise). On the other hand, threatening responses are associated with increased cardiac sympathetic activity and increased total peripheral resistance, resulting in less blood reaching the periphery to prepare the body for certain behaviors such as freezing and avoidance (Mendes et al., 2007; Seery, 2013; Ramsey, 2014). Among the two cases, challenge is more usually associated with positive outcomes such as more productivity and engagement (e.g., Blascovich et al., 1999; Dienstbier, 1989; Jamieson et al., 2010), and threat is more usually associated with negative outcomes such as impaired decision-making and cardiovascular diseases (Jefferson et al., 2010; Matthews et al., 1997; Schnall et al., 1994). The evaluation of resources and demands changes from person to person and is continuously influenced by the changing circumstances (Quigley et al., 2002). Moreover, people can also be trained to differently appraise stressful situations and avoid potential negative outcomes (Jamieson et al., 2012).

While each of the models can be useful to capture different aspects of stress, this thesis borrows the definitions of demands and resources of the JD-R model and studies their
relationship in the context of the BPS model to help quantify the intensity of stress responses.

2.2 Fight or Flight Response

In a stressful situation, the body experiences a series of physiological events driven by the two branches of the Autonomic Nervous System. While the Sympathetic Nervous Systems (SNS) mobilizes the body’s resources in response to a challenge or a threat (e.g., quickens the pulse, deepens the respiration and tenses the muscles), the Parasympathetic Nervous System (PNS) works antagonistically to control this process. Furthermore, processes that are non-critical for immediate survival such as the digestive system, reproductive and sexual activation, and urine production slow down or stop. During this process, people may experience enhanced arousal, improved cognitive functioning and concentration, increased ability to withstand pain, and accelerated motor reflexes, preparing the body to face life-threatening situations. Both SNS and PNS act in coordinated fashion throughout the body achieving homeostasis, but their uncoordinated activation (e.g., excessive SNS, too low PNS) for extended periods of time may lead to a condition in which the body fails to appropriately handle stress, also known as allostatic load (McEwen & Seeman, 2003). In this condition, the body may fail to appropriately engage the stress response in the presence of relevant stressors. For instance, the body may unnecessarily trigger the fight-or-flight responses (e.g., post-traumatic stress disorder), and/or become unable to shut the stress response off after the stressful stimulus has passed. In the long-term, these problems can lead to serious health related problems such as cardiovascular and cerebrovascular diseases, diabetes, impaired fertility, obesity, and immune dysregulation (Cacioppo et al., 2000; McEwen & Seeman, 2003; Sapolsky, 2007; Armaiz-Pena et al., 2009).

2.3 Stress Measurement Approaches

Measuring stress at work has been the focus of interest of psychological and psychophysiological researchers for many decades. Being able to automatically quantify stress during daily life could help people not only to better understand what events elicited the highest stress levels during their daily activity but also to prevent the negative outcomes associated with chronic stress.
From the point of view of psychology, the gold standard of stress measurement is self-report measures that can be collected through retrospective surveys and/or experience sampling (a.k.a., ecological momentary assessments). The psychologists consider that the main advantage of this approach is that it is subjective and, therefore, it can potentially capture the individual personal experience of stress. There are a wide array of surveys to quantify different types of stressors and their frequency during daily life. For instance, the Daily Stress Inventory (Brantley et al., 1987) allows one to quantify the daily stress by counting the number and type of stressors that appeared throughout the day. While retrospective surveys enable capturing very detailed information, they are negatively affected by recall problems. On the other hand, experience sampling methods minimize these problems but severely limit the amount of information that can be captured (e.g., context during a stressful event). A common limitation for both retrospective surveys and experience sampling methods is that they require the full attention of the person, which can be disruptive when considering very frequent (ideally continuous) real-life measurements. Moreover, they assume that the monitored person can appropriately identify and express their own emotions, and that they are willing to do so. These assumptions are not always accurate in real-life (e.g., people with emotional impairments, excessive work overload that negatively impacts accurate self-reflection, or discomfort reporting some experiences).

The main alternative to self-report measures are objective measures. For instance, information about stress responses can also be obtained by analyzing hormones such as cortisol or adrenaline that can be gathered from saliva and blood samples. However, these measures commonly entail costly and slow analysis. An alternative approach consists of monitoring behaviors that are influenced by stress. For instance, Zimmermann et al. (2003) proposed monitoring the use of computer mouse and keyboard to capture changes associated with the affective states of users. However, this approach captures indirect information of stress that is only available when the user is performing a specific behavior at the instrumented location. Finally, an alternative approach that addresses the previous limitations consists of monitoring the physiological responses associated with stress, such as heart rate, blood volume pulse, skin temperature, pupil dilation or electrodermal activity. Although most biosensors are large and bulky, and impractical for long-term use in real
life situations, the recent introduction of more comfortable and wearable devices (e.g., Poh et al., 2010) creates new opportunities for continuously monitoring these signals. This physiologically-based approach is more commonly used in psychophysiology in which stress has been frequently defined as a physiological reaction of the body when a person has to meet certain challenges, and the body moves away from a state of homeostasis. Despite not requiring attention of the person and being able to provide continuous measurements, this method alone cannot capture the large individual differences associated with the stress response, especially when considering the large variability of real-life scenarios. More importantly, similar subjective stress reports may elicit different physiological patterns such as those of challenging and threatening conditions discussed in section 2.1. Thus, a combination of approaches that consider both the person as well as other types of contextual information is preferred.

The work presented in this thesis considers self-reports as the ground truth measurement of experienced stress and explores using supervised learning methods to automatically infer self-report measures from different types of wearable data. To capture self-reports, we use a mixed approach that combines experience sampling to gather affective and work-related information during the day, and retrospective surveys to gather additional contextual information at the end of each day. Moreover, we use wearable cameras to help minimize recall biases associated with the latter approach. To capture wearable data, we considered seven wearable devices and extracted different types of signals from each device. In particular, we extracted physiological (e.g., heart rate variability, electrodermal activity, respiration), behavioral (activity levels from different body locations), and contextual (e.g., amount of light, atmospheric pressure) information that can be comfortably and automatically gathered with existing wearable devices. By combining different approaches, we hope to address some of the major challenges of stress measurement (e.g., disruptiveness of self-reports) while preserving the main benefits of each approach (personalized and continuous measurements).

Table 1 highlights some of the main advantages and disadvantages for each of the discussed methods.
Table 1: Main stress measurement methods and some of their main advantages and challenges

<table>
<thead>
<tr>
<th>Method</th>
<th>Self-reports</th>
<th>Hormones</th>
<th>Behaviors</th>
<th>Physiology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples</strong></td>
<td>Daily Stress Inventory</td>
<td>Cortisol</td>
<td>Smoking</td>
<td>Heart Rate Variability</td>
</tr>
<tr>
<td></td>
<td>Perceived Stress Scale</td>
<td>Adrenaline</td>
<td>Keyboard</td>
<td>Electrodermal Activity</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>Subjective experience</td>
<td>No cognitive attention</td>
<td>Minimally invasive</td>
<td>Continuous</td>
</tr>
<tr>
<td></td>
<td>Captures complexity</td>
<td>Difficult to manipulate</td>
<td>Unobtrusive</td>
<td></td>
</tr>
<tr>
<td><strong>Challenges</strong></td>
<td>Cognitive attention</td>
<td>Intrusive</td>
<td>Indirect</td>
<td>Instrumentation</td>
</tr>
<tr>
<td></td>
<td>Recall biases</td>
<td>Costly and slow</td>
<td>Sporadic</td>
<td>Maintenance</td>
</tr>
</tbody>
</table>

2.4 Ambulatory Self-reports

The Experience Sampling Method (ESM) (Larson & Csikszentmihalyi, 1983), also known as Ecological Momentary Assessment (EMA) (Shiffman et al., 2008; Bolger et al., 2003), is a methodology to gather repeated subjective reports from people during their daily activities. While some uses of ESM date back almost 40 years (Csikszentmihalyi, Larson & Prescott, 1977), continuous advancements in technology have enhanced the approach and contributed to its common usage in real-life studies. For instance, ESM has been successfully used to better understand real-life emotions, social interactions, time productivity, and personality (Conner et al., 2009).

In comparison with traditional annotation techniques such as surveys and diaries, ESM relies on asking questions frequently during daily life to minimize retrospective recall biases that may appear. ESM also minimizes the amount of interaction required with researchers during the data collection which helps reduce the response biases resulting from being observed.

Over the years, researchers have explored using a wide variety of devices to enhance the ESM experience. Some of the most commonly used devices include pagers (Csikszentmihalyi, Larson & Prescott, 1977), Personal Digital Assistants (PDAs), paper booklets, audio recordings, and cameras (Consolvo & Walker, 2003). More recently, thanks to the massive adoption of cellphones, researchers have extensively used phones as a way to trigger, collect and log the responses of participants. Some of the main benefits...
of using cellphones are that people are already used to carrying them, and receiving and sending information with them. Moreover, these devices also offer many opportunities to collect a wide range of information (e.g., social interactions, motion data) that can be used to alter the ESM process in meaningful ways (e.g., only collect information when the person is not moving). However, current approaches of ESM still suffer from some major limitations.

The whole process of receiving a prompt during daily activity is arguably disruptive. As soon as a prompt is triggered, the person needs to reach for the device, provide the requested information, and return it to the original location. If the person is carrying the device inside the pocket (e.g., a phone), this process can significantly bias some of the responses. For instance, if the person is sitting down, it is very likely s/he will need to change the body posture to reach and use the device. This process could not only affect answers to specific questions (e.g., “What is your frustration right now?”) but also may significantly alter physiological readings that are used to better understand emotions (e.g., heart rates are usually higher when standing up). The recommended response time to a specific prompt is less than two minutes (Consolvo & Walker, 2003). Therefore, it is very important to ensure the participants do not spend unnecessary time using the devices.

A fundamental challenge of ESM tools is to effectively deliver the prompts to ensure a high response rate. While there are well-established notification techniques such as sounds and vibrations, it is still difficult to make them noticeable without becoming disruptive. This is especially the case during the busy daily activities in which environmental noises or large body motions may obscure even the most disruptive kind of notifications. As a result, relevant and expensive information may be dismissed and the content of the assessments may be biased towards specific moments in time (e.g., when there is little noise and body motion) significantly affecting the potential generalization of findings.

Finally, one of the main design principles of ESM tools is to minimally disrupt daily activity while collecting as much information as possible. In turn, the length and number of prompts should be as short and infrequent as possible to minimize the burden to participants. To achieve that, researchers usually minimize the response time as much as possible so participants are eager to provide more frequent ratings and leave lengthier questions to other moments in time when participants are expected to be more available.
However, successfully achieving these goals can be quite challenging as postponing questions can easily incorporate well-studied recall biases (e.g., people forget things, report false memories) (Shiffman et al., 2008).

This thesis develops an ESM tool that can be used on different types of wearable devices, and is used to quantify how different form-factors can affect the reporting process. In particular, we compare the traditional phone inside the pocket versus a wrist-worn device and a head-mounted device. Both wrist and head-worn devices are usually more accessible and offer a unique opportunity to minimize the disruption associated with ESM. However, new interaction challenges arise due to the novelty and unfamiliarity of these devices and their interfaces. We also study how different devices can significantly impact the number of missed notifications and response times in a real-life work setting. Finally, while we postpone lengthier questions to other moments of the day, we employ a wearable body camera to capture daily activities and minimize the recall lag.

2.5 Context Gathering

One of the major difficulties of field studies is the occurrence of unexpected and uncontrolled events that may alter or influence a person’s affective state and corresponding physiology. For instance, when monitoring physiological signals such as electrodermal activity or heart rate to monitor emotional states, there are many stimuli that may result in similar autonomic responses. For instance, physical and mental activities may increase these signals. To remove the effects of changes in physical states and better understand other confounding variables (Cacioppo & Tassinary, 1990), scientists also need to unobtrusively capture other sources of information such as context. For instance, knowing that the participant is giving a talk or running on a treadmill may help differentiate between the two activities.

Continuously monitoring a person's contextual activity is closely related to the concept of life-logging, where a person utilizes passive capture devices to record and digitize his life. The seminal papers of Mann (1997) and Picard & Healey (1997) almost twenty years ago, envisioned the scenario in which a person uses wearable computers to capture fine-grained detail of his or her daily activities and affective states, and uses the information to enhance his or her capabilities. Since these papers were first published, there have been
many approaches to efficiently capture daily-life activities and emotions. However, adequately understanding the social and ecological context has remained a key research challenge. Among all of the contextual modalities, images are probably one of the most effective means to capture information. Visual context not only can convey more information than words but also appears among a list of the preferred representations for autobiographical memories (around 80% according to Brewer, 1988).

A successful device that passively captures images in daily-life is SenseCam (Gemmell et al., 2004). Although this device was originally intended to augment human memory, it has been successfully used for other purposes such as improving the communication of autistic individuals (Marcu, Dey & Kiesler, 2012), and helping people to recollect aspects of earlier forgotten experiences (Hodges et al., 2006). SenseCam also uses other sensor modalities such as light-intensity, temperature sensor, and infrared (body heat) detector to gather additional context. However, when logging various measures of the life of a person, it is necessary to also understand how daily activities affect the inner affective state of the person. Teeters et al. (2006) created a system that continuously monitored the facial expressions of the camera-wearer during her daily activity so the person could better self-reflect on her/his emotional states. However, the system did not gather any additional contextual information, so the person could not accurately identify the factors leading to each facial expression. Moreover, facial expressions can be easily manipulated and may not always be correlated with the internal emotional state of the person. In a separate study, Stähli et al. (2009) monitored physiological activity of participants as well as contextual information in the form of scribbles and cellphone activity (e.g., number of messages). Using the information, they created the Affective Diary to help people better reflect on their bodily responses to daily-life activities. Similarly, McDuff et al. (2012) created AffectAura for the reflection of office employees. In this case, they automatically generated the diary by analyzing several physiological signals and associating them with computer activity logs. While emotional information can potentially be associated with many other types of contextual information, few studies have used images as the context source. One of the few exceptions is the StartleCam project (Healey & Picard, 1998), where researchers monitored the physiological responses of computer users to detect relevant moments (e.g., startle responses) and automatically recorded a video of the events leading towards the
physiological change. While *StartleCam* was ambulatory, it required a cumbersome hand-built computer with wireless module, all of which can be replaced today by consumer hardware.

Motivated by previous work, this thesis explores using more practical and accessible devices such as smartphones and new wearable cameras to continuously capture the moments preceding emotional changes throughout daily activity.

### 2.6 Comfortable Physiological Sensing

A first step towards unobtrusively measuring daily stress consists of the development of tools that can monitor relevant cues of stress without creating additional stress. For instance, the current gold standard approach for measuring heart rate (electrocardiogram) requires sticky gels and uncomfortable electrodes attached to the skin which can be quite cumbersome. Moreover, existing physiological sensors require maintenance (e.g., recharging batteries, replacing electrodes) preventing many people from regularly measuring their vital signs.

Arguably one of the least invasive physiological measurement approaches to measure cardiac and respiratory information is photoplethysmography (PPG), which captures color variations of reflected light reflected from, or transmitted through the skin. Traditional measurements of PPG require a dedicated light source in close contact with the surface of the skin such as the finger (Allen, 2007). However, Verkruysse et al. (2008) demonstrated that it is also possible to gather this type of information using a remote camera (>1m) and ambient light. This idea was further explored by Poh et al. (2010 and 2011) who proposed and validated an automated method to robustly extract physiological parameters using a regular computer webcam. Although these studies represent some of the least invasive approaches currently available, they require having a camera pointed at the person which severely limits their possibilities for daily life monitoring.

An alternative approach that has the potential to enable non-intrusive physiological measurements is ballistocardiography (BCG). This method was pioneered by Starr et al. (1939) who showed that the mechanical ballistic forces of the heart elicit subtle body movements with every heartbeat. While the original experiments required a subject to lie down on a suspended supporting structure to amplify and study the motions, continuous
technological advances have enabled BCG measurement in less constrained settings (Giovangrandi et al., 2011). For instance, researchers have successfully gathered BCG information from daily life objects such as a modified weighing scale (Inan et al., 2009), a chair (Pinheiro et al., 2010) or a back pad (Parak, 2012). Moreover, a wide variety of methods have been explored to gather similar information during sleep, when the amount of motion artifacts are minimized. In this setting, researchers have developed sensors to incorporate BCG and respiratory measurements on the bed post (Brüser et al., 2013) or mattress (e.g., Brink et al., 2006; Shin et al., 2008; Migliorini et al., 2010; Paalasmaa et al., 2012) and have been able to accurately extract vital signs such as heart and breathing rates. Furthermore, this information has been shown to be useful in the detection of circadian rhythms and sleep patterns (Migliorini et al., 2010). However, most of the previous approaches require custom-made hardware devices that require installation and alter the sleeping environment. In addition, most of the sensors assume that there is only one person on the bed and that s/he is on a specific region of the bed where the sensor is located. Finally, none of these devices can be easily used when the person gets up from bed. If the same information could be extracted with a device that the person is already wearing during daily life, such as a smartwatch, then the previous limitations could be addressed. Recent work that partially tackles these problems was presented by Kawamoto et al. (2013), in which they demonstrated that a wrist-worn accelerometer could capture respiratory movements of people sleeping whenever the wrist was near the chest (around 44% of the time).

Researchers have also started to consider more wearable approaches which are more appropriate for daily life monitoring. For instance, Kwon et al. (2011) and Dinh (2011) attached a smartphone to the chest and used its accelerometer to monitor heart rate. Similarly, Phan et al. (2008) proposed a different method to extract both heart and breathing rates. While moving us one step closer to unobtrusive physiological monitoring during daily life, they mostly considered the chest location where both cardiac and respiratory motions are more prominent. One of the few exceptions is the work by He et al. (2010, 2011 and 2012), in which researchers created a custom-made device and demonstrated that heart rate could also be extracted from a peripheral location such as the
ear. In He’s thesis (2013), he also showed preliminary results of breathing rate estimation from accelerometer data for a single sample but no validation study was performed.

This thesis systematically evaluates the possibility of measuring heart and breathing rates from different peripheral locations (head, wrist and pocket) with commercially available wearable sensors such as accelerometers and gyroscopes. Furthermore, we quantify the performance of the proposed methods in both laboratory and real-life uncontrolled settings.

2.7 Stress Recognition from Physiological Signals

While research on automated stress recognition has taken many different forms, the systems that have been proposed in the engineering literature typically contain two principle components: 1) a sensor-based architecture that records relevant features and 2) a software-based system that makes predictions about an individual’s current stress level. The sensing modalities can take many forms, including audio and visual modalities, but biosensors provide the most direct access into the physiological changes that accompany stress.

Several automatic stress recognition techniques have been explored in the research literature. In most cases, data are collected in the laboratory where variables that may introduce noise are partially controlled or eliminated. Traditional methods for inducing stress in such settings involve administering electric shocks (Notarius & Levinson, 1979), cognitive tasks (Poh et al., 2010; Setz, et al., 2010; Barreto, Zhai & Adjouadi, 2007) and psychosocial threats (Setz et al., 2010).

Researchers have explored a variety of classification methods, and techniques to minimize physiological differences across people. Barreto, Zhai & Adjouadi (2007), for example, used Support Vector Machines (SVMs) to discriminate between stressful and non-stressful responses (elicited by different versions of the Stroop Test) (Stroop, 1935) in a laboratory setting. The SVMs outperformed other classification algorithms, obtaining an accuracy of 90.1%. Various physiological signals were used in the classification, including electrodermal activity (EDA), blood volume pulse (BVP), pupil diameter (PD) and skin temperature (TEMP). In a separate study, Setz et al. (2010) used EDA to automatically distinguish between cognitive load and stress elicited by arithmetic computations without
and with time pressure and social-evaluative threat, respectively. In this case, Linear Discriminant Analysis (LDA) obtained 82.8% accuracy, outperforming SVMs. Setz et al. (2010) found that the average number of EDA peaks, as well as the distribution of their amplitudes, were the most relevant features to the problem. To account for participant variability, distributions were computed for each participant independently. In another study, Shi et al. (2010) discriminated between stressful and non-stressful responses under social, cognitive and physical stressors. They obtained 68% precision (a.k.a., positive predictive value) and 80% recall (a.k.a., sensitivity) using SVMs with EDA, electrocardiogram (ECG), respiration (RESP) and TEMP. The problem of participant variability in terms of physiological variance was addressed by subtracting a person-specific parameter from the features of each participant. This parameter was estimated as the average feature of all-non-stressful events of the participant. In an effort to automatically recognize stress in less controlled settings, Healey & Picard (2005) monitored ECG, electromyogram of the trapezius (shoulder), EDA and RESP from people during a real world driving task. They used LDA to automatically discriminate between low (at rest), medium (highways) and high (city) levels of stress with 97% accuracy. In this case, the signals from each participant were normalized between zero and one, as proposed by Lykken & Venables (1971). Although the authors could not predict every possible stressor during the experiments, the drivers were instructed to drive through pre-established routes and not listen to the radio which provided some consistency across participants.

This thesis studies the stress of office workers during five days of their regular work day and assesses the value of different types of wearable data (physiological, contextual and behavioral) in the context of self-reported stress recognition.

2.8 Managing Stress with Technology

The main focus of this thesis is to develop methods that can help better understand and measure stress in real-life settings with the motivating goal of helping manage stress when most needed. To help motivate this work, this section briefly overviews work in the area of technology-based stress management interventions.

Many people address unnecessary stress in unconventional ways that provide short-term relief but may increase stress in the long-term (e.g., watching TV, overeating).
Furthermore, these behaviors may also reinforce negative behaviors that may enhance the disruptive effects of stress (e.g., smoking, drinking, and drug consumption). On the other hand, some of the most effective and well-accepted methods to address stress include but are not limited to (Varvogli & Darviri, 2001): progressive muscle relaxation, biofeedback, transcendental meditation, cognitive behavioral therapy, and mindfulness-based stress reduction. Among these interventions, there is no common agreement on which approach works best to manage stress, which may be due, at least in part, to the large variability of people and how they perceive stress. In other words, what works for one person may not work for another.

While the previous methods to address stress do not usually involve the use of technology, recent advancements of technology such as smartphones and wearable devices have encouraged researchers to use them in the context of stress management. Among all the different approaches, self-reflection is arguably one of the methods most frequently explored. The underlying assumption is that through self-reflection, people can gain a better understanding of themselves and thereby gain control over certain affective states such as stress. One of the simplest tools for self-reflection is the diary or journal. When writing everything that occurs during daily life, people can reflect and gain more awareness about significant moments in their life. A relevant example of this approach is the Affective Diary by Ståhl & Höök (2008). In the Affective Diary, the authors combined biosensor data and users’ annotations to automatically create a diary with affectively labeled content. Instead of applying machine learning on the biosensor data to infer the affective states (Picard, 1997), the authors explored open visualizations that changed with the collected data (e.g., higher activity was associated with more movement). This approach is based on the Affective Interaction by Boehner et al. (2005) and allows users to understand, interpret, and experience emotions in their full complexity and ambiguity. In a similar context, Vaara et al. (2009 & 2010) and later on Sanches et al. (2010) explored different types of visualizations that helped users to visualize their stress levels. The three studies agreed that an effective visualization should: 1) be open to interpretation and flexibility while allowing users to recognize themselves, 2) allow interactive visualizations of prior states that help understand the current state, and 3) provide a feeling of fluency that avoids discrete states. In a separate study, Ferreira (2008) used similar visualizations and showed them in real-
time to people undergoing stressful tasks (e.g., public speaking, interview). One of the main goals of the study was to evaluate whether an open visualization, which requires cognitive effort to interpret, would have any impact on the stress of people. After an iterative design process with three different people and evaluating the final visualizations with 6 other participants, Ferreira found that real-time visualizations not only did not increase stress but also empowered users to take control over the stressful situation. Nevertheless, this claim may need further validation as the number of participants was relatively small and the experiment was carried out in a controlled laboratory setting, where the nature of stress is fundamentally different from that observed in real-life situations (Wilhelm & Grossman, 2010).

Designing stress interventions that do not elicit additional stress is probably one of the main design concerns of researchers. Moraveji & Soesanto (2012) examined this problem in the context of user interfaces and devised ten design heuristics to minimize users’ stress. Motivated by psychophysiological research, the heuristics aimed at maximizing the feeling of control (e.g., allowing the user to customize interruptions) while minimizing the feeling of uncertainty (e.g., acknowledging users’ actions and interpretations). When addressing stress in real-life settings, ambient displays (Weiser & Brown 1996) could be especially useful as they allow providing information through the periphery, without requiring too much cognitive effort. Partly motivated by the ambient displays, Sadi (2012) designed and created several self-reflection tools to enhance and improve the decision making in different real-life situations (e.g., web surfing, eating, energy use behavior). These tools, called ReflectOns, fit in as part of interactions that already take place, providing just-in-time information and situational awareness that may improve the decision making process. After creating and evaluating different prototypes, Sadi demonstrated the benefits of providing just-in-time information to influence behavior change and also highlighted the potential benefit of using non-just-in-time feedback to reinforce the just-in-time information. Finally, Sadi also highlighted the difficulty of evaluating such systems, especially determining whether the systems evoke a long-term effect. Factors such as novelty, placebo effect, social demand characteristics, and social acceptance can be major challenges when evaluating this type of system.
Another relevant research area when developing interventions is Persuasive Computing (Fogg, 2002). This area focuses on the technology that influences users’ thoughts and behaviors through persuasion and social influence. A good and simple representative example is the one-click-buying method of Amazon that encourages users to make online purchases with a single click by minimizing the required effort. The amount of studies that rely on Persuasive Computing techniques are continuously increasing and have been applied to influence a large variety of behaviors. For instance, Morris et al (2008) created SuperBreak, an interactive tool with the goal of incentivizing ergonomic typing breaks. In particular, they implemented four different types of breaks: traditional break (simple notification without computer activity), vision-based document reading (allowing users to continue working without keyboard and mouse), passive video presentation (either informative or entertaining content), and an interactive computer vision-based game (forcing users to move their arms and touch virtual blocks). After providing these interventions to 26 office workers for ten days, they found that the interactive computer vision-based game yielded the highest percentage of completed breaks, demonstrating the potential benefits of having interactive interventions. However, the authors acknowledged that different situations may require different types of interventions (e.g., document reading may be better when a deadline is approaching) so it is recommendable to have a repertoire of interventions available. In the context of stress, a relevant study that relies on biosensor data is iHeal by Boyer et al. (2012). The goal of this project is to detect the stress associated with drug cravings through physiological data and provide a just-in-time intervention through the cellphone which prevents the intake of illicit substances (e.g., showing photos of family members). Although their evaluation study with seven homeless veterans did not validate the effectiveness in reducing the disruptive behavior, they found that many participants were concerned about the social stigmatization due to the biosensor. Furthermore, the authors provided insightful schematics on how to combine the different technologies, while protecting the information from participants. In a separate but similar study, Tam (2012) created WellBee, a system that similarly used biosensors to sense stress and avoid stress-related eating. In particular, Tam created two different interventions, one that provided empathic messages (e.g., acknowledging how the user felt, thanking for taking the time to provide information) and another without empathic
messages. After running a study with seven students, Tam did not find any significant reduction of stress-eating events. However, several participants reported increased awareness of stress and mood due to the system. Interestingly, a majority of the participants preferred the non-empathic version because they found the empathic version to be perceived as fake and sometimes condescending. This finding is in opposition to previous studies (e.g., Bickmore & Picard, 2004; Liu, 2004; Picard & Liu, 2007) that found some benefits when using interfaces with empathic messages (e.g., increased willingness to continue using the system, decreased perceived number of interruptions). Another challenge that Tam encountered during the study was that many of the participants did not notice the phone notifications (either through sound or vibration) when they received the interventions. Although this will always be a challenge when delivering just-in-time information in real-life settings, the use of other devices such as head-mounted and wrist-worn devices could partially mitigate the problem. Finally, the effectiveness of the biosensors to provide just-in-time information was not reported by the previous studies but it seems, however, a promising research area that requires further study.

Finally, a relatively new effort to address the problem of stress is led by Moraveji at Stanford University. In a paper with his colleagues (2011) and later on his Ph.D. thesis (2012), Moraveji explored the use of technology (also called Calming Technologies) that helps users to maintain an optimal cognitive, physiological and emotional resting state while performing tasks. In one of the studies, Moraveji monitored the breathing rate of users while working at the computer and triggered an intervention whenever the breathing rate increased above a certain threshold. The interventions were aimed at reducing the respiration rate and the different approaches consisted of different peripheral visualizations (e.g., changing the brightness of the menu bar, translucent horizontal bar moving up/down at the target rate). After evaluating the system with 13 participants during 20 minutes of regular computer activity, participants showed significantly lower breathing rates due to the interventions. However, no evidence of persistent rate change was found. In a separate study, Paredes & Chan (2011) studied the impact of different stress reduction technology-based interventions. In particular, 20 participants were stressed in the laboratory and were provided two of the following interventions: playing games, guided acupressure (through vibro-tactile motors in a bracelet), guided breathing, and social networks (texting with
close finds through text messages). After evaluating the interventions, Paredes & Chan could not find any significant differences, strengthening the argument that different types of interventions may work for different people. However, they found that the social networking intervention obtained higher (but not significant) ratings for likeability, potential benefit, and efficacy. More recently, Morris at MIT recently demonstrated the possibility of leveraging the wisdom of crowds (initially through crowdsourcing platforms such as Mechanical Turk) to help individuals reappraise stressful thoughts and reduce anxiety (Morris & Picard, 2012; Morris, 2015). However, the biggest benefit with their approach came from helping others as opposed to receiving help (Ochsner & Gross, 2005; Dore et al., under review).

Most of the previous studies were conducted in recent years, demonstrating the growing interest in the use of technology to address the problem of stress. Although one of the most common approaches is self-reflection, researchers have started to explore a wide variety of methods that represent important directions for future research (e.g., use of wearable devices to provide just-in-time information, social interactions to help cope with stress). Stress reduction interventions can benefit from findings in many different research areas such as Ambient Displays, Affective Computing and Persuasive Technologies to name a few. However, due to the nature and complexity of stress, there is a need to combine different fields of study such as Psychology, Psychophysiology and Behavioral sciences.

Two of the most common research challenges when developing stress interventions are: 1) avoiding the generation of additional stress through the intervention, and 2) finding more systematic and methodological validation strategies that allow quantifying not only the impact of interventions but also the long-term effects after the interventions. Stress is a complex phenomenon that highly depends on the person and the context and, therefore, it is recommended to have ecosystems of personalized interventions (e.g., Paredes et al., 2014, Sano et al., 2015) that are selectively triggered based on different stressors. In the future, these interventions will also need to evolve with the user and adapt to new contexts so that they can be used effectively to neutralize the negative effects of chronic stress conditions.
2.9 Conclusions

This chapter has provided an overview of workplace stress models and alternative measurement strategies as well as relevant work that addresses some of the main real-life measurement challenges (e.g., ambulatory self-reports and context gathering). Moreover, we have reviewed relevant studies in the context of stress management with technology to help motivate the work of this thesis. While there has been extensive research for each of the areas, we are still far from having a system that can unobtrusively and comfortably measure stress during daily life. Motivated by the previous studies as well as some early research explorations (described in Chapter 3), this thesis examines how the unique advantages of wearable devices (e.g., proximity to the body, different form-factors) can help advance the previous challenges. In particular, Chapter 5 studies how wearable devices can improve ambulatory self-report data collection, Chapters 6 and 7 describe novel methods to provide comfortable physiological measurement, and Chapter 8 studies the relationship between self-reports and wearable data with supervised learning methods.
Chapter 3

Early Exploratory Work

The research described in this thesis expands earlier work in stress measurement, context gathering, and stress management that helped shape and motivate this thesis. This chapter briefly overviews the different explorations, which include both wearable and non-wearable approaches, for each of the previous areas.

3.1 Stress Measurement

3.1.1 Pressure-Sensitive Keyboard and Capacitive Mouse

As discussed in the previous chapter, a commonly explored strategy to non-invasively measure stress consists of measuring behaviors that are influenced by stress such as typing on the keyboard or handling the mouse. Due to the pervasiveness of these two devices in the workplace environment, many researchers have explored their use in the context of stress measurement. For instance, Vizer et al. (2009) created a system that measured keystroke and linguistic features (e.g., number of nouns and verbs, self-reference rates) of 24 computer users, and were able discriminate non-stressful writing with writing after cognitive (arithmetic and memory tasks) and physical (aerobic and resistance exercise) stress with accuracies of 75% and 62.5%, respectively. In a separate study, Lv et al. (2008) used a pressure-sensitive keyboard to recognize 6 emotions of 50 individuals during a laboratory study. In particular, participants of their study had to watch different emotional clips and type a series of keywords after each. Although they obtained an average classification accuracy of 93.4%, stress was not considered as one of their emotions. In the
context of affect sensing from computer mouse, the early work performed by Qi et al. (2001) demonstrated the possibility of sensing frustration of computer users by attaching pressure sensors to the mouse. Finally, Wahlström et al. (2002) monitored the pressure on the mouse as well as other physiological signals in 15 subjects while performing text editing tasks with different levels of time pressure and verbal provocation. One of their major findings was that higher levels of stress yielded increased pressure applied to the button of the computer mouse, as well as more repetitive wrist movements.

![Pressure-sensitive keyboard and capacitive mouse](image)

**Figure 1**: Pressure-sensitive keyboard (left), and capacitive mouse (right)

Motivated by the previous studies, we wanted to explore the feasibility of using a pressure-sensitive keyboard (Dietz et al., 2009) and a capacitive mouse (Villar et al., 2009) to sense the manifestations of stress of computer users (see Figure 1). In particular, we collected data from 24 participants and asked them to perform several computerized tasks in a within-subjects laboratory study. The tasks were as follows:

- **Text transcription**: Participants had to transcribe a text with and without the presence of stressors (e.g., time pressure, extra compensation, different font sizes, and environmental noise).

- **Expressive writing**: Participants were requested to re-experience a relaxing and a stressful recent memory and describe them.

- **Mouse clicking**: Participants were requested to perform the traditional mouse clicking Fitts' law task (MacKenzie, 2009) after the previous tasks.

After each of the tasks, participants had to rate their self-reported stress levels in a 7-point Likert scale with end points "Very stressed" and "Not stressed at all." The results of the
study indicated that increased levels of stress significantly influenced typing pressure and amount of mouse contact of computer users. While >79% of the participants consistently showed more forceful typing pressure, 75% showed greater amount of mouse contact. Furthermore, we determined that considerably small subsets of the collected data (e.g., less than 4 seconds for the mouse clicking task) suffice to obtain similar results, which could potentially lead to quicker and timelier stress assessments.

More details about this exploration can be found in (Hernandez et al., 2014).

### 3.1.2 Wearable Electrodermal Sensors

One of the advantages of wearable physiological sensors is that they are in close contact with the body throughout the day, offering the possibility to provide more frequent and continuous assessments. To further study this, this work explored the possibility of using wearable electrodermal sensors to capture the stress levels of call center employees.

In a completely uncontrolled work setting, nine call center employees were asked to wear a wrist-worn EDA sensor (Poh et al., 2010) on their non-dominant hand for five days, and to report the stress levels associated with each of the calls. Specifically, they were asked to respond to “*How was the last call?*” using a 7 point Likert scale right after each call. The end points of the question were labeled as “*Extremely good*” indicating non-stressful and “*Extremely bad*” indicating very stressful. Figure 2 shows an example of one day of collected data.

![Figure 2: Example of data from one participant that contains calls (dots), stress ratings (darker areas represent more stressful calls), and break times (squares)](image-url)
At the end of the week, we collected information for 1500 calls and found large differences in how individuals reported stress levels, with similarity from day to day within the same participant, but large differences across participants. Moreover, we observed different trends in term of physiological activation and EDA lability (Mundy-Castle & McKiever, 1953). To address these differences, we modified the objective function of linear Support Vector Machines to:

1) Automatically select the training data of participants that were more similar to unseen testing samples in terms of EDA lability. This was achieved by computing the average number of EDA peaks per second for each sample and using clustering methods to find participants whose data was more similar to the testing samples.

2) Incorporate the tendency of some employees to report higher stress levels than others. This was achieved by adapting the class priors of the Support Vector Machines based on the amount of reported stressful calls by each individual.

After dividing the calls into more stressful and less stressful calls, we were able to achieve a recognition accuracy of 78.03% when trained and tested on different days from the same person (i.e., person-specific models), and 58.45% when trained and tested on different people (i.e., general models). However, when we incorporated the proposed modifications, we were able to increase accuracy up to 73.41%, which was fairly close to person-specific models.

More details about this exploration can be found in (Hernandez et al., 2011).

### 3.2 Context Gathering

#### 3.2.1 Behavioral Activity

Traditional methods to effectively capture contextual information in real-life studies involve old-fashioned annotation tools such as use of pen and paper. While this methodology is very flexible and inexpensive, it is also very slow, cumbersome, prone to human error, difficult to share, and limited in its ability to capture rich and complex behavioral and environmental data. Recent advances in mobile technologies, however,
have provided an opportunity to improve this and efficiently and accurately capture other types of contextual information (Ayzenberg, Hernandez & Picard, 2012).

The focus of this exploration was to create an easy-to-use mobile application for collecting, labeling, and sharing complex multimodal (videos, audio and images) behavioral data. In particular, we focused on the problem of effectively capturing stressful challenging behaviors of people diagnosed with Autism Spectrum Disorders (ASD) such as self-injury, property destruction, and repetitive behaviors. The ability to obtain this type of data by therapists, teachers and parents can help with both intervention and behavioral phenotyping efforts.

Through collaboration with behavioral scientists and therapists at the Groden Center (large non-profit school for people diagnosed with ASD), we identified relevant design requirements and used them to create the annotation tool. Figure 3 shows a screenshot of the interface of the tool. Furthermore, we conducted a two-level evaluation: one with users at MIT verifying that the application and its different functionalities worked as designed (debugging), and another with eight teachers assessing usability while working with
children with ASD. Results from a usability scale (Brooke, 1996) showed that the annotation tool was preferred by teachers during the study in comparison with traditional methods (pen and paper-based method). In particular, the speed, accuracy, and ease of use were significant factors in its preference over traditional methods.

More details about this exploration can be found in (Hernandez et al., 2012, and Sano et al., 2012).

3.2.2 Visual Context and Physiological Signals

A key factor towards better understanding the relationship between physiological signals and emotional subjective experience in real-life settings consists of capturing contextual information (e.g., social interactions, environment, and activities). However, due to the large diversity and variety of real-life scenarios, it becomes difficult to develop a one-fits-all solution that captures all the relevant information. Motivated by this limitation, this exploration proposed a system that allowed users to comfortably capture and naturally reflect on their physiological data throughout daily-life activity.

To passively gather information, we designed a wireless system that leveraged the advancements of biosensors and cellphone technology (see Figure 4 and Figure 5). In particular, we leveraged a previously designed multimodal biosensor (Fletcher et al., 2011), which captured electrodermal activity, skin temperature, and 3-axis accelerometer, and
reconfigured a standard mobile phone to be used as a pendant that hangs around a person's neck and gathers visual context from the perspective of the individual.

To visualize the information, we augmented the capabilities of an ordinary mirror and created interactive visualizations that efficiently summarized daily activities. Thus, allowing users to reflect not only on their outer appearance but also on the inner state throughout their day. By monitoring physiological data such as EDA, which tends to be activated by personally significant events, the person could then efficiently retrieve, browse and organize the significant moments that could be captured by the wearable camera. Figure 6 illustrates two visualizations that link the visual context with the physiological information:

A. **Timeline.** A person’s EDA signal from the entire day is presented as a time series plot, along with a sequence of images that are synchronized to the EDA timeline.

B. **Mosaic.** The system selects images associated with the greatest electrodermal activation (amplitude of the signal) for certain periods of time (e.g., hour, day), and arranges them as a mosaic with their sizes proportional to their activation.

This system was tested for over 15 days (6 hours/day on average) by a single user with the main purposes of ensuring everything worked reliably and identifying the main advantages and limitations versus other approaches. Among some other findings, the wearable sensor
system was found to be comfortable and unobtrusive due to the wireless transmission of the data (over the Internet or Bluetooth), with no need to carry additional devices (e.g., laptops, head-mounted cameras), and with an easy to conceal wearable sensor.

Linking both physiological data and visual context in real-life settings not only may help to find more emotionally significant moments of people, but also can enhance the communication between people, catalyze introspection, inform medical diagnoses, and improve scientific understanding of psychophysiology in natural settings.

More details about this exploration can be found in (Hernandez et al., 2013).

### 3.3 Stress Management

#### 3.3.1 Smiles “in the Wild”

Section 3.1.1 has shown that stress can mediate behavioral responses such as computer typing and mouse handling. This section describes an exploration that focuses on facial expressions and leverages self-reflection to help promote positive emotions of a large community.

The meaning of facial expressions has been extensively studied for decades by cognitive, social and clinical physiologists. Among other expressions, smiles have been shown to have a bidirectional link between emotional experience and facial movement in certain conditions (Soussignan & Duchenne, 2002; Strack et al. 1988). Moreover, some studies have shown that the act of laughter releases endorphins, which reduces stress (Bennet et al., 2003) and, consequently, strengthens the immune system. Although smiles are not always indicative of good mood (Hoque et al., 2011), they are more commonly associated with positive feelings such as well-being and happiness, and it is easily understandable why they have become the social convention in photographs. In fact, smiles have become so prevalent that people purposely induce them with standard words and sentences (e.g., “Say cheese!”), and modern commercial cameras automatically detect them to decide when to take the optimal picture.
This exploration created an interactive technology that enables a new type of live portrait of a community, creating a time-changing location-based emotional footprint. In particular, we created an interactive installation that automatically encouraged, recognized and counted smiles of participants strolling by. We deployed four of the systems at major locations on a college campus for ten weeks. The online portrait continuously showed the collected information in a variety of visualizations and interactive graphs. For instance, one of the visualizations overlaid the amount of smiles of each location on the campus map, with the “hotter” regions indicating a higher smile count (see Figure 7). We also displayed the live-feed captured from the camera using projectors and large screens (see Figure 8). The live-feed overlaid a yellow neutral face if the person was not smiling (“smile intensity” < 50%) and a green happy smiley otherwise. The interface also displayed a smile-barometer (left-side) that depicted the aggregated smile estimation for everyone present in the image. For example, if there are two people smiling with “smile intensity” values being 70% and 80%, the smile-barometer would have a value of (70+80)/2 = 75%.

We evaluated the interaction of people and the effectiveness of the system in a college campus during a ten-week installation, and found self-reported mood improvements in a survey of 300 people. In particular, the survey responses indicated that most people interacted with the system when they had some free time and/or when they were in a positive mood, which are both expected to be negatively correlated with workplace stress. Additionally, many respondents stated that they briefly felt better after interacting with the system or seeing others interacting with it. Moreover, many people used the system as a
self-reflection tool. For instance, a person reported "I became a little more aware of what my projected mood was and I smiled to make it better," and another "It was definitely a great way to remind yourself to smile - just like seeing someone smile or a baby might do!" There were also few occasions in which the community enjoyed playing with the systems and exploring (and exploiting) its limits, which made the interaction more enjoyable (see Figure 9 for two examples). Overall, the survey responses indicated that the system was a positive mood booster.

Figure 9: (Left) students printed several smiling faces and hung them to increase the smile-barometer, and (right) the software detected the face of a dog.

The system captured the average intensity of smiles and the amount of people interacting with the system every 2 seconds, providing insightful information to better understand the community. For instance, even though there were fewer people around over the weekend, they all smiled more intensively (see Figure 10). Also, we found that the intensity of smiles reached its lowest on Tuesdays (although not significant) which is consistent with the temporal distribution of self-reported stress levels obtained in the 5-day workplace stress study of this thesis (see Figure 47). We also found that smile parameters were largely correlated with academic events and festivities of the college. For example, smile intensities were measured to be the lowest during the exams period and the highest the day after graduation, which are both correlated with stress.
In summary, this exploration was both a scientific experiment and an interactive installation. Even though we did not measure stress directly, the previous findings suggest that the number and intensity of smiles are potential useful metrics to indirectly capture stress. Moreover, we have shown how a simple and positive installation can be used to instigate positive mood and help manage emotions.

More details about this exploration can be found in (Hernandez et al., 2012).

### 3.3.2 Head-mounted Wearable Devices

Among other wearable devices, this thesis studies how head-mounted wearable devices such as Google Glass can help improve the measurement of stress in real-life settings. There are, however, unique aspects of this type of devices that could be useful to better manage emotions.

Google Glass can be connected with existing state-of-the-art wearable devices that can comfortably monitor several physiological signals. For instance, we have created a custom-made Android program that connects Google Glass with the Q™ sensor through Bluetooth and enables continuous measurement and visualization of the captured data. Being able to wirelessly collect different physiological parameters and have them automatically synchronized with the same device is critical for the understanding of emotions during natural settings (see left of Figure 11). Moreover, the see-through display of the devices enables real-time self-reflection of affective information, which can potentially empower the wearers to change their behavior and better influence certain emotional states when most needed.
As shown in the previous section, cameras offer the opportunity not only to capture indicators of stress but also to help manage it. To start exploring a wearable version of this, we connected Google Glass with the custom-made smile recognition software discussed in the previous exploration, allowing real-time monitoring and visualization of surrounding smiles (see center of Figure 11). When displaying the average intensity of the smiles of people, for example, users can increase their awareness and provide more control over certain situations. For instance, during a public speaking scenario, the speaker may not be able to look at every face of the audience. However, by displaying aggregated metrics such as the average intensity of smile on the see-through display, the speaker could know when the audience is losing engagement and appropriately modify the speech. In the long-term, this type of technology could help simplify complex information (e.g., ambiguous facial expressions, aggregate the responses of many people) and potentially help reduce anxiety in certain situations. It is important to note, however, that even simplified information can become disruptive during stressful events such as public presentations.

Capturing and processing continuous affective information in real-time with Google Glass (either through other wearable devices or the approaches proposed in this thesis) enables providing timely and gentle emotion management interventions. For instance, Google Glass could display a breathing guide (a line that rises and falls, slowly, guiding the user towards deep in- and out-breaths), once the sensor detects increased levels of
relaxation, the display would fade away (see right of Figure 11). This intervention is very similar to traditional biofeedback exercises which are widely used for emotion regulation. However, Google Glass offers the possibility of comfortably and unobtrusively delivering the intervention when most needed (e.g., when dealing with computer problems).

Finally, when considering several Google Glass devices, they could potentially communicate with each other and exchange affective information to facilitate communication in certain scenarios. For instance, some of the most prevalent and disruptive symptoms of Autism Spectrum Disorders (ASD) include stressful challenging behaviors (e.g., self-injury, repetitive behaviors) and impaired verbal communication. If teachers, therapists or family members could have access to real-time information of the internal state of people with ASD, they could potentially prevent the occurrence of challenging behavior and gain deeper understanding of the emotional states of the individuals. Recent work by Riobo et al. (2014) has started exploring some aspects of this.

More details about this exploration can be found in (Hernandez et al., 2014).

3.3.3 Driving Scenario

To design effective stress management interventions it is critical to consider the characteristics of a particular setting. This section describes how the stress of drivers can enable several interactions in the car with the goal of helping to manage stress.

Driving can be an emotionally stressful experience. Some of the main stressors are the lack of control, the potential negative impact of accidents, and the high cognitive load that is required (Eyben et al., 2010; Grimm et al., 2007). While certain amounts of stress help the driver to remain alert and attentive, too much or too little can negatively impact driving performance. When abnormal levels of stress are detected, the car can use the information to automatically adapt its interactions with the driver and increase individual and social awareness, helping the driver to better manage stress. Some of the possible interactions are as follows:

- **Adaptive music.** Music has been widely used for emotion manipulation. If the car detects that the driver is overly stressed, it may recommend lowering the volume or listening to more relaxing audio. However, auto-selecting songs may elicit additional stress due to the lack of control so a compromise between the two methods may be the
best solution. In general, solutions aimed at reducing stress should give participants a greater sense of control, not less.

- **Empathetic GPS.** Nass et al. (2005) performed a relevant study where the voice of the GPS matched the induced emotional state of a driver (subdued or happy) and found that congruent emotional states improved driving performance. Similarly, the GPS could also match linguistic features that change during stress episodes (e.g., Vizer et al. 2009) to more effectively convey information.

- **Calming temperature.** During intense stress episodes, it is common to experience a sensation of increased heat. The car could potentially detect early indicators of stress and automatically adapt the temperature and its intensity to help alleviate this sensation.

- **Corrective headlights.** One of the changes associated with stress is the loss of peripheral vision, also known as tunnel vision. This is especially dangerous during the night where vision is already limited. The car could compensate for this by auto-adjusting the field of view of headlights.

- **Reflective dashboard.** Many technology-based stress reduction interfaces provide physiological information to the user so s/he can reflect on it and take control of the situation (e.g., MacLean et al., 2013). An appropriate interface to achieve this in the car could be changing the color of the dashboard based on physiological changes. For instance, green and red colors could indicate a more relaxed or stressed driver, respectively. We have further explored this idea and created a small prototype that wirelessly connected the Affectiva Q™ sensor with an array of LED lights inside a Remote Controlled car. We then asked several volunteers to wear the sensor and drive the car through a circuit with several obstacles as fast as possible (see Figure 12). Although this is only a small step towards the final goal, we were able to see how the lights turned red every time drivers crashed against obstacles and/or had to make pronounced turns. Furthermore, the feedback system made some participants more aware of their affective states and several slowed down when the reddish color appeared. Figure 13 shows an adaptation of the prototype in the car dashboard.

- **Communicative paint.** The same type of information could also be provided to other road users to improve driving safety. For instance, if a cyclist or a pedestrian detects that a driver in a car is overly stressed, s/he could intentionally keep larger separation
and prevent adding additional stress. The communication of the information could be done through LEDs such as the previous prototype or even by changing the color of the whole car. To further explore the latter idea, we built a prototype of a small car and painted it with thermochromic paint, which has the ability to change color based on different temperatures. We used the same setup as the previous prototype but instead of connecting the biosensor to the LEDs, we connected it to a thermoelectric/peltier mini module. This setup enabled us to control the temperature of the surface based on the physiological readings of the person. Although we were able to change the color fairly quickly, the amount of current (1A) required to manipulate one square inch may not be easily scalable to a full-size car.

- **Large-scale applications.** Collecting stress data of multiple drivers and associating it with other kinds of information can enable other compelling interactions. For instance, some days a driver might prefer “the least stressful” route home, even if it is not the fastest route. The collective stress level can be combined with mapping/routing tools to geographically identify routes that minimize stress whilst taking you to your destination (see Figure 14). This capability could be especially useful for people who have high sensitivity to stress or who just need to be calmer before going home. Furthermore, analyzing the stress levels of hundreds of drivers around the city at different times of day and different days of the week could help find stressful locations and help design more livable cities and commutes that improve the well-being of citizens.

This exploration overviewed a number of potential opportunities to help manage stress in a driving scenario and has started prototyping some of them. In the future, this type of interaction could not only be used to reduce stress of drivers but also to bring empathy to the driving experience while improving driver safety and increasing social awareness.

More details about this exploration can be found in (Hernandez et al., 2014).
3.4 Conclusions

This chapter has overviewed several ways in which technology has been used for stress measurement, context gathering, and stress management, highlighting some of the main problems and how different approaches can help address some of them.

The first exploration demonstrated that pervasive computer keyboards and mice can be used to comfortably and non-invasively capture relevant aspects of stress (e.g., change of pressure and amount of contact). However, the amount of stress assessments is fundamentally limited by the amount of interaction users have with such devices. Thus, if a computer user is speaking on the phone or reading an article, no information can be gathered with the keyboard or the mouse; thus, other types of devices (e.g., microphones) would be needed to perform stress assessments. To improve the continuity of potential assessments without the need of instrumenting every location, we explored using wearable EDA sensors in a stressful work scenario such as a call center. In this case, we developed
methods that recognized a majority of the stressful calls by only analyzing the physiological responses of employees. However, it is important to note that one of the fundamental limitations of EDA is that high arousal levels of both positive and negative events can elicit similar-appearing responses. While the type of arousal experienced in call center settings is expected to be more frequently associated with negative arousal, other workplace scenarios may similarly contain both types. To effectively capture both types, additional sources of information need to be considered (Cacioppo & Tassinary, 1990). This thesis provides a first evaluation of how different types of physiological (e.g., respiration, heart rate, heart rate variability), contextual (e.g., ambient temperature, atmospheric pressure) and behavioral (e.g., motion data from different body locations) wearable signals can be used in a different real-life work setting.

To help capture contextual information such as behavioral activity, we designed and created a tool that enables the effective capture of rich behavioral activity in complex and challenging scenarios. However, to maximize the frequency and objectivity of collected data, automated tools that do not require the attention or active intervention of the person are usually preferred. To address this, we proposed a wearable approach which repurposed a smartphone to help capture rich visual context during daily life and used a gesturally-controlled mirror to effectively browse both physiological and contextual information. After this exploration was performed, new wearable cameras have become commercially available (e.g., Narrative Clip) which similarly capture the visual context of daily activity. Instead of using the prototypes we developed, this thesis will rely on the new commercial devices and use them to help participants reflect on their data and to link them with the physiological responses to improve stress measurement and understandings.

Finally, this chapter has reviewed several explorations in the context of stress management that have helped motivate a need for better automated stress recognition. The different explorations highlight potential applications where technology might help a person minimize unnecessary stress for various positive outcomes. For example, we have highlighted different potential stress management applications in a driving scenario, where reduction of driver stress also has the potential to increase driver safety.

To sum up, there are a wide variety of challenges when measuring and using stress information in real-life. Some of the main challenges include but are not limited to
providing continuous measurements, addressing individual differences, and capturing the large variability of real-life scenarios. Based on the learned experiences from each of the explorations, the remainder of this thesis studies how different types of wearable devices (e.g., head-worn, wrist-worn) can help minimize some of the challenges.
Chapter 4

Real-life Workplace Study

To better study some of the challenges of real-life stress measurement, the analysis presented in the remainder of this thesis (except for Chapter 6) is based on data collected during a real-life workplace study in which several participants underwent their regular work day while wearing several wearable devices and engaging in frequent experience sampling. This chapter provides a description of the experimental protocol and tools used for the data collection as well as some of the main design considerations. The chapter is divided as follows. First, we review the selection of wearable devices and the types of information they capture. Then, we describe the experimental protocol and the questions participants were asked throughout the study. We then describe how we anonymized and protected the privacy of participants. Finally, we describe the recruitment process and provide a preliminary data overview.

4.1 Wearable Devices

The main motivating goal of this work is to gain a deeper understanding of how wearable devices can be used to improve stress measurements. Exploring this question requires the consideration of devices that can effectively capture relevant aspects of physiological changes, stress self-reports, and contextual information.

Due to the large variety of wearable devices available in the market, we followed a bottom-up approach in which we first determined our needs and then found the most
appropriate devices for them. Based on some of the main challenges described in the previous chapters, we selected the following parameters as the main selection guidelines:

- **Unobtrusive.** Successful life-log capture devices become transparent and do not alter the user's behavior during daily-life activity. Among other characteristics, the devices should be wearable, non-intrusive, comfortable, and require minimal user interaction. This is critical in our study as we want to measure work-related stress and not the stress elicited by the devices.

- **Rich sensor capabilities.** The stress response is associated with a set of physiological responses. To accurately capture it, we need devices that contain multiple and varied sensing capabilities that can not only capture the different parts of the stress response but also enable collecting self-report levels.

- **Quality and resolution.** Wearable data during daily life can contain a large array of artifacts. For instance, electrodes may detach, batteries die, and strapping mechanisms may become loose. It is important that the selected devices offer the highest quality of data possible while being flexible and adaptable to different types of people (e.g., genders, body morphology). To achieve some of the goals of the study, there are also some quality and resolution requirements, such as high sampling rates for the motion sensors.

- **Battery life.** While battery life is almost always a problem with wearable devices, it is especially critical in our experiment as we do not want to add additional stress to the participants by having to continuously charge the devices. To minimize the potential burden, we tried to prioritize devices with a battery life that could last for a whole day with one battery charge. However, ensuring a high data quality (e.g., high sampling rates) while capturing a large number of signals can be quite challenging.

- **Control and scalability.** To provide full control of the acquisition and storage of sensor information as well as providing maximum flexibility when creating the self-reporting tools, we prioritized devices that allowed us to install our own custom applications. Furthermore, we selected devices sharing a common development platform such as Android to significantly reduce the costs associated with multi-device support.
Considering the previous factors, we selected a set of seven wearable devices to be worn by each participant during their regular work day. Table 2 shows a summary of the devices as well as the information that was recorded by each of them. The selected devices can be separated into three major categories based on their main purpose during the study.

- **Self-reporting tools.** A fundamental part of stress measurement is to gather ground truth self-reported stress levels. While commonly explored approaches involve the use of a cellphone, this work also considers other wearable devices with different form-factors that may be more appropriate. In particular, we used the Google Glass (Glass) which is head-worn, and the Gear Live (Gear) which is wrist-worn, and compared their performance with a Galaxy S4 smartphone which was carried inside the trouser pocket.

- **Physiological sensors:** As mentioned earlier, an important part of stress is the physiological set of responses associated with it. To capture some of the main aspects we used the BioPath which captures electrocardiography (250 Hz) and respiration (25 Hz) as well as heart and breathing rates (1 Hz) from the chest, two QTM sensors which collected electrodermal activity and skin temperature from the two wrists (8 Hz), and the Gear smartwatch which sporadically estimated heart rate from wrist Blood Volume Pulse.

- **Context and behavior gathering tools:** A key aspect when studying stress in a natural setting is to capture and understand the context in which stress is happening. The main sensor we used to capture context is the Narrative Clip which takes a photo every 30 seconds. This information was used to capture daily visual context that can help understand the origin of physiological changes (Hernandez et al., 2013). Additionally, each of the selected devices also provided other types of contextual information. For instance, Glass captured the amount of light (1 Hz) of the environment, the Galaxy S4 recorded the atmospheric pressure (5 Hz), temperature (1 Hz) and humidity (1 Hz), and the Gear estimated the number of steps taken by the user during the day. Moreover, all of the sensors except the Narrative Clip also captured behavioral motion data which can be used to better understand the level of activity and potentially other physiological data (see Chapters 6 and 7). However, we mainly focused on the motion data collected by the Glass, the Gear, and the Galaxy S4 (100 Hz).
Table 2: List of wearable devices each participant carried during the study

<table>
<thead>
<tr>
<th>Device</th>
<th>Location</th>
<th>Recorded data</th>
<th>Placement considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Glass</td>
<td>Head</td>
<td>3-axis accelerometer</td>
<td>The device was adjusted every morning to ensure participants could see the display and the touch sensor was activated.</td>
</tr>
<tr>
<td>(Google, Inc.)</td>
<td></td>
<td>3-axis gyroscope</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Light</td>
<td></td>
</tr>
<tr>
<td>Gear Live</td>
<td>Wrist</td>
<td>3-axis accelerometer</td>
<td>This device was placed on the non-dominant hand. The band was placed as tight as possible while still remaining comfortable.</td>
</tr>
<tr>
<td>(Samsung, Inc.)</td>
<td></td>
<td>3-axis gyroscope</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heart rate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of steps</td>
<td></td>
</tr>
<tr>
<td>Pocket</td>
<td>Wrist</td>
<td>3-axis accelerometer</td>
<td>Participants were asked to have the phone inside the front pocket of their trousers and were encouraged to have the touch display facing outwards to prevent accidental touch interactions.</td>
</tr>
<tr>
<td>Galaxy S4</td>
<td></td>
<td>3-axis gyroscope</td>
<td></td>
</tr>
<tr>
<td>(Samsung, Inc.)</td>
<td></td>
<td>Atmospheric pressure</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Humidity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Temperature</td>
<td></td>
</tr>
<tr>
<td>BioPatch</td>
<td>Torso</td>
<td>Electrocardiography</td>
<td>We used new Kendall 535 foam electrodes every day and attached the device to the torso as close as possible to the heart. Skin was pretreated with Uni-Patch pre-TENS wipes (Coviden™) to minimize irritations.</td>
</tr>
<tr>
<td>(Zephyr Tech., Inc.)</td>
<td></td>
<td>Respiratory motions</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Wrist</td>
<td>3-axis accelerometer</td>
<td>The electrodes were placed on the upper side of both wrists to minimize collisions with the sensor. We used new CAT CS5-I gelled electrodes (LeadLock Inc.) every day to minimize artifacts due to motion of the sensor. Skin was also pretreated with Uni-Patch pre-TENS wipes.</td>
</tr>
<tr>
<td>(Affectiva, Inc.)</td>
<td></td>
<td>3-axis gyroscope</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Electrodermal activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Skin temperature</td>
<td></td>
</tr>
<tr>
<td>Narrative Clip</td>
<td>Torso</td>
<td>Photos</td>
<td>The device was clipped as close to the neck as possible and facing forward to capture the visual context in front of the person.</td>
</tr>
<tr>
<td>(Narrative, Inc.)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 Experimental Protocol

This section describes the experimental protocol approved by the Institutional Review Board of the Massachusetts Institute of Technology (COUHES).

During the experiment, participants were asked to wear all the devices during five days of work. The overall experiment was divided into three main phases:

1. **Starting phase.** Participants met with the researchers to learn about the experiment as well as how to use the different sensors during daily life (e.g., how to carry them, how and when to charge them). After providing written consent, participants provided demographic information as well as responded to some surveys about stress levels and emotional awareness (more details in the following section). Participants also received a document containing all the guidelines provided during the starting phase (see Appendix A for more details).

2. **Data collection.** The data collection lasted five work days. At the beginning of each day, participants met with the researchers who ensured the devices were appropriately worn (see placement considerations for each device on Table 2). Throughout the day, participants received several prompts through different devices asking about their emotional levels. At the end of each day, participants met with the researchers to return the sensors and report any relevant comments (e.g., problems with the sensors). Both starting and ending times for each day were flexible as long as participants tried to gather around 8 hours of wearable data.

3. **Ending phase.** After the five days of data collection, participants met with the researchers to respond to some questions about the usability of the different devices and to provide comments about the overall experiment.

To ensure the battery life of the sensors lasted for the entire day, participants received battery chargers and were instructed to charge three of the devices (the phone, the Gear, and the Glass) during lunch time. They were asked to leave them charging for at least 1 hour which was enough to ensure 8 hours of continuous wearable data. When the devices were being charged, the custom-made data logger automatically paused the sensors to maximize charging speed. If participants forgot to charge the devices and the devices reached a low battery level (below 10%), the data logger would trigger a notification.
instructing participants to charge the devices (see Figure 15). This notification would automatically disappear when the devices were connected.

All of the devices were synchronized to a computer on a daily basis and were recharged during the night to ensure they were fully charged for the following day. Moreover, sensor data was visually inspected by the researchers at the end of each day to ensure the maximum quality of sensor readings and quickly address any potential problems (e.g., battery problems, loose electrodes).

Figure 15: Notification displayed on the Google Glass, the Gear Live and the Galaxy S4 when battery levels were under 10%

4.3 Questions and Surveys

To better characterize the participants of the study, we included several surveys. This section overviews the different types of survey questions for each of the phases of the experiment.

During the starting phase of the study, participants had to respond to the following surveys:

- **Perceived Stress Scale (PSS).** We used the 10-item version of the original 14-item Perceived Stress Scale survey (Cohen, Kamarck & Mermelstein, 1983) which has been widely used and validated in many different settings (e.g., smoking cessation, frequency of colds). The main purpose of this survey is to capture the stress levels that a person may have undergone during the previous month. Note that this survey is more adequate to capture the long-term feeling of stress instead of the short and acute type of stress which we are mostly interested in this work. However, long-term stress levels may also influence how people perceive and respond to short-term stressors. Among some other questions, participants had to rate on a 5-Likert scale the following
questions: “In the last month, how often have you been upset because of something that happened unexpectedly?” and “In the last month, how often have you felt that things were going your way?” The end points of the questions were labeled as “Never” and “Very often.”

- **Toronto Alexithymia Scale (TAS).** This 20-item instrument is used to measure alexithymia (Bagby, Parker & Taylor, 1994). Those with higher alexithymia scores have more difficulty identifying and describing their own emotional experience. As this study considers self-reported stress levels as the gold standard, it is critical to identify if a participant may have problems reporting their own feelings. This scale has been successfully used in previous studies demonstrating that people with high alexithymia may have more problems reporting their electrodermal arousal levels (Ayzenberg, 2012), which are also measured in our study. Among some other statements, participants had to rate how much they agree on a 5-Likert scale to the following ones: “I often do not know why I am angry,” and “I am often puzzled by sensations in my body.” The end points of the questions were labeled as “Strongly disagree” and “Strongly agree.”

- **Big Five Inventory (BFI).** This 44-item instrument is widely used to measure the personality traits of participants in five main personality dimensions: openness, conscientiousness, extroversion, agreeableness, and neuroticism (John, Donahue & Kentle, 1991; John & Srivastava, 1999). These factors are not mutually exclusive and, therefore, the same person can rank high on several of them simultaneously. Some of these dimensions have been shown to influence how people perceive and experience stress (Vollrath, 2001) and, consequently, can have a direct impact on the findings of this work. As part of the different items, participants had to rate the following statements on a 5-Likert scale: “I see myself as someone who is talkative,” and “I see myself as someone who gets nervous easily.” The end points of the questions were labeled as “Strongly disagree” and “Strongly agree.”

- **Other information.** Participants also had to respond to some general questions about demographics (e.g., gender, age and height), health (e.g., cardiac, respiratory or musculoskeletal conditions), and use of similar types of devices as those used in the study during their regular lives (e.g., glasses, watches). This information was recorded
to better understand the potential differences in terms of data quality, algorithmic performance, and self-reports.

At the end of each day, participants were presented with the images captured by the Narrative Clip device, and were asked to provide some additional information about their daily activities. The types of questions can be grouped into two categories:

- **Contextual information.** Participants were asked to provide additional information for the times in which a prompt was triggered (either submitted or ignored) during their daily activities. These questions were then used to better understand the ratings and stress levels. The questions included information about activities, social interactions, and body postures during the five minutes preceding the triggering of the prompt.

- **Daily overview.** Participants were asked to provide information about their whole day. Some of the information included relevant stressful events, caffeine intakes, sensor problems, as well as general ratings for their perceived pleasantness and energy levels, amount of daily demands and resources, and stress during their day. To further capture their overall stress levels with a standardized method, participants were asked the short 4-item PSS scale which was slightly modified to ask questions about the day instead of the month. Among some other items, participants had to rate on a 5-Likert scale the frequency of the following questions “*During the day of today, how often have you felt that you were unable to control the important things in your life?*” and “*During the day of today, how often have you felt difficulties were piling up so high that you could not overcome them?*” The end points of the questions were labeled as “Never” and “Very often.”

Finally, at the end of experiment participants were asked some general questions about their experience. In particular, there were two types of questions:

- **Usability of the devices.** For some of the devices we wanted to learn more about the user experience. In particular, we asked several questions about the usability of the Glass, Gear, phone, and the Narrative Clip. These involved general questions such as if they would wear the device during their daily life, whether wearing the device affected their social interactions, and how comfortable the device was during their daily
activities. This information is very relevant to understand the potential benefits and disadvantages of using such devices in the context of experience sampling.

- **Free-form interview.** The last part of the experiment consisted of a 5 to 15 minute free-form interview in which the experimenter asked some general questions about their experiences (e.g., relevant stressors, main problems with the sensors, things that could be improved), the wearable devices (e.g., which ones they preferred and why), and the kind of insights participants would find interesting to learn from the data.

Figure 16 provides an overall overview of all the survey types for each part of the study and Appendix B contains the different questionnaires with all the items.

![Figure 16: Summary of surveys and questions for the different phases of the study](image)

### 4.4 Experience Sampling Questions

During each day of data collection, participants were prompted through several wearable devices to respond to several questions about their emotional experience preceding the prompt. In particular, participants were asked to rate their emotional arousal and valance (Russell, 1980; Bradley & Lang, 1994), their perceived amount of demands and resources (Demerouti 2001; Bakker & Demerouti 2007), and their general feeling of stress during the five minutes preceding the prompt. Finally, they were also asked to rate the disruptiveness of the prompt. Based on previous research, stressful events are expected to be associated with high arousal and negative valence, challenging scenarios are expected to be associated with high demands and high resources, and threatening scenarios are expected to be associated with high demands and low resources.
Table 3 contains the questions, the application interface, and the definitions provided to the participants to help in answering and understanding the questions. This information was also included in the introductory tutorial that participants received at the beginning of the experiment. The following chapter provides more information about how our custom-made application triggered the prompts and how participants interacted with each of the devices to self-report during their daily activity.

4.5 Ensuring Privacy of Data

Monitoring people with devices can easily invade their privacy, especially when this is performed in real-life scenarios. This section describes the protocols we followed to ensure the privacy of participants was protected.

Despite the large number of devices, no audio or video was recorded throughout the experiment. However, one of the devices collected still images every 30 seconds from the point of view of the person. While we did not anticipate that one image every 30 seconds would capture much sensitive information, participants were given the opportunity to delete any photo they considered inappropriate after reviewing their images at the end of each day. The rest of the sensor data (e.g., physiological, contextual and behavioral) were anonymized at the time of data collection and were locally stored on each of the devices during the day and transferred to a password-protected computer during the night. To ensure maximum privacy, none of the sensors were connected to the Internet during the study, and participants were not allowed to use the devices for any other purpose besides the ones of the study.

Participants were also instructed to only wear the devices while at work and to cover the camera if they considered it would compromise their privacy and/or the privacy of others (e.g., going to the restroom). Moreover, they were recommended to remove the devices if wearing them negatively affected their task performance (e.g., during a meeting, while commuting from one building to another, going to the gym). Finally, participants were also asked to post a sign on their office doors indicating that anonymized images inside the office could be used for research purposes and to contact the researchers if additional information was needed (see Appendix C for more details).
Table 3: Questions asked for each of the prompts and definitions of the used concepts

<table>
<thead>
<tr>
<th>Question and interface</th>
<th>End-points and definitions</th>
</tr>
</thead>
</table>
| **How were you feeling during the previous 5 minutes?** | **Very energetic/low energetic axis.** Energy or arousal is the physiological and psychological state of being awake or reactive to stimuli. Examples of emotions with high energy are angry or excited. Examples of emotions with low energy are relaxed or depressed. Note that high energy can be either positive or negative.  
**Very pleasant/very unpleasant axis.** Pleasantness is the attractiveness (positive valence) or averseness (negative valence) of an event, object, or situation. Examples of emotions with positive valence are happy or excited. Examples of emotions with negative valence are angry or depressed. Note that both negative and positive valence can have either higher or low energy. |
| **What situation best reflects your previous 5 minutes?** | **High/low demands.** Demands are those physical, psychological, social or organizational aspects of the job that require sustained physical and/or psychological (cognitive and emotional) effort or skills and are therefore associated with certain physiological and/or psychological costs. Examples of high demands are high work pressure, unfavorable physical environment, and emotionally demanding interactions with other people. Note that job demands may not necessarily be negative.  
**High/low resources:** Resources are those physical, psychological, social, or organizational aspects of the job that are either functional in achieving work goals, reduce job demands and the associated physiological and psychological costs, stimulate personal growth, learning, and development. Examples of high resources are organizational support, performance feedback, good material, job autonomy, positive climate, etc. |
| **How stressed were you feeling during the previous 5 minutes?** | **Not at all/extremely stressed.** Stress can be defined as a reaction from a calm state to an excited state for the purpose of preserving the integrity of the organism. While there are different types of stress, we focus on the negative emotional feeling associated with work overload. Examples of potentially stressful situations are giving a speech, submitting a paper or disagreements with the boss. The middle point of the scale would be “Moderately.” |
| **Was this prompt disruptive?** | **Not at all/extremely disruptive.** A disruption is a major disturbance, something that changes your plans or interrupts some activity, event or process. Examples of disruptive interruptions could happen when interacting with others and/or interrupting the work flow. The middle point of the scale would be “Moderately.” |
4.6 Recruitment and Compensation

Participants of the study were members of a large technical research laboratory in the Massachusetts Institute of Technology. Participants were recruited over e-mail and the criteria for inclusion were the following: 1) skin free of damage at the sites of the electrodes, 2) no use of prescription glasses (contact lenses were ok), 3) being a non-smoker, and 4) remaining at the lab for most of the work day (around 8 hours). The requirement for no glasses ensured that the glasses did not add excessive weight to regular frames of glasses and minimize the burden of participants. The requirement for non-smokers was to minimize the existence of potential respiratory problems.

To minimize the amount of people dropping the study, we followed a scaled monetary reward system. The total compensation for completing the whole study was up to $200 (in the form of an Amazon gift card). The payments were distributed as follows: $15 for the 1st day, $25 for the 2nd day, $35 for the 3rd day, $45 for the 4th day, and $55 for the 5th day. Additionally, there was a bonus of $25 for completing the whole study successfully.

4.7 Preliminary Data Overview

Fifteen participants (7 females and 8 males) completed the whole study successfully. Another participant started the experiment but, after a few hours of wearing the devices, decided to drop the study. The collected information of this participant was excluded from the analysis. More information about this case will be provided in the following chapter.

The average age of participants was 29.66 years (Standard Deviation = 6.42) with a minimum of 18.70 and a maximum of 41.95 years old. The average weight was 153.13 pounds (STD = 44.31) with a minimum of 90 and a maximum of 260 pounds. The average height was 5.57 feet (STD = 0.39) with a minimum of 5 and a maximum of 6.33 feet. While most of the people did not have any diagnosed cardiac, respiratory or musculoskeletal problems at the beginning of the study, two of the participants started to frequently cough during the study. One of the participants was later diagnosed with pneumonia. Despite the coughing, both of the participants desired to continue with the study. All of the data were collected in a period of ten consecutive days. While most participants completed the experiment in five consecutive work days, a few of them (such as those with coughing
problems) had to take one or two days off to recover and/or attend to unexpected business out of the work environment.

Thirteen out of the 15 participants were graduate students, one was a research assistant, and another one was an administrator. When participants were asked to briefly describe their work environment, most of the participants mentioned spending large amounts of time in front of the computer writing text and code, responding to e-mails, and making phone calls. Some of the participants also reported working on electronics and performing laboratory assays. Their work occurred in closed office spaces (with none or two other office mates) and shared collaborative spaces. There were also a large number of work meetings, casual social interactions, and classes. Some of the main stressors described during the study included: final exams, paper deadlines and reviews, financial issues, and public presentations to sponsors and colleagues. Each of the following chapters will provide additional information about the relevant aspects of the studied population.
Chapter 5

Wearable Experience Sampling

This thesis considers self-reports as the gold standard measure of stressful experience. To effectively gather this type of information without disrupting daily life researchers have widely used the Experience Sampling Method (ESM). While the most recent approaches rely on using a cellphone to trigger prompts and record information, wearable devices now offer new opportunities for improving this method. This chapter studies the effects of using three classes of wearable and mobile devices for the purpose of ESM. In particular, we developed an Android ESM tool that can be similarly used on different types of devices (e.g., smartwatches, head-mounted devices, smartphones) and quantify how the device form-factor can significantly impact relevant aspects of the reporting process (e.g., amount of missed prompts and response times). The chapter is divided as follows. First, we review some of the advantages and disadvantages of considering different locations for ESM. Second, we describe the proposed ESM tool, the interactions, and prompting criteria. Third, we provide quantitative and qualitative analysis of the main study to compare the different devices. Finally, we provide some discussion and concluding remarks.

5.1 Considered Locations

In this thesis, we wanted to explore the potential advantages of using different types of wearable devices in the context of experience sampling. In particular, we compare the traditional "phone inside the pocket" approach against wearable devices that are carried on different body locations (the wrist and the head). While there are many other body parts that we could have considered, we believe these two locations represent separate trends for
wearables that could offer different benefits for ESM. Below we highlight the main characteristics for the wrist-worn and head-worn locations.

- **Wrist-worn.** An ongoing trend for wearable devices is the development of wrist-worn devices that closely resemble bracelets or watches. We use the Gear Live smartwatch (henceforth referred to as Gear) which is equipped with many sensors similar to those on a smartphone. This type of device offers a unique opportunity to provide information to the users in an easily accessible and concealable location. While this type of device is not as widespread as cellphones are, many people are already used to wearing a traditional watch which potentially minimizes the burden associated with wearing them. The interactions with the Gear are very similar to the ones with traditional phones (e.g., touch surface and display to receive and provide information) but the screen size of 1.63-inch (320x320 pixels) is fundamentally limited by its body location. In contrast, the smartphone we use in this study (Samsung Galaxy S4) has a screen size of 5-inch (1920x1080 pixels).

- **Head-worn.** A relatively new trend that is quickly growing in recent years is the integration of head-worn devices such as the Oculus Rift or the Microsoft Hololens. In this chapter we explore this technology through the use of the Google Glass head-mounted device (Glass) which, among other sensors, includes a see-through display located just above the right eye (640x360 pixels), a touch surface on the right side of glasses, and a bone-conductive speaker above the right ear. This type of device offers unique opportunities to not only provide quicker and more intimate information to the user, but also capture insightful behavioral information (e.g., eye gaze, head gestures, facial expressions). As this wearable form-factor is relatively new and unfamiliar to users when compared to smartphones or smartwatches, the types of interactions are less well-defined and standardized. However, the device offers the opportunity to explore new types of interactions, such as head gestures.
5.2 Experience Sampling Tool

While there exists a great range of ESM tools available in the market, they are mostly designed to be deployed on smartphone platforms. Therefore, we designed and developed a novel ESM application that could be deployed on different types of devices, ensuring the maximum level of control and aesthetic similarity. This section provides more details about the application and the main differences across the devices.

5.2.1 Implementation

We used the Android development platform to create our ESM tool. Android is already being used in a wide variety of devices, such as smartphones and wearable devices, and offers the benefit of significantly reducing the costs associated with multi-device support. Our tool included two of the most commonly used types of questions in behavioral and psychological sciences: Grid and Likert scales questions.

- **2D-Grid.** This type of question enables the user to point at a specific location on a 2D-Grid to select his/her preferred choice. In the case of our application, we created two different grid questions. The first one asked the user to report their affective state in terms of emotional valence (x-axis) and arousal (y-axis), which is based on the commonly used Circumplex model of emotions (Russell, 1980). The second asked the user to report their current job demands (x-axis) and resources (y-axis), which is based on the Job Demands-Resources model to quantify the wellness of the work environment (Demerouti, 2001; Bakker & Demerouti, 2007).

- **Likert-scale.** This type of question enables the user to pick a point on a rating-scale. In the case of our application, we used a 5-point Likert scale to ensure the different options
could be easily read and accessed on the smaller smartwatch screen. For the purpose of our study, we created two Likert scale questions. The first question was: "How stressed are you feeling right now?" and the second question was: "How disruptive was this prompt?" Both Likert scales had end points “Not at all” and “Extremely.” Only the end points were labeled.

Every time the application triggered a prompt, the questions were presented in the same order (as shown in Figure 17), facilitating the ease of use.

5.2.2 Interaction

While the underlying implementation and appearance was the same across devices, the specific form-factors of each wearable device resulted in different interaction patterns. This was especially true for the head-mounted device which was the least familiar to the participants. This section describes the interactions performed by users in order to report their answers.

- **2D-Grid.** When using the smartphone and the smartwatch, the user touched their display to select the preferred answer. While the user continues to touch the display, a 3-second countdown starts and a progress circle is shown around the finger of the user. At the end of 3 seconds, the answer is automatically submitted. If the user interrupts the countdown and stops touching the display, the progress circle will disappear and a small white smaller circle will appear on the latest touched point, indicating that the countdown has been reset. The time of the countdown can be easily configured but we found 3 seconds was enough to prevent accidental submissions, especially when taking the phone from the pocket. While Glass also provides a display, its touch pad only works on one dimension, which does not easily allow for pointing on a 2D-Grid. However, due to its location on the head and onboard motion sensors, the device offers the opportunity of using head gestures to point and report users’ selections. Therefore, we used the Android libraries to estimate the pitch and yaw of the head and mapped the head orientation in real-time to a virtual pointer on the display. To start the response process, the user needs to tap with one finger on the side of the Glass and the virtual pointer would appear in the middle of the 2D-Grid. Then, head movements are
intuitively associated with the movement of the pointer. During this process, the user can tap again in order to re-center the pointer. Finally, the response is submitted by tapping with two fingers simultaneously. Both the one finger and two finger tap interactions were pre-defined gestures in Glass; the selection of which one was used to re-center the pointer and which one to submit the response was made to minimize accidental submissions (the two finger tap was deemed less likely to happen accidentally). To further minimize accidental submission, one finger and two finger taps needed to be 1 second apart in order to be registered.

- **Likert-scale.** When using the smartphone and the smartwatch, the user similarly uses touch interactions to pre-select the preferred answer and then a virtual submit button to complete the process. In the case of Glass, the user would need to swipe one finger forward/backward over the touch pad of the device to select different answers. To submit the final answer, the user would similarly use the two finger tap gesture. While we could have also used the head gestures for this type of question, we decided to only use finger swiping gestures so we could more closely resemble the interaction with other devices and better study the different types of interactions.

Every new question and user action triggers an auditory (Glass) or haptic (smartphone and smartwatch) feedback, which are designed to be as unobtrusive as possible while still being noticeable. Once a prompt is triggered, the application automatically turns the display on and a subtle notification appears. If the prompt is not answered, this notification is repeated every 30 seconds to ensure the user notices the prompt or did not forget to respond to it. If the prompt is not answered within a pre-defined amount of time (3 minutes in our study), it is automatically dismissed and the display of the device is automatically turned off. If the person is too busy to respond to the prompt, s/he can delay it by a certain amount of time (5 minutes). To do so, the user needs to either click the physical button located on the side of the watch, the volume down button of the phone, or swipe two fingers down on the Glass (as depicted on Figure 18). The application also records time stamps for every user interaction and other relevant events such as the time when the prompt is triggered, the
time when the first user interaction occurs, and the time when the reports are provided. Table 4 provides an overview of the main interaction differences for each device.

![Image of a phone, watch, and glasses]

Figure 18: Prompts could be postponed for 5 minutes by pressing the volume down button of the phone (left), the side button of the watch (middle), and by swiping two fingers down on the side of the glasses (right).

<table>
<thead>
<tr>
<th>Device</th>
<th>Display Shape</th>
<th>Feedback</th>
<th>Location</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone</td>
<td>Rectangle</td>
<td>Haptic</td>
<td>Pocket</td>
<td>Finger touches</td>
</tr>
<tr>
<td>Watch</td>
<td>Square</td>
<td>Haptic</td>
<td>Wrist</td>
<td>Finger touches</td>
</tr>
<tr>
<td>Glass</td>
<td>Rectangle</td>
<td>Auditory</td>
<td>Head</td>
<td>Head gestures, Finger swipes</td>
</tr>
</tbody>
</table>

### 5.2.3 Prompting Criteria

Deciding when to trigger a prompt to collect information is directly influenced by the purpose of the study and potential use of the reports. In our study, we wanted to gather information throughout the day to capture stress fluctuations as well as compare how people reported on each of the three devices. Therefore, we decided to trigger the prompts at random times and devices, which also minimized users’ anticipation for the prompts.

To ensure the different devices would not prompt at the same time, we pre-generated files with the triggering times of each device and transferred them to each of the devices before the study. Then, the ESM tool automatically loads them and uses them to trigger the prompts.
In our study, the triggering times were automatically generated with a custom-made MATLAB script that followed the following constraints:

- **Time distribution.** The time between prompts needs to follow a uniform distribution between 30 minutes and 60 minutes when considering all the devices together. Therefore, the user should not get more than one prompt in a 30 minute period and should receive approximately one prompt every 45 minutes.

- **Time variability.** The standard deviation of the triggering time for each device has to be at least 3 hours ensuring the prompts are distributed throughout the day.

- **Device variability.** No more than two consecutive prompts can happen on the same device in order to minimize the anticipation of users.

The seeds of the random numbers were also randomized to ensure the orderings changed every time MATLAB was restarted. Figure 19 shows a representative example of triggering times for one day.

![Figure 19: Example of daily scheduled prompts for the three devices of one of the participants](image_url)
5.3 Results

During the whole duration of the experiment, there were a total of 627 prompts triggered across all participants and devices. This section systematically compares the differences across devices.

5.3.1 Response Rates

As mentioned earlier, one of the challenges of ESM is to ensure that participants do not unnecessarily miss prompts because they were unnoticed and/or because they were too busy. This section further explores whether different form-factors yielded different response rates.

The left and right graphs of Figure 20 show the average number of prompts triggered throughout the study and the percentage of these that were successfully answered, respectively. The graphs were computed by calculating the average for each participant and device separately and then aggregating the ones corresponding to each device. Green lines indicate the standard error across participants. As can be seen on the left graph, even though the timing of the interruptions was designed to be uniformly distributed across devices, there were significantly fewer watch prompts than with the other devices (Two-sample t-Test; t(43) = 2.98, p = 005). This difference is due to the amount of time each of the devices was working during the experiment, which is affected by several factors. At the beginning of the study, participants were instructed to wear the three ESM devices throughout the work day and charge them during lunch time. However, due to the unpredictability of their schedules (e.g., meetings during lunch, late breakfast), many participants ended up charging the sensors only when the low battery warning was triggered, resulting in longer charging periods. This problem occurred more frequently for the Gear as its battery life was more limited. Moreover, we noticed that some skin moisturizers sometimes would transfer to the bottom of the watch, partially occluding some of the charging ports and, consequently, preventing some of the devices to properly charge.

When examining the number of prompts that were successfully answered, we can observe that the total number of answers was very high when considering all the devices (82.30%), yielding a total of 111 unanswered prompts. Furthermore, the Glass and the Gear prompts were answered significantly more often than the phone prompts (around 13%
more, \( t(43) = 2.71, p = 0.009 \). At the end of the study, the majority of participants reported feeling comfortable with the number of prompts they received and mentioned that they could have probably provided more reports before feeling disrupted. These comments suggest that the differences in terms of response rates were mainly due to participants not noticing some of them, especially on the phone. This finding is consistent with the hypothesis that both the Glass and the Gear are in closer contact with the body and, therefore, offer a more direct access to the attention of the person, irrespective of the feedback method (auditory and haptic, respectively).

Figure 20: (Left) Average number of prompts received for each participant across devices, and (right) percentage of these prompts that were answered. Green lines indicate standard error.

Figure 21: (Left) Average total response time for the prompts with each device, (middle) average time between the triggering of the prompt and the first user interaction, and (right) average time between the first user interaction and the submission of the final answer.
5.3.2 Response Times

The response time of a prompt captures a tradeoff between the burden imposed on a participant and the amount of data a researcher can collect, which are both critical to the design of successful ESM studies. In this section we present an analysis of how the response times varied across the devices.

Figure 21 shows the average number of seconds between the triggering of the prompts and their completion (left), between the triggering of the prompt and the first interaction of the user with the application (center), and between the first interaction and the completion of the questions (right). Similarly, the information was computed for each participant and then aggregated for each of the devices.

As can be seen on the first graph of Figure 21, participants took significantly less time to report on the Gear (42 seconds, t(43) = 2.43, p = 0.019) and around the same amount of time with the Glass and the phone (52 seconds). When considering the different parts of the response (center and right graphs), we can see that phone prompts took around 14 more seconds on average than the other two devices (t(43) = 3.67, p<0.001) before the first interaction happened. While the location of the device on a more peripheral location played an important role, it is important to note that the time to fetch a phone inside the pocket is significantly smaller than 14 seconds. However, this time captures the amount of time it took to participants to stop their daily activity and focus on answering the prompt, the phone having taken the longest. While 14 seconds may not seem very long, many things can happen in a few seconds during daily life. Indeed, one of the participants reported that something very stressful happened between receiving the notification and starting to respond, which added confusion when completing his stress report. Therefore, minimizing the time needed to access the device is critical to ensure the highest quality of the response. These differences were further supported by some of the comments gathered at the end of the study (e.g., “The phone was by far the worst, because I had to take it out of my pocket,” “Taking the phone out from the pocket was cumbersome”).

Finally, the third graph of Figure 21 shows that responding to the questions with Glass took significantly longer than with the other two devices (around 14 seconds on average, t(43) = 5.35, p<0.001). Feedback provided by the participants suggested that this was partly due to the familiarity of the Gear and phone, while the new types of interactions on
Glass slowed down the interaction. One of the participants explained, “I had to learn how to use the glasses – I was not used to the head gestures.” To further explore this, Figure 22 shows the average total response times of the first two days versus the last two days for the three devices. As can be seen, participants were faster providing their reports by the end of the study, especially for the Watch ($t(28) = 1.76$, $p = 0.09$) and significantly more for the Glass ($t(28) = 2.55$, $p = 0.017$). While we incorporated a practice session at the beginning of the experiment to minimize this effect, longer tutorial sessions or trial lengths could have helped to address this. Note, however, that there may also be some learning effects associated with the type of questions we asked, since people usually tend to get faster at reflecting and reporting their emotional states over time.

![Figure 22](image)

Figure 22: (Left) Average total response time for the prompts with each device when considering the first two days versus the last two days of the study.

To better understand the challenges of interactions, participants were also asked to report on a 5-Likert scale how easy it was to interact with each of the devices (with end points “Very challenging” and “Very easy”) at the end of the study (first column of Figure 25). Indeed, their responses were inversely correlated with the time it took them to use each of the devices. While device familiarity is an important factor to explain this difference, some participants also experienced unexpected problems with some of the designed interactions. For instance, some participants experienced problems with the two finger tapping gesture to submit responses on Glass. One participant reported “I had problems with the double tapping because of my hair. I realized that one finger was tapping over the hair and the device only detected one finger,” and another participant reported similar problems in
which one of the fingers was placed on the non-touch sensitive area (close to the display). These comments highlight some of the challenges associated with the design of devices that can work for a wide variety of populations and how hair length, in this case, can be an important design factor when designing head-mounted interactions. Some participants also reported experiencing difficulties when responding to 2D-Grid questions on the Gear. In particular, a participant stated “I found it difficult to point out things [on the Gear] because my finger was on top of it” and another “The watch was the hardest [to point] because my finger may be too fat,” highlighting that finger and screen sizes are important design factors when designing smartwatch interactions.

Overall, these results illustrate how analyzing the different parts of response time can yield different conclusions and how different factors such as location, familiarity, and types of interaction can influence the response time.

5.3.3 Response Distribution

In this section we present analysis of how the provided answers varied across the devices. Figure 24 shows boxplots of the answers for all the questions across each of the devices. Note that the 2D-Grid questions were separated into two separate graphs as they captured information along 2 dimensions. As can be seen, there are some differences across questions, but there is high consistency within each of the questions. To further explore if there were significant differences, we performed a non-parametric Kruskal-Wallis test for each of the questions, and found that none of the comparisons rejected the null hypothesis that the responses belonged to the same distribution \((p>0.058)\), indicating that the distributions were not significantly different.

When more carefully inspecting the 2D-Grid responses (Figure 24 a-d), a small but relevant difference indicates that Gear reports may have used a smaller range of values. To amplify these differences, Figure 23 shows the average range of values for each of the devices and the two types of questions. As can be seen, the 2D-Grid reports on the Gear tended to use around 6% less of the range than the other two devices (Two-sample t-Test, \(t(43) = 1.42, p: 0.162\)). This finding is consistent with the previous comments about using the finger with limited screen sizes. While the distributions and ranges of the answers were not significantly different, it is also important to keep in mind the type of analysis that will
Figure 24: Boxplots for each of the responses across devices. The boxplot shows the maximum, minimum, median and 75th percentiles

Figure 23: Average range of responses for the 2D-Grid (left) and 5-Likert scale questions (right) across participants for each of the devices

be performed. For instance, if researchers are interested in the extreme points of the 2D-Grid questions (e.g., people with very low arousal and valence), they may need to correct for the screen size of different devices. The range differences were not observed in the Likert-type questions probably because it was easier to point at discrete answers and double check the answer before its submission, indicating that different types of interactions and questions can also help reduce the impact of the limited screen size.

At the end of the study, participants also provided ratings about how accurate they thought their reports on each device were (see Figure 25). While no significant differences were observed, participants thought that using the phone would probably yield more accurate results, followed by the Gear and then the Glass. As expected, these results are positively correlated with how difficult they thought the interactions were with each of the devices.

Overall, these findings seem to provide support that our ESM application was able to appropriately address the different types of device form-factors and interactions providing distributions of reports that did not significantly change across devices. This was especially
challenging for the Glass as the interactions with the device are less well-established than those with the other two devices.

5.3.4 Usability of Devices

An important factor to consider with ESM studies is whether the prompting devices can not only effectively capture information but also be useful and wearable during daily activities. To further explore the users' feelings towards each of the devices, participants were asked to report any problems they encountered on a daily basis and to respond to a usability survey at the end of the study (see Appendix B for more details).

Figure 25 shows the distribution of responses for each of the usability questions. In particular, participants were asked to rate whether the devices were comfortable to use, whether they affected their social interactions, whether wearing the device increased their stress levels, and whether they would continue using the device during their daily lives. As
can be seen, both the Gear and the phone received more positive ratings than the Glass across all of the questions. This difference was more significant when reporting about the potential use of the device in the future.

When considering device comfort, the Gear received slightly better results than the phone, and the Glass scored below the average. However, when examining the distribution of responses, there appear to be two clusters of opinions. Among the participants who provided more negative scores, physical discomfort was a common concern. While we were anticipating it would take a few hours to get used to the new form-factor of Glass, especially with people who did not wear glasses normally, some of the problems persisted until the end of the study. Part of the discomfort was associated with the specific form-factor of the device, as several participants commented: “The Glass was a little bit tight on me”, “I did not like [the Glass] because it hurt me a lot. Maybe my ears had something weird...”, “[The Glass] is painful, I wear glasses sometimes and they're not that uncomfortable.” Two of the participants personally addressed this problem by adding some soft padding around the ear on the side where most of the electronics were housed. Some other participants also mentioned that part of the discomfort was associated with the reduction in the field of vision, with a participant stating “I do not like the idea of [the Glass] sitting in your peripheral vision,” and another one “The main problem of [the Glass] is that I was losing the upper part of my view... if anything, I would like wearables that increase my field of view.” During the first day of study, one participant also reported having headaches due to the Glass. However, this problem did not occur during the following days. The participants who provided more positive scores in terms of comfort did not seem to experience this type of problem and one of them reported “Surprisingly [the Glass] was not uncomfortable,” and another one “I got used to having [the Glass] on my face.”

A similar split was observed when participants were asked whether the Glass affected (either positively or negatively) their social interactions. In this case, several participants expressed concerns about how they looked and how other people felt about them. One participant explained “Among all the devices, I did not like the Glass because everyone knows you're wearing it,” and another mentioned “When I was by myself I forgot about [the glasses] but then I became self-conscious about them when I was walking around.”
One participant further explained that part of the problem was that wearing the Glass sparked unpleasant conversations about privacy (e.g., “People would feel I was taking pictures of them [with the Glass] and did not enjoy the conversation.” “I went to several meetings with other people, and I would take off the [Narrative] clip and the Glass because I did not enjoy being the topic of conversation”). Indeed, the only participant who dropped the study after a few hours explained that he could not afford having those types of conversations, as his job required him to meet new people very frequently. The participants that provided the more positive ratings seemed to enjoy this type of conversations. One of the participants mentioned “[The Glass] was a nice ice breaker,” and another explained “[Wearing Glass] was a good opportunity to speak about self-tracking.” In either case it is clear that the Glass made people more self-conscious. Several participants mentioned that receiving notifications on the devices during social interactions was very disruptive, especially for the case of Glass. One participant further explained “I found it very annoying receiving notification through Glass when speaking with people because it was so noticeable.” Finally, two participants reported feeling more distant during social interactions due to the specific form-factors of Glass (e.g., “[The Glass] sort of creates distance between the person you are communicating with and yourself”).

Both device comfort and potential impact on social interactions could partly explain the negative Glass ratings for future use as well as elicited stress due to the device. While the Gear and the phone received more positive comments overall, participants also experienced some problems with them especially when considering the female population. Even though the band of the Gear was adjustable, even the smaller possible configuration was still too large for four of our female participants. Moreover, two of them thought the device was too heavy for their wrist. On the other hand, the same band felt a bit tight for two of the male participants. With regard to the phone, several female participants mentioned they did not enjoy wearing it inside the front pocket of their trousers. One participant stated: “Phones inside the front pocket are not enjoyable as women’s trousers are not designed for it.” While different device form-factors (e.g., elastic bands, smaller phones) could have partially addressed the problems, these are still important design considerations when using different devices in ESM studies.
5.3.5 Visual Context

While not directly associated with the prompts and ESM tools, this thesis also explores using a wearable camera to capture daily activity and contextualize longer questions at the end of day. This section provides similar qualitative comments about how people felt with the camera and the reviewing of the photos. Figure 26 shows the distributions of the qualitative responses.

At the end of each day of data collection, participants had to review their daily images and use them to further annotate 5-minute regions during the day (before each of the triggered prompts) and daily events (e.g., relevant problems with the sensors, caffeine intake). When participants were asked whether they found the images useful to annotate the data, the majority provided very positive ratings (4.5 on average in a 5-Likert scale with end points “Not at all” and “Extremely”). This is partly to be expected as the 5-minute regions were randomly distributed across the day which did not necessarily capture salient events of the day. However, it is important to note that participants were not offered any other contextual cue as an alternative method so we cannot quantitatively compare these ratings with other types of data (e.g., phone calls, calendar events). Overall, participants found the device to be very comfortable (average score of 4.3) and did not think it added additional stress (average score 2). When asked about potential social stigma and potential future use, the average response was around the middle point. In contrast with the Glass, some participants mentioned that other people did not notice the wearable camera which could be due to its reduced size and/or the fact they were wearing more noticeable devices. When participants were asked whether they were concerned the device would capture private information/moments, their responses were also around the middle point. There were three instances throughout the study, however, in which some participants contacted the researcher because they believed the camera captured some sensitive information (private e-mails, restroom, and interactions with some people). It is important to note that participants were offered the possibility to delete sensitive photos at the end of each day which could have positively affected these ratings.

While this work uses the wearable camera to mainly enhance the memory of participants, this type of device also offers the opportunity to help people reflect and learn new things about their daily activity. To further explore this idea, we asked participants
whether they learned anything new with the images. A large majority of participants found
the images to be a great method for self-reflection. One participant reported “[Reviewing
the images] was a nice way to reflect on the moments of the day that I was most productive,
see with what people I did not interact during the day, check the places that I visited more
frequently, and be reminded of small details of the meetings.” Some of the comments also
highlighted that the images were used to quantify their amount and quality of social
interactions. For instance, one participant mentioned “I learned that I have more social
interaction during the day than I thought I had” and another remembered “During one of
the days, I had a meeting with a collaborator and I realized that the camera did not capture
him. This made me think that I may not have addressed him properly and that I should have
faced him more.” Some other participants mentioned that reviewing the images affected
their perception of productivity (e.g., “My idea of how the day went was not what I
perceived, after seeing the photos I felt I was more productive,” “I realized I spent less
time working that I might have guessed”). There were also some comments about how
wearing the devices influenced some of their daily behavior. For instance, one participant
mentioned “I think I might have gone less to the restroom to not to have to remove the
devices,” and another one stated “Wearing the camera made me immensely productive.
Instead of looking at e-mail during the day, I checked them first thing in the morning and
at the end of the day [when I was not wearing the sensors]... so those moments would not
be captured.” Finally, there were also a minority of participants who reported to not learn
new things from the images, probably due to a combination of several factors such as their
type of work, the quality of the captured images, and their memory skills. For instance, one
participant mentioned “There was nothing new, just my workday routine,” another one
explained “Most of my work during the day looks exactly the same (staring at my screen),”
and “I do not think I learned anything that my calendar could not tell me.”

5.4 Discussion

The previous section compared relevant factors of ESM across different wearable devices
in the context of a real-life study.

Among some of the main findings we found that phones, which are commonly used in
ESM, received the more positive scores in terms of ease of interaction, potential future use,
and perceived accuracy of the reports. However, participants of our study missed a significant number of prompts and took significantly longer to start interacting with the application once the prompt was triggered. These results are in contrast with the Glass and Gear, which significantly reduced the number of missed interruptions and the time before the first interaction. To interact with the Glass, we used novel head and finger gestures which did not affect the distributions of the answers but significantly increased the interaction time. The specific form-factor of the Glass also received some negative scores in terms of potential future use and participants highlighted different scenarios in which the device could have added some unnecessary stress, especially during social interactions. The Gear received more positive scores across the different usability factors and slightly outperformed the phone in terms of device comfort. However, some specific form-factors such as the limited screen size may influence (although not significantly) the range of the responses.

In the context of stress measurement (the main purpose of this thesis), the Gear seemed to prevail. This device not only enabled effective and quick gathering of self-reports during the day but also minimized the burden of participants with a concealable and accessible form-factor that did not artificially increase stress levels. However, it is important to note that different types of devices may be more appropriate in different experimental conditions. For instance, certain work environments may not involve social interactions which would minimize part of the social stigma associated with head-mounted wearable devices. Also, if the reporting tasks are limited by a small screen (such as reporting on a 2D-Grid) then a smartphone or Glass device may be preferable. Certain demographics may also be reluctant to wearing some types of devices in certain locations (e.g., phones inside the pocket for female participants). Finally, this study explored the use of head gestures and touch interactions. However, different findings may be obtained when considering different types of interactions. For instance, using voice as an input mechanism may be more invasive but also may yield similar response times across the three devices.

This chapter evaluated the ESM tool in a population of 15 people during five work days yielding a total of 627 prompts. While we still obtained some significant findings, the sample size and the demographics of our participants may not necessarily generalize to other populations. In our experiment, all of the participants were part of a technical research
institution and, therefore, they tended to be more familiar with the state-of-the-art of technology. Because of this, participants may have been more open or critical to certain aspects of the devices which could explain the divided opinions for some of the questions. Participants of this experiment were also requested to not wear eyeglasses during the study in order to minimize the burden associated with them. As a result, only two out of the 15 participants used regular eyeglasses during their regular daily life which could have biased some of the qualitative Glass ratings. Indeed, two of the participants who provided positive ratings about the Glass comfort were used to wearing glasses during their daily lives. Similarly, only three of the participants were used to wearing a watch during their daily lives. Another limiting factor of our study is that participants were not allowed to use the devices for their personal use. While this is a reasonable request when providing expensive ESM tools to participants, it can also affect the perceived utility and potential use of the devices. For instance, if participants could have used the Glass to take photos or receive phone calls, they could have provided more positive scores. Indeed, one participant noted “I did not like that I could not use the wearable devices. Without using them, it is difficult to know whether I would wear them during daily life.” Finally, a few months before the study took place, the Glass device received some media criticism due to its onboard camera and its potential to invade the privacy of others (e.g., taking photos without people awareness). While we disabled the camera for the purpose of the study, some participants were still concerned about this. Indeed, one of the participants described “I felt very embarrassed with the Glass probably because there is a lot of baggage with it.” Future studies with larger sample sizes, more varied populations and potentially other device form-factors will help quantify the generalization of our results.

Overall, this chapter has shown that there is not a one-fits-all solution for ESM. Different devices offer different benefits which may make them more adequate for different scenarios. An interesting area of future work would be to systematically explore what specific scenarios may make more sense to use one or another form-factor. Moreover, it is very probable that people may be carrying more than one wearable device in the future, creating new opportunities to explore a multi-device ESM that more effectively lowers the burden of participants. For instance, if a person is intensively using his/her hands on a specific task (e.g., manufacturing, writing a thesis), the system could deliver the prompt
through the Glass and use voice interactions to completely free the hands. However, if a person is speaking with someone and/or is surrounded by other people (e.g., in a classroom), more concealed devices that deliver more unobtrusive prompts may be preferred. Finally, we would also expect that different people may prefer different prompting channels (e.g., people already wearing glasses may prefer the head-worn form-factor) and, therefore, learning and adapting to users’ preferences can be an active field of research to further enhance the ESM.

5.5 Conclusions

This chapter proposes an application to perform experience sampling specifically designed to work on smartphones, wrist-worn and head-worn Android devices. With a real-life deployment, we have demonstrated that each device form-factor offers unique opportunities to improve the ESM process. We have highlighted several critical design factors as well as areas of potential improvement. With the continuous development of technology, we expect these devices will become more ubiquitous, creating new opportunities to effortlessly capture ambulatory self-reports and improve scientific discovery.
Chapter 6

Physiology Parameter Estimation with Wearable Motion-based Sensors

The stress response has a well-studied set of physiological changes that can be monitored for the purpose of stress measurement. Traditional approaches to measure parameters such as heart and breathing rates in the midst of daily activity require attaching electrodes to the skin or strapping a band around the chest, which is cumbersome for daily life monitoring. However, recent advances in technologies have enabled the creation of comfortable wearable devices that are in close contact with the body and can run continuously during daily activities. This chapter presents a series of mostly controlled experiments studying how low-cost motion sensors located on different peripheral body locations (wrist, head and pocket) can be used to capture subtle motions associated with the beating of the heart (also known as ballistocardiography; BCG) and with respiration. We show that we can create an automated algorithm that often derives accurate estimates of heart and breathing rates from these subtle motions, provided that the person is reasonably “still.” The chapter is divided as follows. First, we start by describing the controlled experiment. Then we sequentially motivate, propose and evaluate the methods, and discuss the main findings for each of the considered locations. Finally, we conclude the chapter with an overview of the main findings and their potential implications.
6.1 Controlled Experimental Details

To study each of the locations we performed several laboratory experiments with subtle differences for each. This section describes the experimental details that were common across the studies.

6.1.1 Apparatus

Among all of the sensing capabilities of wearable devices, motion sensors such as accelerometers are arguably the most pervasive ones. Most currently available wearable devices contain this type of sensor to help capture and understand behavior (e.g., steps, activities) and potential sources of artifacts for other sensing modalities (e.g., electrodermal sensors). This work studies a much less explored application in which the person wearing the motion sensors appears to be “still” and relevant motions are not easily observable to the eye.

While the form-factor of the considered wearable devices changed from head to wrist to smartphone, we mainly used two types of motions sensors (accelerometers and gyroscopes) to capture vital signs and compared their measurements with FDA-cleared physiological sensors. The following provides more details about each category of sensors.

Motion Sensors

Accelerometer. The 3-axis accelerometer captures the acceleration applied to a specific device (meters/second²) along the X, Y and Z axes. Accelerometer measurements include the force of gravity and have been widely used in the context of activity recognition (Lara & Labrador, 2013). Furthermore, this sensor seems to be one of the preferred choices for measuring the BCG in different contexts (e.g., Dinh, 2011; He et al., 2012; Kown et al., 2011; Phan et al., 2008; Rienzo et al., 2012).

Gyroscope. The 3-axis gyroscope captures the rate of rotation (radians/second) of the specific device along the X, Y and Z axis of the device. Unlike an accelerometer, a gyroscope is not affected by gravity but may have a cumulative error (drift) over time. A common application of this sensor is the stabilization of aerial vehicles, but it can also be
combined with other sensors to provide accurate localization (e.g., Kothari et al., 2012). To the best of our knowledge this sensor has not been validated in the context of BCG measurement, probably due to the limited rotational motions of typically considered body locations (e.g., chest, ear).

To assess different sensor locations and devices, we used the Google Glass head-mounted wearable device, the Samsung Galaxy Gear smartwatch and a Samsung Galaxy S4 smartphone. To simultaneously log information from the motion sensors of the different wearable devices we created a custom Android application. The average sampling rates were 50 Hz for the Google Glass and 100 Hz for the other two devices. To ensure a constant sampling rate and consistency across studies, data from all devices were cubically interpolated at a sampling rate of 256 Hz. This frequency was the one used by most of the FDA-cleared physiological devices (see more below) and, therefore, enabled us to preserve the maximum amount of gold standard information.

**Physiological Sensors**

To obtain gold standard cardiac and respiratory information, we used the following FDA-cleared sensors.

**Cardiac activity.** For comparison with measures of BCG at the head, we used the FlexComp Infiniti by Thought Technologies that captured Blood Volume Pulse (BVP) from the finger at a constant sampling rate of 256 Hz. For comparison with wrist and other peripheral locations, we used the Alive Technologies sensor to measure ECG from single-lead chest-worn electrodes at a sampling rate of 300 Hz. Conductive gel was used to improve the signal quality. Additionally, for comparison with the wrist location, we also collected BVP with the wrist-worn wearable device Empatica E3 (www.empatica.com) at a sampling rate of 64 Hz.

**Respiratory activity.** For all the studies, we used the FlexComp Infiniti that captured respiration from a chest belt sensor at a constant sampling rate of 256 Hz.
6.1.2 Heart and Breathing Rate Estimation

To estimate heart rate (HR) from ECG measurements, we used the peak detector described in Pan and Tompkins (1985). The detected peaks were visually verified for each of the experiments in order to ensure a fair comparison with the measurements from the motion sensors. Heart rate was then computed as 60/(average number of seconds between peaks) beats per minute.

To estimate HR and breathing rate (BR) from the other sensors and modalities (e.g., BVP, respiration signals), we looked at their responses in the frequency domain. In particular, we extracted the frequency response with the Fast Fourier Transform and identified the frequency with the highest amplitude response. The band of frequencies used for HR were [0.75-2.5] Hz corresponding to 45 and 150 beats per minute, and the band of frequencies used for BR were [0.13-0.75] Hz corresponding to 8 and 45 breaths per minute. The final estimated HR and BR corresponded to the maximum frequency multiplied by 60 (beats and breaths per minute, respectively). Similar frequency analysis was used when estimating heart and breathing rates from motion data.

When computing physiological parameters from the motion signals, the methods that we propose provide cardiac and respiratory physiological waves similar to those of BVP and respiration signals. Therefore, we estimated heart and breathing rates in the frequency domain as well. Computing these parameters in the frequency domain instead of the time domain allowed us to 1) partially address the problem of missing peaks due to non-constant sampling rates of accelerometer and gyroscope, 2) deal with the non-linear phase responses of the filters we employ, and 3) avoid addressing the problem of peak detection with noisy signals.

6.1.3 Protocol Overview

For each of the three studies, we collected data from 12 participants (6 females and 6 males) in a controlled laboratory setting, yielding a total of 36 participants. After obtaining written consent, participants were asked to keep still and breathe spontaneously while remaining in three different positions (standing up, sitting down and supine) for a minute each. These positions have been shown to influence the subtle motions associated with the beating of
the heart and respiration and represent some of the most common postures that can be observed during daily life. To generate a larger dynamic range of physiological readings, participants were also asked to repeat the three positions after pedaling a stationary bike for one minute. Thus, this procedure yielded 72 minutes of well-characterized data over a range of heart and breathing conditions for each of the locations. To increase the number of samples, we divided the data into intervals of 20 seconds with a 75% overlap, yielding 648 data samples per considered location.

The previous experiment lasted around 25 minutes and participants received a $5 monetary compensation in the form of an Amazon gift card. While the main protocol was very similar when studying different locations, there were some significant differences that will be covered by each of the following sections. This experiment and following variations were approved by the Institutional Review Board of the Massachusetts Institute of Technology.

6.2 Head Motions

In this part of the analysis we focus on the motion-sensitive capabilities of a head-mounted wearable device. Google Glass is a wireless head-mounted device equipped with a touch pad, a see-through display, and most of the sensors available in smartphones (see Figure 27). Although the device was not designed for physiological measurement, its unique location on the head of the person provides an opportunity to unobtrusively monitor physiological information during daily activities. In particular, we develop methods allowing motion-sensitive sensors in Glass to be used to capture subtle head motions of the wearer that are associated with the mechanical activity of the heart and the respiration of the wearer.
6.2.1 Experimental Additions

Besides monitoring information from the accelerometer and the gyroscope, we also considered the use of a head-worn camera to monitor subtle periodic motions of the captured images. This is in contrast to the work of Balakrishnan et al. (2013), in which they measured the heart rate of a person in front of a static camera by monitoring subtle head motions. To the best of our knowledge, the work presented here is the first to use the egocentric view of a wearer to gather his or her own physiological data.

The video was recorded at a constant frame rate of 30 Hz at a resolution of 1280x720 pixels (the default settings of Glass). Each of the pixels yields a vector in RGB color space. To extract motion signals of the camera, we tracked distinctive 2D feature points in the video. To do so, we first detected feature points (Shi & Tomasi, 1994) in each frame and tracked them using a Kanade-Lucas-Tomasi feature tracker (1981). We then fit a homography matrix (Hartley & Zisserman, 2004) to the point correspondences using RANSAC (Fischler & Bolles, 1981). Note that we assume that all tracked points correspond to static 3D points, in which case their offsets are solely explained by the camera motion. Finally, the vertical and horizontal motion (up to a scale) of the camera were extracted from the matrix. While more accurate motion estimation methods exist (e.g., Hartley & Zisserman, 2004), they require addressing additional challenges such as calibrating the camera that may attenuate the physiological information.
6.2.2 Proposed Methods

A challenge in extracting physiological parameters during daily activity with wearable devices is to develop algorithms that require low-computational power and run in real-time. Therefore, we constrained our new methods to use combinations of efficient signal processing techniques. In this section, we provide details about the proposed processing steps to estimate the pulse and respiratory waves from a specific stream of sensor data.

**Pulse Wave**

Given a specific sensor modality with sensor readings as a time series of vectors (e.g., 3D vector for accelerometer and gyroscope, 2D vector for camera), the estimation of the pulse wave was divided into the following steps. First, a moving average window of 3 samples was subtracted from each dimension of the vector, allowing the removal of signal shifts and trends. Second, a band-pass Butterworth filter of order 4 with cut-off frequencies of 10 and 13 Hz was applied to each dimension to isolate BCG changes. To aggregate the different components of the signal, i.e., dimensions of the vector, we then computed the square root of the summation of the squared components (i.e., Euclidian norm) at each sample. This aggregation gives equal weight to each of the dimensions and makes our approach more robust to different body postures. Finally, a band-pass Butterworth filter of order 2 with cut-off frequencies of 0.75 and 2.5 Hz was applied, yielding the final pulse wave.

Figure 28 shows an example of an accurate pulse wave estimation from gyroscope data of a person wearing the head-mounted wearable device while lying down. The top graph shows the 3-axis gyroscope over a period of 5.5 seconds and clearly shows the BCG changes. The middle graph shows the pulse wave obtained by BVP (blue) and the pulse wave obtained after applying the described methods on the gyroscope data (red line). Despite the signal lag associated with the employed filters, we can see that the estimated pulse wave is very similar to the wave of reference and is able to capture the changes associated with the beats and their reflections.
Respiratory Wave

To estimate the respiratory signal from data of a specific sensor (the same as what we used to derive the pulse signal), we performed the following steps independently for each sensor modality. First, an averaging filter was applied to each of the components. The window length was set to be the duration of a respiration cycle at a maximum breathing rate (45 breaths per minute in our case). Second, a band-pass Butterworth filter of order 4 with cut-off frequencies of 0.13 and 0.75 Hz was applied to each dimension. Since different dimensions of the sensor reading (e.g., X and Y axis of accelerometer) may change in different directions depending on the body position, we applied Principal Component Analysis to reduce this influence. We then computed the Fast Fourier Transform of each
principal component and selected the most periodic signal, where periodicity of the signal was estimated by computing the maximum magnitude observed within the operational frequency range.

Figure 29 shows an example of an accurate respiratory wave estimation from accelerometer data of a supine participant. As in Figure 28, the two waves are closely aligned. The bottom graphs of Figure 28 and Figure 29 show the Fourier Spectrum over the whole 20 seconds for an estimated (left) and reference (right) signal. As can be seen, the frequency responses for the two waves are closely aligned and their peak frequency is approximately the same: 1.1 Hz (corresponding to a heart rate of 62 beats per minute) for the pulse waves, and 0.41 Hz (corresponding to a breathing rate of 25 breaths per minute) for the respiratory waves.
6.2.3 Results

In this section, we use the segmented data to compare performance across the three modalities and body postures. We then evaluate the benefit of combining the three modalities. Finally, we explore the effects of dividing the data into intervals of different durations.

Comparison across Modalities

To evaluate the utility of each sensor modality, we extracted heart and breathing rates from each of the samples and computed the same performance metrics used in (Poh et al., 2011). Figure 30 shows the Bland-Altman plots (Bland & Altman, 1986) for each physiological parameter using the different modalities. In particular, each graph shows the agreement of the 648 pairs of measurements color-coded by participant. The graphs also show the mean error and the 95% limits of agreement (i.e., 1.96 standard deviations above and below the mean). As can be observed, most of the graphs have a significant concentration of points along the zero values of the y-axis, illustrating a close agreement between the measurements. Table 5 and Table 6 show a summary of quantitative metrics for heart and breathing rates, respectively, across all the 648 segments (72 minutes) of data computed from the 12 participants. We observe the same trend across the three types of sensors for both heart rate and breathing rate: our estimation has a high accuracy in comparison to gold standard, while the errors vary across different sensors. The small mean errors from different modalities provide strong evidence to the feasibility of our approach. When comparing the three sensors individually, the gyroscope yielded the best performance for both heart rate and breathing rates, achieving a mean absolute error of 0.82 beats per minute (STD = 1.98) and 1.39 breaths per minute (STD = 2.27), respectively. Notably, the accelerometer (upon which prior BCG work is based) was never the best performing modality for heart rate or respiration. While the camera modality outperformed the accelerometer for estimating breathing rate (achieving a ME of 1.55 breaths per minute), its performance for heart rate estimation was the worst (ME = 7.92 beats per minute).
Table 5: Heart rate estimation with Glass

<table>
<thead>
<tr>
<th>Sensor</th>
<th>ME</th>
<th>STD</th>
<th>RMSE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>0.82</td>
<td>1.98</td>
<td>2.14</td>
<td>0.99</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>2.51</td>
<td>7.03</td>
<td>7.46</td>
<td>0.91</td>
</tr>
<tr>
<td>Camera</td>
<td>7.92</td>
<td>13.4</td>
<td>15.56</td>
<td>0.58</td>
</tr>
<tr>
<td>All</td>
<td>1.19</td>
<td>3.42</td>
<td>3.62</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 6: Breathing rate estimation with Glass

<table>
<thead>
<tr>
<th>Sensor</th>
<th>ME</th>
<th>STD</th>
<th>RMSE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>1.39</td>
<td>2.27</td>
<td>2.66</td>
<td>0.75</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>2.29</td>
<td>3.43</td>
<td>4.12</td>
<td>0.41</td>
</tr>
<tr>
<td>Camera</td>
<td>1.55</td>
<td>2.59</td>
<td>3.02</td>
<td>0.69</td>
</tr>
<tr>
<td>All</td>
<td>1.16</td>
<td>2.04</td>
<td>2.35</td>
<td>0.79</td>
</tr>
</tbody>
</table>

ME = Mean absolute error (beats/breaths per minute)
STD = Standard deviation of the absolute error
RMSE = Root mean squared error
CC = Pearson’s correlation coefficient

The process of extracting motion measurement from video is more complex and less direct than it is for the other two sensors. For instance, different factors such as the depth of the scene or the number of feature points that can be tracked in the environment have a direct impact on the performance. Furthermore, the sampling rate of the camera was significantly lower than the other two sensors. Therefore, high frequency changes such as subtle head movements due to heart activity may not be as accurately captured as the low frequency movements associated with respiration. While these results still show that it is possible to extract heart and respiratory information from a head-mounted camera, future work will need to focus on performing a more systematic comparison across different factors (e.g., motion estimation method, depth of the scene and objects, amount of tracked points).

Postural Changes

Body posture impacts the intensity and quality of BCG movements. In this study, participants were measured from three body postures: sitting, standing, and supine. Table 7 and Table 8 show the mean absolute error (beats and breaths per minute, respectively) for each of the different positions and sensors. When estimating heart rate, the most challenging position was sitting down, which is in accordance with the results described in He et al. (2012). The results obtained with the gyroscope in this study outperform the results of He et al. for the sitting (ME = 1.27) and supine (ME = 0.84) conditions but not for the standing (ME = 0.72) position. However, the range of heart rates they observed in their study was considerably smaller (55 to 95 beats per minute) in comparison to what we elicited (56 to 133 beats per minute). The different results for the camera sensor may be due to a combination of several factors such as the influence of body posture, the accuracy of motion estimation as well as the relative pose of the camera with respect to the wearer’s
head. When estimating breathing rate, the most challenging position was standing up for all the modalities. Apparently, respiratory movements have less influence on head motion while standing. Overall, even the most challenging posture positions yielded low error for the best-performing modality.

Figure 30: Bland-Altman plots for heart (top) and breathing rates (bottom) using gyroscope (left), accelerometer (center), and camera (right). Each graph shows the agreement of 648 pairs of measurements. Data from different participants are represented with dots of different colors. Mean error is depicted with slashed red and 95% limits are depicted with dashed green lines. (HR: Heart Rate, RR: Breathing Rate, GYR: Gyroscope, ACL: Accelerometer, RGB: Camera)

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Sitting</th>
<th>Standing</th>
<th>Supine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>1.18</td>
<td>0.85</td>
<td>0.44</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>3.30</td>
<td>2.18</td>
<td>2.06</td>
</tr>
<tr>
<td>Camera</td>
<td>4.51</td>
<td>10.17</td>
<td>9.10</td>
</tr>
<tr>
<td>All</td>
<td>1.48</td>
<td>1.17</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 7: Mean absolute error of heart rate with Glass (N = 216 per column)

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Sitting</th>
<th>Standing</th>
<th>Supine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>1.13</td>
<td>1.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>1.87</td>
<td>3.17</td>
<td>1.82</td>
</tr>
<tr>
<td>Camera</td>
<td>1.32</td>
<td>1.89</td>
<td>1.45</td>
</tr>
<tr>
<td>All</td>
<td>0.94</td>
<td>1.77</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 8: Mean absolute error of breathing rate with Glass (N = 216 per column)
**Combination of Modalities**

Differences in performance across modalities are partly due to the different types of information being captured by each of the sensors. For instance, while the accelerometer data captures linear accelerations, the gyroscope captures rotations of the device. Furthermore, while some of the sensors may be affected by sampling rate artifacts (e.g., accelerometer and gyroscope), other sensors may provide more constant sampling rates but less accurate information (e.g., camera). Therefore, a combination of different modalities may help to provide more reliable estimates at the cost of computational complexity. To explore this idea, we extracted the heart and breathing rates of each modality separately and computed the median as the final estimate. The bottom rows of Table 5 and Table 6 show the results using this aggregation technique. While the heart rate estimation using the gyroscope was still better than the one obtained combining all the modalities, the breathing rate estimation with all three modalities yielded better results than with any of the other modalities alone (reducing the ME to 1.16 breaths per minute, STD = 2.04). Although not explored in this chapter, we expect larger improvements by combining physiological signals in less controlled settings where different modalities might provide complementary information about motion.

**Observation Windows**

While combining several sensors may partially address the problem of motion artifacts, an underlying assumption of the proposed methods and evaluation is that the person is holding a motion-less position for the majority of a certain observation window (20 seconds for the previous results). However, being able to remain still for large periods of times during daily life may not be always possible and shorter and more available observation windows may be preferred. To explore the performance of our methods for different lengths of observation windows, we split the collected data into segments of different durations following the same criteria described above and assessed their performance. Figure 31 shows the number of samples obtained for each of the observation windows (left), and the absolute mean error of the different approaches to estimate heart rate (center) and breathing rate (right). As can be seen, heart rate can be computed with a ME of 4 beats per minute with an observation window of only 5 seconds from gyroscope. This error goes below 2
beats per minute for observation windows equal or larger than 10 seconds, reaching its minimum at 25 seconds. The decrease of performance with smaller windows is expected due to several factors. When computing heart and breathing rates, especially in the frequency domain, longer observation windows are preferred to provide more accurate estimates. Moreover, different body locations reflect physiological changes at different times which can bias our estimates.

Although we expected ME would always decrease with longer observation windows, there was a subtle rebound effect for windows above 30 seconds for the accelerometer and gyroscope data. Visual inspection suggested that this rebound was due to missing beats due to the non-uniform sampling rates of these two sensors, which became more significant when reducing the amount of samples. This effect was not observed when estimating breathing rate, where larger observation windows always improved performance. This finding is in accordance with the previous observation as breathing rate operates in a lower frequency range and the non-uniform sampling rate did not negatively affect the signal.

Using the longest observation window (60 seconds) and a combination of all the sensors, the ME was reduced to 0.6 breaths per minute (STD = 1.19).

Note that the aggregation of the three modalities measured by Glass performs more or less the same as the best modality when estimating the heart rate, and consistently leads to slightly better results than the individual modalities when estimating breathing rate. These results are promising for real-life monitoring where the head of the person is more likely to remain still for shorter periods of time than longer ones. In practice, the duration of the observation window and the accuracy is a trade-off that needs to be carefully chosen when deciding on a specific population and/or application. For instance, children may have more problems remaining still than adults and so that shorter and less accurate windows may be preferred. However, longer and more accurate readings may be more accessible and adequate in certain scenarios such as sleeping or practicing meditation.
6.2.4 Discussion

The previous sections have shown that it is possible to capture physiological parameters from acceleration, gyroscope and camera sensed from a head-mounted wearable while the user is relatively "still." The results from different modalities are consistent and the mean absolute errors are small, which further justify our methods. Furthermore, some of our results were improved by combining several modalities and changing the observation windows. While we expect the combination of sensors will yield improved assessments of physiological parameters during daily life, there are critical differences among sensors that need to be considered.

Both accelerometer and gyroscope require considerably less energy than the camera and, therefore, allow for longer periods of monitoring without charging the batteries. With the current version of Google Glass, we were able to continuously record gyroscope and accelerometer data for around 8 hours with an average sampling rate of 50 Hz. These two sensors directly capture complementary aspects of the wearer’s head motion. For instance, while driving a car, the accelerometer readings may be influenced by external forces such as changes of speed; however, the gyroscope will provide cleaner signals associated with the rotation of the head of the driver. In this specific case, the accelerometer and the gyroscope can provide meaningful information about both the context and the physiology. The camera requires significantly more energy (the current battery lasts for approximately 20 minutes of continuous monitoring), but also provides some critical benefits. The
location of the camera of the head-mounted device in this chapter is located above the right eye (see Figure 27). This setup offers the opportunity of capturing the environment from the wearer’s perspective. This information is useful not only for extracting physiological parameters as demonstrated in this part of the analysis, but also for measuring rich contextual information that can help to infer influences on the physiological responses. Furthermore, linking physiological information with visual context can be useful in a wide variety of applications such as catalyzing introspection (Hernandez et al., 2013), augmenting human memory (Hodges et al., 2006), and improving social communication (Marcu, Dey & Kiesler, 2012).

One of the main challenges when estimating physiological parameters from motion in real-life scenarios is the presence of large body motions due to physical activity. Daily activities such as walking or speaking with other people involve large body movements that might obscure or distort more subtle heart and respiratory motions. Although this study did not directly address this issue, we evaluated our methods for a large range of observation windows. For instance, with only a 5-second observation and using the gyroscope, we were able to provide estimates of heart rate with a ME of 4 beats per minute, which is adequate to provide a gross estimation of heart rate. This observation window is significantly smaller than windows reported on similar studies (e.g., >70 seconds in Balakrishan et al. (2013), >20 seconds in He et al. (2012), 30 seconds in Poh et al., (2010)).

An important area of research is to understand how often these observation windows are accessible during daily life activity. In a relevant study, Rienzo et al. (2012) monitored sternal seismocardiogram of 5 participants during 24 hours and found that there were more than 100 5-second segments per hour with good quality acceleration data during the day and three times higher during the night. These numbers were quickly reduced with longer observation windows. These results are promising for non-intrusive physiological assessments. However, the location and types of sensors of this study are different, and can have an impact on the statistics. Chapter 7 assesses the availability of good quality data captured with a head-mounted device and other relevant devices during daily life activity.
6.3 Wrist Motions

In recent years, researchers have been actively working on the development of less intrusive physiological sensors and the creation of wearable devices for daily life monitoring. Among the different form-factors, an ongoing trend is the development of wrist-worn devices (e.g., Empatica, MyBasis, Mio Alpha) that can monitor vital signs by non-intrusive methods such as photoplethysmography (Allen, 2007; PPG). Being able to measure vital signs from the wrist presents several benefits for daily life monitoring as the sensors are always in close contact with the person and can be easily accessed and concealed. As illustrated in Figure 32, this section explores using motion information on the wrist in order to derive measures of cardiac and respiratory activity. In particular we address the following research questions: “How can we use the currently available motion sensors within wrist-worn wearable devices, namely accelerometers and gyroscopes, to accurately measure cardiac and respiratory movements?”, “How do these results compare to traditional approaches and state-of-the-art wearable devices?”, “Does combining measurements from motion and other traditional approaches improve performance?”, and “How well do the proposed methods perform in a real-life setting to provide in-situ non-intrusive physiological assessments?”.

Figure 32: We present a novel approach for measuring cardiac and respiratory parameters from wrist motions using a smartwatch (Samsung Galaxy Gear), even when the wrist is not held against the chest.
6.3.1 Experimental Additions

In addition to the laboratory experiment (see section 6.1.3), a second naturalistic experiment was performed with three participants (ages 27, 28 and 30) wearing sensors during real nights of sleep. The participants were recorded for around 6 hours during each of two nights of sleep. ECG gold standard and wrist measurements (E3 and smartwatch) were time aligned by starting the recordings simultaneously (see Figure 33 for the setup). For this part of the analysis, only HR estimation was considered as we did not want to alter the sleeping environment with the wired respiration chest belt. Moreover, in order to minimize the effect of having novel wristband devices or ECG electrodes, participants slept with all the sensors for two days before the recordings took place.

![Figure 33: Setup for the data collection during the sleep experiment. ECG electrodes were attached to the skin near the heart and on the abdomen. The Galaxy Gear was worn on the left wrist.](image)

6.3.2 Proposed Methods

**Pulse and Respiratory Wave Estimation**

To extract HR and BR from a specific stream of motion data (e.g., 20 seconds of accelerometer or gyroscope data, or a combination of data from both sensors), several processing steps were followed. First, each of the components (e.g., X, Y, Z axis of the accelerometer) were normalized with z-scores in order to give them the same relevance. Then, for each sensor (e.g., accelerometer, gyroscope or a combination of sensors) we
estimated the pulse and respiratory waves from which heart and breathing rates could be easily calculated.

To estimate the pulse wave, we first applied an averaging filter and subtracted it from each of the components. The window duration of the filter was empirically set to 1/7th of a second which effectively removed signal shifts and trends due to body motion while preserving BCG information. Then, a band-pass Butterworth filter of order two, with high and low cut-off frequencies of 4 and 11 Hz respectively, was applied to isolate the BCG changes. The different components of each sensor were then aggregated with a square root of the summation of the squared components to make the estimations robust to different body postures. Finally, a band-pass Butterworth filter of order two with cut-off frequencies of 0.66 and 2.5 Hz was applied to obtain the final pulse wave.

To estimate the respiratory wave, we first applied an averaging filter to each of the components independently. The size of the window was set to be the duration of a respiration cycle at a pre-defined maximum breathing rate of 40 breaths per minute, which enables removing the higher frequency cardiac motions. We then selected the component with the most periodic signal to become the final respiratory wave. In this case, the periodicity level was defined as the maximum amplitude observed within 0.13 and 0.66 Hz in the frequency domain.

After extracting the pulse and respiratory waves, HR and BR were estimated in the frequency domain. Figure 34 shows a representative example of pulse and respiratory wave estimation from accelerometer and gyroscope data of a participant while standing up.

These methods and parameters were motivated by the ones used to monitor head motions. However, the methods used in this study require fewer processing steps and lower computational cost (e.g., lower filter orders and no need for Principal Component Analysis).
Combination of Sensors

For part of the analysis we considered combining different sensor modalities. To combine motion modalities (i.e., accelerometer and gyroscope), we time-aligned all the components and used them as input to the algorithms described above. To combine the BVP with the motion-based modalities, we performed a weighted combination of the HR estimates obtained by the two types of sensors. The weights were set to be the normalized absolute magnitudes of the frequencies associated with their HR estimations. For instance, if the maximum frequency response when combining motion-based sensors was 1 Hz (60 beats per minute) with a magnitude of 0.5 dB and the maximum frequency response when using the E3 was 1.5 Hz (90 beats per minute) with a magnitude of 1 dB, the resulting estimation was 1.33 Hz, corresponding to 80 beats per minute.
6.3.3 Results

Laboratory Setting

The validation experiment yielded six one-minute sessions for each of the 12 participants: two sessions in each position (sitting down, standing up and lying down). However, the E3 data of one participant was corrupted for several parts of the experiment due to a loose sensor. The data for that one session was excluded from the HR analysis. To evaluate the performance, we divided each of the one-minute sessions into twenty-second segments with a 75% overlap (N = 594 and N = 648 samples for HR and BR, respectively). The average HR of the segments was 76.70 beats per minute with a standard deviation of 14.26 (minimum of 49 and maximum of 130), and the average BR was 16.63 breaths per minute with a standard deviation of 4.02 (minimum of 7 and maximum of 26).

Table 9: Heart rate estimation with Gear

<table>
<thead>
<tr>
<th>Sensor</th>
<th>ME</th>
<th>STD</th>
<th>RMSE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>2.01</td>
<td>5.89</td>
<td>6.22</td>
<td>0.91</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>1.39</td>
<td>3.85</td>
<td>4.09</td>
<td>0.96</td>
</tr>
<tr>
<td>E3 BVP</td>
<td>0.95</td>
<td>2.74</td>
<td>2.90</td>
<td>0.98</td>
</tr>
<tr>
<td>Gyro.+Accel.</td>
<td>1.27</td>
<td>3.37</td>
<td>3.60</td>
<td>0.97</td>
</tr>
<tr>
<td>Gyro.+Accel.+E3</td>
<td>0.88</td>
<td>1.85</td>
<td>2.04</td>
<td>0.99</td>
</tr>
<tr>
<td>Gyroscope+E3</td>
<td>1.17</td>
<td>2.77</td>
<td>3.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Accelerometer+E3</td>
<td>0.92</td>
<td>2.05</td>
<td>2.24</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table 10: Breathing rate estimation with Gear

<table>
<thead>
<tr>
<th>Sensor</th>
<th>ME</th>
<th>STD</th>
<th>RMSE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>0.38</td>
<td>1.19</td>
<td>1.25</td>
<td>0.95</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>0.97</td>
<td>2.20</td>
<td>2.40</td>
<td>0.82</td>
</tr>
<tr>
<td>Gyro.+Accel.</td>
<td>0.55</td>
<td>1.80</td>
<td>1.88</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 11: Mean absolute error of heart rate with Gear and E3 (N = 216 per column)

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Sitting</th>
<th>Standing</th>
<th>Supine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>2.97</td>
<td>2.04</td>
<td>1.01</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>1.91</td>
<td>1.12</td>
<td>1.14</td>
</tr>
<tr>
<td>E3 BVP</td>
<td>0.78</td>
<td>1.30</td>
<td>0.77</td>
</tr>
<tr>
<td>Gyro.+Accel.</td>
<td>1.59</td>
<td>1.10</td>
<td>1.11</td>
</tr>
<tr>
<td>Gyro.+Accel.+E3</td>
<td>0.88</td>
<td>0.93</td>
<td>0.83</td>
</tr>
<tr>
<td>Gyroscope+E3</td>
<td>1.33</td>
<td>1.39</td>
<td>0.80</td>
</tr>
<tr>
<td>Accelerometer+E3</td>
<td>0.99</td>
<td>0.92</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 12: Mean absolute error of breathing rate with Gear (N = 216 per column)

<table>
<thead>
<tr>
<th>Sensor</th>
<th>Sitting</th>
<th>Standing</th>
<th>Supine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>0.22</td>
<td>0.72</td>
<td>0.19</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>0.41</td>
<td>1.97</td>
<td>0.54</td>
</tr>
<tr>
<td>Gyro.+Accel.</td>
<td>0.22</td>
<td>1.24</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Comparison across Modalities: We compare the performance of our approach using each of the sensor modalities alone (i.e., accelerometer, gyroscope and BVP from the wrist) as well as the different combinations.

Table 9 shows the mean absolute error, standard deviation, mean squared error and correlations for each case when estimating the heart rate. As can be seen, the accelerometer sensor alone achieved a mean absolute error of 1.39 beats per minute, outperforming the gyroscope sensor (ME = 2.01) significantly (Two-sample t-Test, t(1186) = 2.14, p = 0.032). The combination of both sensors slightly outperformed the accelerometer (t(1186) = 0.56, p = 0.576), achieving a mean absolute error of 1.27 beats per minute. These results demonstrate that both gyroscope and accelerometer contain relevant ballistocardiographic information and that our proposed methods can isolate them with high precision during relatively stationary postures, even when the hand was not near the chest. When comparing these sensors with the E3 sensor, the results are slightly worse than the E3 (ME = 0.95) but not significantly (t(1186) = 1.81, p = 0.071). The combination of all sensors significantly outperformed the motion-based sensors (ME = 0.88; p < 0.004), highlighting that both motion-based sensors and BVP provide complementary information for the estimation of HR. While the combination of all sensors was better than when only using the E3 sensor, the difference was not significant (t(1186) = 0.51, p = 0.610). The mean absolute error obtained for the excluded participant was 2.12 beats per minute when combining accelerometer and gyroscope and 11.90 beats per minute when using the too-loosely placed E3. Indeed, the fact we were not able to use the E3 sensor due to it being too loose to read the BVP highlights an important advantage of the motion-based measurements. That is, sensors that rely on motion-based estimations (e.g., smartwatch in our study) do not need to be tight so long there is a contact point with the body (e.g., band, clock). In contrast, BVP monitors that use traditional LED light sources need to be in close contact with the skin to provide accurate ratings. Figure 35 (left) shows a Bland-Altman plot with the heart rate measurements from the combination of the gyroscope and accelerometer for the 594 pairs of measurements (excluding the ones associated with loose E3). The data from the different participants are shown in different colors. The mean error was 0.18 with 95% limits of agreement –7.23 to 6.86 beats per minute.
Table 10 shows the different results when estimating breathing rate with the accelerometer, gyroscope and in combination. While both sensors yielded accurate results (less than 1 breath per minute error), breathing rates were significantly more accurate ($t(1186) = 4.87, p < 0.001$) with the gyroscope ($ME = 0.38$) than the accelerometer ($ME = 0.97$) probably due to the rotational motion of the arms when contracting and expanding the chest during breathing. In this case, the combination of both sensors ($ME = 0.55$) did not yield better results than the gyroscope alone, which may be due to a combination of several factors. For instance, the gyroscope measurements are already very accurate making it difficult to improve performance. Moreover, the algorithm to extract the respiratory wave selects the component with the most periodic signal, which does not always benefit from adding extra signals. Note that our respiration estimates were not compared with the E3 or any other wrist-worn wearable sensor as there are not, to the best of our knowledge, sensors offering breathing rate estimation from the wrist. Figure 35 (right) shows a Bland-Altman plot with the breathing rate measurements from the gyroscope for the 648 pairs of measurements. In this case, the mean error was 0.15 with 95% limits of agreement -2.28 to 2.58 breaths per minute.
**Impact of Body Postures:** Since BCG measurements are influenced by posture, we compared the performance of our methods during three different postures: sitting down, standing up and lying down. Table 11 shows the mean absolute error for all the different sensors when estimating HR from the wrist. As shown in previous studies (e.g., Alametsä et al., 2008; He et al., 2012) sitting was the most challenging posture to obtain clean BCG information. The estimations of HR while sitting were slightly worse ($\text{ME} = 1.59$ beats per minute when combining accelerometer and gyroscope readings) than the estimations obtained during different body postures ($t(592) = 1.64$, $p = 0.102$). When combining the different motion-based modalities, both standing and lying down yielded similar performance ($\text{ME} = 1.10$ and $\text{ME} = 1.11$, respectively). The performance pattern was not the same when estimating HR from BVP, in which standing yielded significantly worse performance ($\text{ME} = 1.30$ breaths per minute, $t(592) = 2.22$, $p < 0.027$) than the other postures ($\text{ME} = 0.78$ and $0.77$ beats per minute for supine and standing). Partly due to this difference in performance across body postures and modalities, combining the three sensors outperformed each of the modalities alone. Interestingly, the combination of accelerometer and BVP slightly outperformed ($p < 0.316$) any other combination for the standing position, yielding a mean absolute error of 0.92 beats per minute.

Table 12 shows the mean absolute error for all the motion-based modalities when estimating BR. In this case, the position with worst performance was standing. This result is also in accordance with findings of the previous section. As can be seen, the gyroscope is significantly better than the accelerometer for all the postures ($p < 0.031$), which can help explain why the combination of the two sensors did not improve overall performance.

**Sampling Rates:** In this study we collected accelerometer and gyroscope data at a sampling rate of 100 Hz. However, when recording for long-periods of time (e.g., days or months) the amount of stored data can quickly increase exceeding the currently available storage space of existing wearable devices. For instance, the sensor we used in our study (Samsung Galaxy Gear) has a storage memory of 4 GB, which could potentially store the equivalent of approximately 6.4 days of continuous monitoring with a sampling rate of 100 Hz. To assess whether we can reduce the sampling rate and still obtain reliable results, we undersampled the recorded data and re-computed the results. Figure 36 shows the mean
absolute error for heart (top graph) and breathing rate (bottom graph) estimation using gyroscope, accelerometer and their combination at different sampling rates (from 5 Hz to 50 Hz). As can be seen, similar performance to the one reported in the previous sections can be achieved with a sampling rate of only 20 Hz. This is partly to be expected as our methods monitor movements well under the frequency of 11 Hz. Sampling data at 20 Hz instead of 100 Hz enables storage of 5 times more data than before (i.e., 32 days of continuous monitoring). Due to the change of sampling rate, the battery life is also extended. For instance, while the Gear battery life lasted for around 6 hours when recording both accelerometer and gyroscope at a sampling rate of 100 Hz, the battery lasted for around 9 hours at a sampling rate of 20 Hz. Note, however, that while 20 Hz seems enough to gather accurate estimations of heart and breathing rates, higher sampling rates may be recommended for the analysis and estimation of other physiological measures such as heart rate variability (Malik et al., 1996), which requires very precise temporal resolution of peaks (not addressed in this thesis).

![Graph showing mean absolute error for heart rate and breathing rate estimation](image)

Figure 36: Mean absolute error for heart rate (top) and breathing rate estimation (bottom) when considering different sampling rates, ranging from 5 to 50 Hz

**In-situ Sleep Measurement**

To preliminarily evaluate our findings in a real-life scenario we analyzed sensor data of two consecutive nights of sleep of three participants who voluntarily agreed to be recorded. All the sensors were switched on right before the participant went to bed for the night. To increase the likelihood that all participants were sleeping, we excluded the first hour of recorded data, yielding a total of 31.57 hours of sleep data.
Table 13: Heart rate estimation from Gear and E3 during sleep (N = 95,510)

<table>
<thead>
<tr>
<th>Sensor</th>
<th>ME</th>
<th>STD</th>
<th>RMSE</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyroscope</td>
<td>1.02</td>
<td>3.96</td>
<td>4.08</td>
<td>0.93</td>
</tr>
<tr>
<td>Accelerometer</td>
<td>1.45</td>
<td>5.30</td>
<td>5.79</td>
<td>0.88</td>
</tr>
<tr>
<td>E3 BVP</td>
<td>1.08</td>
<td>3.10</td>
<td>3.28</td>
<td>0.96</td>
</tr>
<tr>
<td>Gyro.+Accel.</td>
<td>0.95</td>
<td>3.48</td>
<td>3.61</td>
<td>0.95</td>
</tr>
<tr>
<td>Gyro.+Accel.+E3</td>
<td><strong>0.81</strong></td>
<td><strong>2.37</strong></td>
<td><strong>2.51</strong></td>
<td><strong>0.97</strong></td>
</tr>
<tr>
<td>Gyroscope+E3</td>
<td>0.83</td>
<td>2.44</td>
<td>2.57</td>
<td><strong>0.97</strong></td>
</tr>
<tr>
<td>Accelerometer+E3</td>
<td>0.97</td>
<td>2.85</td>
<td>3.01</td>
<td>0.96</td>
</tr>
</tbody>
</table>

ME = Mean absolute error (beats per minute)
STD = Standard deviation of the absolute error, RMSE = Root mean squared error
CC = Pearson’s correlation coefficient (p < 0.001 for all correlations)

Comparison across Modalities: For the analysis we segmented the sleep measurements into 20-second segments with 95% overlap and used the same algorithms to estimate HR for each segment independently. Since cardiorespiratory movements can be easily obscured by large body motions, we implemented a rule to detect and exclude segments with “excessive” motion. In particular, we computed the first derivative of the raw accelerometer data, aggregated the squared components with a square root summation, and excluded the segments that went above a certain threshold. This threshold was empirically set to 0.15 for part of the analysis, which was the maximum value observed during the validation study. After excluding segments with large motions, we were able to preserve around 85.87% of the recordings (yielding N = 97,510 samples, corresponding to 26.94 hours of in-situ sleep measurements). Table 13 shows the performance metrics obtained with the different motion-based sensors, the E3’s BVP, and the different combinations. As can be seen, the mean absolute errors are comparable to the results shown on Table 11 for the supine position. Thus, preliminary findings from wearing the sensor for sleeping at home matched the quality obtained from the laboratory environment. Similarly, the gyroscope (ME = 1.02) yielded significantly better performance than the accelerometer alone (ME = 1.45; Permutation-Test, p<0.001), and their combination outperformed each individually (ME = 0.95, p<0.001). Furthermore, the combination of all the sensors yielded significantly better performance (ME = 0.81, p<0.001) than E3’s BVP or motion-based sensors separately, providing further support that light-based and motion-based sensors contain complementary information.
Figure 37: Representative example of heart rate estimation with motion-based sensors on the wrist (dashed-green) and with chest ECG electrodes (blue) during one night of sleep. Raw accelerometer (middle) and gyroscope (bottom) readings. Red areas in the top graph indicate regions with "excessive" accelerometer motion.

Figure 38: (Left) Percentage of excluded segments for different artifact thresholds. (Right) Mean absolute error (beats per minute) for different combinations of sensors.

Figure 37 shows a representative example of one night of sleep for one of the participants. The top graph shows the gold standard heart rate (obtained with the Alive sensor) and the heart rate estimation obtained when combining motion-based sensors. Due to the overlap during the segmentation, HR at each point was computed as the average of HRs obtained from each of the segments that contained that data point. As can be seen, both estimates are closely aligned for the whole duration of the recording. Red areas on the top graph indicate the sections that were excluded for the analysis due to "excessive" motion artifacts.
(threshold 0.15). For most of these areas, the effect of motion artifacts can be seen as abrupt sporadic changes of HR estimates. Note, however, that when analyzing signals over a long period of time, these sporadic peaks can be easily removed by encoding contextual rules that enforce smoothness of the changes. Nevertheless, for the purpose of this study, we wanted to avoid adding additional layers of complexity and to enable a fair comparison across modalities. Raw accelerometer and gyroscope readings are shown on the middle and bottom graphs, respectively. These graphs illustrate that the gyroscope is less affected by motion than the accelerometer, which may be one reason why it, when used solo, outperforms the accelerometer.

**Artifacts**: During the previous analysis we excluded segments that contained motion artifacts above a pre-defined threshold based on our validation study, which enabled us to estimate heart rate for around 86% of the data. To assess the impact of this parameter on performance, we computed the mean absolute error for different artifact thresholds. Figure 38 shows the amount of excluded information for different thresholds (left) and performance for different combinations of sensors (right). As expected, using smaller thresholds to remove motion artifacts yields better performance but also limits the amount of data that we could analyze. For instance, when using a threshold of 0.05, we can obtain a mean absolute error of 0.67 beats per minute when using the combination of all the sensors but we can only provide estimates for around 77% of the data. As can be seen on the right graph, the gyroscope has more tolerance to artifacts than the accelerometer and, therefore, shows better performance. On the other hand, the E3’s BVP sensor seems to degrade more slowly with the inclusion of larger motion artifacts than motion-based sensors. The combination of both motion-based sensors always outperformed each of the motion sensors alone, and outperformed the E3 for lower artifact thresholds (smaller than 0.33). Finally, the combination of all modalities always yielded better performance than the others, irrespective of the artifact thresholds. Note that the results described in the previous section correspond to the results obtained when the artifact threshold is 0.15.
6.3.4 Discussion

Motivated by the previous study with Google Glass, in this section we have proposed low-complexity algorithms that estimate heart and breathing rates from wrist-motion sensors. These algorithms use traditional signal processing techniques to ensure minimal processing power in anticipation of designing a real-time vital signs monitor.

Among some of the main results, we have shown that traditional sensors for BCG such as accelerometers can sense heart and respiratory activity from the wrist, a more peripheral location than traditionally studied locations (e.g., on the chest (Dinh, 2011; Kown et al., 2011), ear (He et al., 2012), below the feet (Inan et al., 2009; Wiard et al., 2011)) even when the sensor is not in contact with the chest. This finding has the potential of enhancing the capabilities of most currently available wrist-worn wearable devices, which already incorporate motion sensors for the purpose of artifact detection and behavioral understanding (e.g., number of steps). We have also found that accelerometers outperformed the gyroscopes when estimating HR but gyroscopes outperformed the accelerometers for BR estimation. Furthermore, the combination of both sensors yielded slightly better performance for HR estimation. Both the gyroscope and the accelerometer have also been compared with a state-of-the-art wrist-worn device that measures blood volume pulse from the wrist for HR estimation. While results from the light-based BVP measurements were better than from the motion sensors in the validation study, that was not the case for the real-life sleep setting. Furthermore, the combination of all the sensors yielded better results than any solo sensor. While motion-based BCG measurements usually contain more noise than light-based PPG measurements, the recording of BCG requires less energy and does not need to be tightly attached to the body to provide accurate readings. We have also provided results across different sampling rates and discussed their implications in terms of performance, battery power and storage space which are critical aspects in the design of wearable devices. Finally, we have shown preliminary results supporting generalization of these methods in real-life settings such as sleeping at home.
6.4 Smartphone Motions

The previous results demonstrate that motion sensors embedded in head and wrist-worn wearable devices can capture heart and breathing rates accurately. While the results were very promising, these type of devices are still not widely used and some people may find them cumbersome or stigmatizing, especially if they are not used to wearing glasses (see Chapter 5). Motivated by these challenges, this section explores the possibility of using motion sensors inside currently available smartphones to see how they can measure breathing and heart rates while being carried in different locations on the body (e.g., trouser pocket, shoulder bag) and when used during regular phone activities (e.g., watching a video, listening to a conversation), as depicted in Figure 39.

![3-Axis Accelerometer](image)

81 Beats per Minute

22 Breaths per Minute

Figure 39: Motion sensors of a smartphone can be used to recover heart and breathing rates of the users during stationary positions and activities such as listening on the phone

6.4.1 Experimental Additions

This study contained two separate parts. The first part consisted of the same experimental protocol as the other studies (see section 6.1.3). During this part, participants were asked to carry a phone in their pocket and, when in the standing position, participants were also asked to carry two bags with phones inside; one in the left hand and the other hanging from the right shoulder. This procedure enabled capture of different phone carrying behaviors,
which are more common for specific demographics such as females (Ichikawa, Chipchase & Grignani, 2005; Cui, Chipchase & Ichikawa, 2007). For the second part of the experiment, participants were asked to perform several traditional phone-related activities while sitting down. Specifically, participants had to watch a video, listen to a conversation, and browse the Internet for a minute each. During this part of the experiment, participants were instructed to hold the phone as they would during their regular phone activity. Figure 40 shows an overview of the location of the phone for all the conditions.

![Figure 40: Location of the smartphone during each part of the experimental protocol. Participants remained still while standing up, sitting down, and lying down, which were repeated before and after exercising, followed by watching a video, listening on the phone, and browsing the Internet while sitting down](image)

6.4.2 Proposed Methods

**Pulse Waveform**

A moving average window (N = 15 samples, sampling frequency: 256 Hz) is subtracted from each of the axes to detrend the data. Then, each of the components is set to have zero mean and unit variance so they have the same weight and the analysis is more robust to different device orientations. A band-pass Butterworth filter (cut-offs at 7 and 13 Hz, N = 1) is then applied to isolate the BCG motions of each component. The resulting components are then aggregated with a squared root summation of the squared comments. Finally, another band-pass Butterworth filter (0.66 – 2.50 Hz, N = 1) is applied to obtain the final pulse waveform.
**Respiratory Waveform**

Similarly to the previous processing steps, a moving average window (8.5 seconds) is first subtracted from each of the components which are then z-scored. To remove additional noise, we used Independent Component Analysis (Jade implementation; Cardoso, 1999). To isolate the respiratory motions, the resulting components are then band-pass filtered with a Butterworth filter ($0.13 \text{ Hz} - 0.66 \text{ Hz}$, $N = 1$). Finally, we automatically selected the most periodic component as the final respiratory waveform. As in previous sections, periodicity of the signal was estimated by the maximum magnitude achieved in the frequency domain within the previously used frequency range.

These algorithms were motivated by the methods of the previous two studies as well as (Poh et al., 2011), which considered different datasets and modalities to extract physiological parameters. The methods were adjusted (e.g., different window sizes) to be able to correct undesired motion artifacts and to capture more subtle motions (e.g., lower filter orders).

### 6.4.3 Results

Table 14 and Table 15 show a summary of the phone results across locations and activities for HR and BR, respectively. Figure 41 shows representative Bland-Altman plots for some of the conditions.

**Phone in Pocket with Different Body Postures**

The left-most columns of the tables show the performance of our methods when the phone was inside the pocket of the participant. These results include estimations across three different body postures – sitting, supine, and standing (pre- and post- physical exercise). As can be seen, both HR and BR estimations show low mean absolute error rates with some differences across the two measurements and body postures. For HR estimation, the mean absolute error achieved during the supine position was significantly better than during the other two positions (Two-sample t-Test, $p<0.005$). This finding is aligned with the original studies (Starr et al., 1939), which required participants to lie down to minimize the amount of unexpected motion. On the other hand, while the standing position yielded the worst performance across body postures, it still provided reasonably accurate results. For BR
estimation, the mean absolute error rate achieved during the sitting position was significantly better than the other two positions (p<0.001). We believe this is the case due to the proximity of the phone to the stomach, which is more directly influenced by respiratory motions. Overall, BR estimations were more accurate than HR estimations due to their larger amplitude and lower frequency range. The motions were more easily transmitted through the body and sensed by peripheral accelerometers in the phone.

Table 14: Heart rate estimation with the Phone

<table>
<thead>
<tr>
<th></th>
<th>Pocket</th>
<th>Bag</th>
<th>Hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>1.97</td>
<td>1.16</td>
<td>3.37</td>
</tr>
<tr>
<td>Supine</td>
<td>2.93</td>
<td>3.00</td>
<td>6.67</td>
</tr>
<tr>
<td>Stand</td>
<td>3.53</td>
<td>3.21</td>
<td>7.46</td>
</tr>
<tr>
<td>Hand</td>
<td>11.56</td>
<td>13.98</td>
<td>13.98</td>
</tr>
<tr>
<td>Shoulder</td>
<td>4.85</td>
<td>4.97</td>
<td>7.40</td>
</tr>
<tr>
<td>Watch*</td>
<td>2.37</td>
<td>4.97</td>
<td>7.40</td>
</tr>
<tr>
<td>Listen*</td>
<td>3.38</td>
<td>5.48</td>
<td>8.25</td>
</tr>
<tr>
<td>Browse*</td>
<td>3.71</td>
<td>8.25</td>
<td>15.96</td>
</tr>
</tbody>
</table>

Table 15: Breathing rate estimation with the Phone

<table>
<thead>
<tr>
<th></th>
<th>Pocket</th>
<th>Bag</th>
<th>Hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sitting</td>
<td>0.26</td>
<td>0.92</td>
<td>2.16</td>
</tr>
<tr>
<td>Supine</td>
<td>0.50</td>
<td>2.46</td>
<td>3.41</td>
</tr>
<tr>
<td>Stand</td>
<td>0.56</td>
<td>2.62</td>
<td>4.03</td>
</tr>
<tr>
<td>Hand</td>
<td>2.05</td>
<td>4.51</td>
<td>3.96</td>
</tr>
<tr>
<td>Shoulder</td>
<td>2.26</td>
<td>3.40</td>
<td>3.71</td>
</tr>
<tr>
<td>Watch*</td>
<td>3.38</td>
<td>4.32</td>
<td>0.71</td>
</tr>
<tr>
<td>Listen*</td>
<td>2.44</td>
<td>0.58</td>
<td>0.98</td>
</tr>
<tr>
<td>Browse*</td>
<td>1.47</td>
<td>2.58</td>
<td>2.95</td>
</tr>
</tbody>
</table>

ME = Mean absolute error (beats/minutes), STD = Standard deviation of the absolute error
RMSE = Root mean squared error, CC = Pearson’s correlation coefficient (p < 0.001 for all correlations)
N = 216 for each column except for * which N = 108

Standing: Pocket vs Hand-bag vs Shoulder-bag

While the participants were holding the standing position, three different phones at separate locations were simultaneously monitoring motion data: one inside the pocket, another inside a bag in the left hand, and another inside a bag hanging from the right shoulder. When comparing results across locations, the phone inside the bag hanging from the shoulder yielded slightly better mean absolute error for BR and significantly better for HR estimation (p<0.001). We believe this result is due to a combination of several factors. When hanging a bag from the shoulder, the accelerometers can capture more accurate
ballistocardiographic and respiratory motions that are more prominent along the vertical axis (Starr et al., 1939) and around the chest location. Moreover, when hanging a bag from the shoulder the amount of contact with the body is large and its range of movements are more constrained than when holding the bag from the hand. Overall, these factors can enable easier propagation of the subtle cardiorespiratory motions. During the standing position, one of the participants had difficulties remaining relatively still which had a detrimental impact on some of the results. The data points associated with this participant correspond to the purple dots of the middle graphs of Figure 41.

Figure 41: Bland-Altman plots for heart (left) and breathing rates (right) of the conditions that yielded the best mean absolute error when the phone was inside the pocket (top), inside the bag (middle), and on the hand (bottom). Mean error is depicted with slashed red and 95% limits are depicted with slashed green lines. (HR: Heart Rate in beats per minute, BR: Breathing Rate in breaths per minute, Accel: Accelerometer). N = 216 for top and middle graphs, and N = 108 for bottom graphs.
**Phone Activities**

While the phone may remain most of the time in one of the locations already considered (Ichikawa, Chipchase & Grignani, 2005; Cui, Chipchase & Ichikawa, 2007), the main purpose of the device is to facilitate performing activities such as communicating with other people. In this section, we provide preliminary analysis of HR and BR estimation while the phone is being used during three, relatively stationary, common activities: watching a video, listening to a conversation, and browsing the Internet.

In terms of HR estimation, both watching and listening yielded slightly worse but still comparable results to the ones obtained when the phone was inside the pocket. However, the mean absolute error was significantly worse when browsing information than the other two activities ($p<0.001$). Among all the different conditions, this is the only one in which the user is actively manipulating the device and, therefore, is affecting the accelerometer readings more directly. Indeed, subtle touch interactions such as zoom-in, zoom-out and finger taps elicit very similar acceleration patterns to the ones associated with ballistocardiographic beats and could potentially confuse the algorithms. While more complex approaches could be used to detect and cancel the effects of these non-cardiorespiratory motions, larger motions such as the ones observed during daily activities (e.g., walking) could easily obscure the subtle motions associated with the heart beats.

In terms of BR estimation, listening to a conversation with the phone next to the ear yielded significantly better results than the other two activities ($p<0.001$) which was comparable to the results obtained when the phone was inside the pocket during the sitting down position. Interestingly, the performance during the browsing activity was still very accurate, indicating that touch interactions were not confused with respiratory motions probably due to the lower frequency range of the latter.

### 6.4.4 Discussion

The previous section demonstrates that physiological parameters such as heart and breathing rates can also be recovered from a phone via accelerometer measurements while the person is carrying it in different locations or using it during different activities. However, most of the considered conditions involved positions and activities without too much motion (e.g., lying down, watching a video). As shown in the browsing condition,
for example, motions such as those presented when interacting with the phone can impair the performance of the proposed methods. Therefore, our methods would only be applicable to provide sporadic assessments during the day when the amount of motion is small (e.g., reading a book, watching TV). Chapter 7 focuses on extending this research and assesses the accuracy and utility of the proposed methods in a real-life workplace scenario. Note, however, that alternative methods that cancel motion artifacts could also be developed, but they are beyond the scope of this work.

6.5 Conclusions

This chapter has systematically demonstrated the possibility of using wearable motion-sensitive sensors embedded in head-mounted, wrist-worn and smartphone devices to capture physiological parameters of the wearer. In particular, we have proposed several real-time algorithms to process motion from different sensors and locations, provided validation of heart and breathing rate estimation with FDA-cleared sensors in mostly controlled laboratory settings, and quantitatively compared relevant factors (e.g., combinations of sensors, body postures, sampling rates, observation windows).

When studying head motions, we have shown that the gyroscope outperformed the other sensors, including the accelerometer upon which prior BCG measurements are mostly based. We believe this improvement is partly due to the above-eye location of the sensor and its capability to capture amplified rotational movements of BCG. Moreover, we have demonstrated that analyzing the data from a head-mounted camera is a novel and promising method for collecting physiological information of the wearer, with the benefit of also providing insightful visual context. Finally, as each of the modalities captures different aspects of motion, their combination offers the opportunity to improve performance.

When studying wrist motions, we have shown that accelerometers outperformed the gyroscopes when estimating heart rate but gyroscopes outperformed the accelerometers when estimating breathing rates, thereby illustrating the benefit of capturing different types of motions. Moreover, we compared motion-based methods with the current state-of-the-art light-based methods on the wrist for HR estimation, and demonstrated that both methods are complementary and can improve overall accuracy. Furthermore, we provided
preliminary results supporting the generalization of our methods in a real-life setting such as while at home sleeping.

Finally, when considering more peripheral motions, we have demonstrated that the same physiological parameters can be gathered (although less accurately) with the motion of sensors currently available on smartphones, when carrying them inside the pocket and/or while using them to perform relatively "still" activities. The results are very encouraging but also worrisome as it offers the possibility for intrusion of privacy. For instance, an application could be used to covertly monitor the physiological responses of individuals to personalize advertisements. While this may be a potential application in the future, it is important to appropriately inform users about how the data is being used. Currently, most phone users are not aware that cardiovascular health information can be conveyed simply by carrying or holding a phone that contains accelerometers. Findings like the ones presented in this chapter urge us to reconsider how this type of data is monitored, stored and transmitted to enforce transparency and protect user’s privacy of their health data.

Future efforts will consider evaluating other modalities and developing novel methods to more effectively combine them while also considering contextual information such as body postures or activities. We have also started to work on more sophisticated methods that can handle large motions associated with daily activities, which is fundamental to apply the proposed methods in real world settings. In the future, we will also be analyzing other relevant physiological parameters such as heart rate variability (e.g., Brüser et al., 2013) as it has been shown to be associated with cognitive load (Moriguchi et al., 1992) and stress (Hjortskov et al., 2004; Moses, Luecken & Eason, 2007). The key to this parameter is to obtain highly accurate timing of the heartbeats and, therefore, uniform sampling rates and filters with linear phase or without phase delays are preferred.

In summary, the studies presented in this chapter demonstrate a new capability for providing accurate real-time heart-rate and respiration measures from motion-sensitive sensors available in today wearable and mobile devices. However, there are several research challenges that need to be addressed before we can provide continuous physiological measurement. As these methods continue to advance, we hope they will be used to create passive and comfortable assessments that foster greater health and wellbeing during daily life.
Chapter 7

Opportunistic Heart Rate Assessment from Motion Sensors during Real-Life

This chapter extends the research presented in the previous chapter and examines the possibility of using wearable-motion sensors in real-life settings to estimate heart rate from peripheral body locations (head, wrist and inside the pocket). In particular, we use the methods proposed in Chapter 6, which were developed in relatively “still” laboratory conditions, and evaluate them in our real-life workplace dataset (Chapter 4), involving 15 participants during five days of work. Instead of developing complex methods that compensate for daily activity motion, this chapter focuses on opportunistic measurements during the day when the person is relatively “still.” The chapter is divided as follows. First, we start by describing the methods used for the analysis (sensors, devices, analysis, and processing algorithms). Then, we overview the laboratory results and evaluate the methods in our real-life dataset. Finally, we provide some discussion and concluding remarks.

7.1 Methods

7.1.1 Apparatus
This chapter considers the measurement of ballistocardiographic (BCG) signals with wearable motion sensors on three peripheral body locations (pocket, wrist and head) which capture three of the most commonly used locations for wearable devices. In particular, we use the Galaxy S4 smartphone (Samsung, Inc.) to capture motions from the pocket, the
Gear Live smartwatch (Samsung Inc.) to capture wrist motions, and the Google Glass (Google, Inc.) to capture head-motions.

Each of the devices is equipped with both a 3-axis accelerometer and 3-axis gyroscope sensor that captures linear accelerations (meters/second^2) and rotational movements of the devices (radians/second), respectively. To retrieve and log motion information, we created a custom Android program that collected data at an average sampling rate of 100 Hz. However, all sensors readings were down-sampled to 50 Hz to minimize the virtual and disk memory required during the analysis. As shown in section 6.2.3, 50 Hz still captures enough information to make accurate heart rate assessments from peripheral BCG signals.

7.1.2 Gold Standard Heart Rate Measurement

To collect gold standard heart rate measurements, we used a single lead BioPatch (Zephyr Tech, Inc.) with Kendall 535 foam pre-gelled electrodes that captures electrocardiography (ECG) measurement from the torso (250 Hz). The device also generates heart rate estimates (1 Hz) and confidence values for each of the estimates. These values range from 0 to 100 indicating not confident to very confident, respectively. Only estimates with more than 50% confidence were used to evaluate the performance of the motion-based methods (as described later about 70% of the data had a confidence score above 50%). The BioPatch as well as the other wearable devices were synchronized with the same clock at the beginning of each day and locally stored the sensor information on each device.

7.1.3 Heart Rate Estimation from Motion

To estimate heart rate from peripheral BCG signals we used the methods described in sections 6.2.2, 6.3.2 and 6.4.2 for the head-mounted, wrist-worn and phone devices, respectively. These methods share similar processing steps but have different parameters (e.g., filter orders, frequency ranges) to differently capture the subtle BCG differences inherent to each of the locations. The main processing steps of the algorithm are summarized as follows:

1. **Signal pre-processing.** The motion signals (3-axis accelerometer, 3-axis gyroscope or a combination) are de-trended with averaging filters and each of the components is z-scored. These steps remove small and slow motion artifacts (e.g., gyroscope drifts) and
give the same weight to each of the components, making the estimates more robust to different device orientations.

2. **BCG isolation.** A band-pass Butterworth filter, with cut off frequencies specific to the device, is used on each of the components to isolate and amplify BCG motions. The minimum and maximum cut off frequencies across devices ranged from 4 to 13Hz.

3. **Component aggregation.** The resulting filtered 3-axis motion data signals are aggregated with a squared root summation of the components, providing the same weight to each of the components, and making the estimations more robust to different device orientations.

4. **Pulse wave and heart rate estimation.** Finally, a second band-pass Butterworth filter with cutoff frequencies of 0.75 Hz and 2.5 Hz (corresponding to 45 and 150 beats per minute) is applied to extract the final pulse wave. Using this wave, the heart rate is estimated by finding the frequency with highest amplitude in the Fourier domain and multiplying it by 60 (beats per minute).

As in our previous analysis, 20-second window segments with a 75% overlap were used to over sample and divide the streams of data. While more complex methods could be developed to attempt to model and remove real-life motion artifacts, in this first study we wanted to replicate the procedure applied in laboratory experiments in order to establish a baseline comparison.

### 7.1.4 Motion Level

One of the fundamental limitations when measuring BCG signals is that large motions can quickly obscure the subtle BCG motions, especially when considering locations far from the chest. To assess how the previous methods handle different levels of daily motion, we created a metric that captures the amount of motion within each of the 20-second windows of data. In particular, we computed the first derivative of the raw accelerometer data, aggregated the squared components with a square root summation, and computed its standard deviation. This metric is consistent with the pre-processing steps described above (it also removes signal drifts of the sensors) and captures the range of quick motions that may negatively impact the performance of the methods. Using this criteria, a “still”
segment only containing BCG information will show significantly smaller standard deviation values than segments with large and apparent motions such as those of daily activity.

To use this metric for the analysis, we will define different levels of acceptable motion and only use the segments with values below the level. In other words, if the motion level is set to 0.3, we will only consider 20-second segments for which the standard deviation of the first derivative of the accelerometer data are below 0.3.

7.2 Results

This section provides an overview of the findings of the study. We start by reviewing some relevant information about the previously performed laboratory studies to help contextualize the rest of the analysis. Then, we systematically evaluate the performance of each of the methods during real-life. Finally, we quantify the distribution and frequency of “still” moments for each of the devices in our study.

7.2.1 Laboratory Overview

This chapter leverages methods developed in laboratory conditions and tests them in a real-life work environment. For each of the laboratory experiments (see Chapter 6), 12 participants were recruited to perform three body postures (standing up, sitting down and lying down) during two separate minutes (one before, and another after, physical exercise) while capturing motion data from each of the considered locations. Table 16 shows a summary of the best results obtained for each of the studies in the context of heart rate estimation. While each of the previous analyses explored different modalities and sensor combinations, the table includes only the best performing combination and the mean absolute error when estimating heart rate. To simplify the analysis, the rest of results will only consider the best performing combination for each of the locations and leave for future work a comparison across sensors of potential combinations. That is, we will use the
Table 16: Summary of laboratory findings for heart rate estimations

<table>
<thead>
<tr>
<th>Device location</th>
<th>Best performing combination of sensors</th>
<th>Mean absolute error (beats per minute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pocket</td>
<td>Accelerometer</td>
<td>2.17</td>
</tr>
<tr>
<td>Wrist</td>
<td>Accelerometer + Gyroscope</td>
<td>1.39</td>
</tr>
<tr>
<td>Head</td>
<td>Gyroscope</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Figure 42: Distribution of motion levels observed in the laboratory experiments for the pocket (top), wrist (middle), and head (bottom) accelerometer sensors in which only ballistocardiographic signals were observed (N: 648 per histogram)

The gyroscope sensor for the head-worn device, the accelerometer for the phone, and a combination of accelerometer and gyroscope for the wrist-worn device.

Using the motion level metric defined before, Figure 42 shows the histogram of motion values observed during the laboratory experiments. As can be seen, the head BCG motions yielded larger amplitudes, and both wrist and pocket BCG motions were more subtle. This result is to be expected as both the wrist and the phone are in more peripheral locations, further away from the chest, where the motions are less prominent. In the remaining analysis we will use these ranges for each of the devices to help improve the resolution of the result and better study the impact of different motion levels.

7.2.2 Real Life Heart Rate Estimation

After excluding segments of data when participants were charging or not carrying the devices, there were around 488, 445 and 441 hours for each of the sensors inside the
smartphone, the wrist-worn, and the head-worn wearable devices, respectively. On average, each participant contributed around 6 hours (STD = 1.29) of useful recordings per day per sensor, and around 31 hours (STD = 5) by the end of the study. The differences across devices were due to several factors such as devices running out of battery, forgetting to charge some of the devices, and occasional malfunctioning of some of the devices. After segmenting the data into 20-second pieces, we obtained a total of 1.2 million segments for each of the two sensor modalities (gyroscopes and accelerometers). From these segments, only 70% of them had a heart rate estimation with a confidence value above 50% (yielding around 822 thousand segments).

Figure 43 shows the average mean absolute error (green lines) and the standard errors (green lines) in beats per minute across the 15 participants (black dot lines) for different motion levels. As can be seen, there is large variability across participants, devices, and amounts of motions, suggesting that better individual-level results could be gained by setting individual thresholds. However, all the curves show a decrease in terms of performance when increasing the motion level, supporting that our motion level estimation (see section 7.1.4) effectively captures the negative impact of motion in our methods. Among the three devices, the head-worn device is the one that provided the lowest mean absolute errors, followed by the wrist-worn and then the phone. While the performance values are worse than the ones observed in laboratory experiments, the differences across devices are still consistent, further supporting that more peripheral locations are more challenging to measure and contain more subtle information.

The methods yielded very accurate results for some of the participants (below 5 beats per minute of mean absolute error for different motion levels) which are more comparable to the results obtained in the laboratory experiments. However, there were other participants for which the methods did not work as well, even for low motion levels. While it is very challenging to account for every possible factor in an uncontrolled real-life study, these differences were positively correlated with the amount of motion each person experienced during their daily activity. Thus, participants who remained most of their time sitting at their desk in front of the computers, yielded the lowest mean absolute errors and participants who were more active usually yielded the worst performance.
Figure 43: Average mean absolute error (blue line) and standard error (green lines) across all participants (dot black lines) for the pocket (left), wrist (center), and head (right) locations. The total number of considered segments for these graphs were 227K, 205K, and 203K, respectively.

Figure 44: Distribution of motion levels from motion sensors in the pocket (left), wrist (center), and head (right) when considering the whole dataset. The total number of considered hours were 488, 445 and 441, respectively.

Figure 45: Temporal distribution of potential assessment from motion sensors in the pocket (left), wrist (center), and head (right) locations, when considering motion levels that yielded a mean absolute error of $\leq 5$ beats per minute. Each row represents a different participant, grey rectangles indicate different days, blue areas indicate aggregated distribution, and grey dots indicate separate heart rate assessments. The results were generated from 1374 hours of accelerometer data.
7.2.3 Distribution of “Still” Moments

The previous section demonstrates that it is possible to opportunistically estimate heart rate of people when the amount of motion is relatively small. This section studies the frequency with which this happened during daily life.

Since the data were segmented into 20-second segments with an overlap of 75%, there is the potential to generate a prediction every 5 seconds. Indeed, each 5-second piece is potentially covered by 5 sliding 20-second segments. To better capture a potential usable assessment moment, we assigned each 5-second segment the smallest motion level associated with one of the associated data windows. Figure 44 shows the distribution of potential assessments under different motion levels when considering the 1374 hours of collected data. As can be seen, the availability of “still” segments during the day also varies across devices and participants, but is largest for the most participants with the head-worn device.

The wrist location yielded the larger amounts of motion and, therefore, the fewer segments of “still” data. This is partly to be expected as the sample of participants here needed to constantly use their hands to perform their daily work. Among the three sensors, the head-worn sensors showed the largest amount of “still” samples throughout the different days, further supporting the benefit of using such type of devices in this context.

To understand how these segments are distributed over time, Figure 45 shows the average distribution of segments for all the recorded data when considering accelerometer motion levels of 0.0106, 0.0199, and 0.058 for the pocket, wrist and head locations, respectively. These values correspond to the motion levels with an average mean absolute error of 5 beats per minute in Figure 44, which yielded 6%, 2.7% and 20.6% of the data captured by each of the devices. The graph shows the differences in terms of distribution across the different days of each participant (grey rectangles). Also, it can be easily seen that there are clusters in which lower motion levels are observed and represent the varying activity levels throughout the day. Grey dots indicate that at least one isolated 5-second piece qualified for the assessment during that time, indicating that even though continuous measurements were more challenging, opportunistic assessments were still possible.
7.3 Discussion

The previous section demonstrates that it is possible to opportunistically provide heart rate assessments during daily activity relying on peripheral wearable motion sensors.

Among some of the main findings, we observed that head-mounted devices can yield the most accurate as well as the most frequent assessments when considering the three locations. To a lesser extent, the wrist and the phone yielded less frequent and accurate assessments. While these results may suggest that head-worn devices are the most promising location for daily life BCG measurement, it is important to also consider other relevant factors such as the availability and pervasiveness of certain devices. For instance, it is estimated that the amount of smartphones users worldwide will be 2.16 billion by 2016 (according to eMarketer³). Even if these devices could only provide one or two accurate assessments each day, it could quickly generate a large amount of health data with potentially invaluable medical importance. In this case, the research challenge is to accurately distinguish the noisy and accurate data points. In this chapter, we have shown that a simple motion metric based on the standard deviation of the first derivative can effectively discriminate between different levels of accuracy. However, we expect this metric will need to change when considering more complex and sophisticated methods.

The findings presented on this chapter are also very dependent on the specific studied population. Participants in our study spent a large part of their work activity working in front of the computers which is convenient due to the stillness of the head. However, other populations such as those working on ambulatory jobs such as transportation or package delivery may have less “still” data during their daily activity. For these cases, more complex methods that can model the sources of motion or different body locations may be more appropriate.

This chapter explored the use of previously developed methods in controlled laboratory scenarios without introducing any adaptation. While the results presented in this study are very promising, they were still not as accurate as those observed in the controlled settings. This difference is always to be expected when considering the complexity of daily activity and the amount of motion associated with it. While we expect these motions can be detected

³ http://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694
and corrected, other types of motion are more subtle and closely resemble BCG signals. For instance, a person working at the computer can artificially simulate the beating of the heart when tapping their feet due to nervousness or when following the beat of background music. Despite these challenges, the results of this chapter are still very encouraging and demonstrate that it is still feasible to capture subtle cardiac vibrations during daily life.

7.4 Conclusions

Wearable motion sensors are ubiquitous and offer a unique opportunity not only to track apparent behavioral activity but also to provide comfortable and passive physiological assessments during the day. This chapter provides a first step towards realistic, in-situ assessments of these measures in the context of different peripheral and challenging body locations. While there are still many challenges to address before we can have continuous assessments, we are excited about the possibility of making health information more widely accessible with low-cost motion sensors.
Chapter 8

Stress Recognition from Wearable Data

The previous three chapters have shown that wearable devices can be used to improve the collection of ground truth self-reports as well as to comfortably measure physiological parameters during the day. This chapter develops supervised learning methods to automatically infer self-reported stress levels from different types of wearable data in the context of our real-life workplace study (described in Chapter 4). The chapter is divided as follows. First, we start reviewing some of the main challenges associated with real-life stress measurement that are present in our main study. Then, we review the collected survey and self-reported stress data, since it is intended to serve as a gold standard. We continue by describing how we characterize the wearable data and the different preprocessing steps. Finally, we review the predictive value of the different types of signals measured and discuss the implications of this work.

8.1 Real-life Stress Measurement Challenges

Researchers have widely studied stress in controlled laboratory settings in which potential sources of noise and interference are minimized or eliminated. This section overviews some of the main challenges that appear when studying stress “in the wild.”

One of the most important and challenging aspects of studying real-life stress is how to elicit regular and accurate gold standard for experience of stress. While there exist a
wide variety of stress measurement approaches, self-reports are commonly accepted as the gold standard. When considering real-life settings, a method is required that can effectively facilitate collection of self-reports without eliciting additional stress. Furthermore, one needs to find a set of questions that can quickly and unambiguously capture the specific type of studied stress. Note, however, that even if the previous two challenges are appropriately addressed, self-report measures may be still vulnerable to other problems such as recall and cognitive biases or difficulty to quantify an emotional state due to alexithymia, work overload, stigma of a situation, or other reasons. To help minimize these challenges, this thesis has built and deployed custom-made wearable experience sampling software across several kinds of devices (see Chapter 5) and also explores the use of regression analysis to remove possible sources of error from the self-report data (section 8.4.5).

When considering physiologically-based methods in real-life, there are several factors that can impair the quality of the readings. For instance, many wearable physiological sensors can be negatively affected by body motion, with noise introduced when walking or changing body posture. These motions not only detrimentally affect the readings but also may create false readings that can be confused as accurate physiological measures. Moreover, different body postures or environmental conditions can change the range of physiological parameters. For instance, heart rate is usually higher when standing than when sitting down, and skin conductance is usually higher in warm and humid environments. To partially address this challenge, we capture different types of wearable data (physiological, behavioral and contextual signals).

Finally, a major challenge when studying stress is the large variability observed within and across people. A major part of the variability is due to the way people appraise stressful events. For instance, the same event may be reported as stressful for some people but not for others, and even the same event may elicit different levels of experienced stress in the same person depending on the circumstances. This challenge becomes worse when considering real-life settings in which the range of experienced emotions and possible circumstances are more varied than in controlled laboratory settings. Other sources of variability are due to the individual differences associated with the specific type of information being collected and analyzed. For instance, physiological readings can be
affected by multiple factors such as gender, age and biological rhythms. While understanding all sources of variance is very challenging, this study analyzes the findings both at an individual level and at group levels. Furthermore, we explore the use of normalization methods that amplify daily deviations from personalized baselines.

8.2 Surveys

As part of our experiment, participants had to fill in the Toronto Alexithymia Scale (TAS) which helps quantify the difficulty some people have in identifying and describing their own emotions. Therefore, scoring high on this scale can be associated with poor quality of the emotional self-reports (see more details on section 4.3). In our study, the average TAS score was 43.33 (STD = 10.38) with a minimum of 23 and a maximum of 60. All the participants scored below the cut-off threshold (61) for alexithymia. However, four of the participants scored between 52 and 60. While these people could be considered as potentially alexithymic, the results presented in this chapter did not vary significantly for these people.

Participants also had to fill in the Perceived Stress Scale (PSS), which helps quantify the amount of stress that participants are experiencing during their daily life. The average PSS score across the participants was 18.8 (STD = 6.66) with a minimum of 8 and a maximum of 28. These scores were significantly correlated with the TAS scores (Pearson’s correlation coefficient, $r = 0.63$, $p = 0.01$), suggesting that experiencing higher levels of stress is related to the ease to identify and describe emotions. PSS scores were also significantly negatively correlated with conscientiousness and positively correlated with the neuroticism dimensions of the Big Five ($r = -0.60$, $p = 0.02$ and $r = 0.76$, $p<0.01$, respectively). This result is consistent with previous studies (e.g., Schwebel & Sult, 1999; Vollrath, 2001; Murphy et al., 2013). The conscientiousness dimension was also significantly correlated with the TAS score ($r = -0.79$, $p<0.01$) and was slightly different for each gender (Two-sample t-Test, $t(13) = 1.84$, $p = 0.09$). Females scored higher on average for conscientiousness. No other significant differences were observed.
Figure 46: Distribution of stress levels by each of the self-reported affective states (left) and work conditions (right) during the study. Top graphs show the original stress self-reports, and bottom graphs show the derived stress self-reports. Each dot represents a separate self-report and its color indicates the stress level, ranging from dark red indicating the highest stress level (5) to dark blue indicating the lowest stress level (1). Black circles indicate the average response. (N = 512 for each of the graphs)
8.3 Self-reports

During the study, there were a total of 516 successfully completed prompts. Each of the prompts included two questions about affect (arousal and valence), two questions about the work conditions (resources and demands), and one question about the stress levels. Figure 46 shows the distribution of stress levels for the emotional (left) and work-related (right) questions. Each dot represents a separate self-report and its color indicates the stress level, ranging from dark red indicating the highest stress level (5) to dark blue indicating the lowest stress level (1). The average ratings for stress, arousal, valence, resources and demands were 2.57/5, 56.7/100, 56.9/100, 51.7/100 and 57.2/100, respectively. As can be seen on the top two graphs of Figure 46, stress ratings were negatively correlated with valence ($r = -0.5, p<0.01$) and resources ($r = -0.24, p<0.01$) and positively correlated with demands ($r = 0.45, p<0.01$). The correlation with arousal was not significant ($r = 0.07, p: 0.13$) which is to be expected as both high and low levels of stress can be associated with high levels of arousal (e.g., frustration and joy, respectively). Similar correlations were found when comparing daily stress levels with daily valence ($r = -0.59, p<0.01$), arousal ($r = -0.13, p = 0.27$), resources ($r = -0.33, p<0.01$) and demands ($r = 0.5, p<0.01$) provided at the end of each day. Moreover, the correlation between the daily four-item PSS and the daily stress reports were significant ($r = 0.37, p<0.01$), indicating consistency across surveys.

When considering the correlations across the other components, we found the following significant correlations: valence and arousal ($r = 0.26, p<0.01$), valence and resources ($r = 0.45, p<0.01$), arousal and resources ($r = 0.34, p<0.01$), arousal and demands ($r = 0.36, p<0.01$), and resources and demands ($r = 0.17, p<0.01$). Similar correlations were found when comparing the daily responses: valence and arousal ($r = 0.59, p<0.01$), valence and resources ($r = 0.58, p<0.01$), arousal and resources ($r = 0.54, p<0.01$), and arousal and demands ($r = 0.33, p<0.01$). However, the correlation between demand and resources in a daily basis was not significantly correlated ($r = 0.11, p = 0.34$). Finally, the four-item PSS collected at the end of each day was significantly correlated with daily valence ($r = -0.59, p<0.01$), arousal ($r = -0.36, p<0.01$), and resources ($r = -0.48, p<0.01$).
At the end of the study participants were asked to rate how difficult they found it was to answer to each of the prompt questions from 1 ("Very challenging") to 5 ("Very easy"). The average ratings for arousal, valence and demands were 3.87, 3.93 and 3.33, respectively, indicating that they were relatively easy to answer. However, the average rating for resources was significantly lower 2.60 (Two-sample t-Test, $t(73) = 3.49$, $p<0.01$), indicating that participants experienced more difficulty quantifying the amount of resources.

Finally, Figure 47 shows the distribution of stress ratings for different days of the week. To better capture the temporal dynamics, the reports were normalized for each person (between 0 and 1) before its visualization. While no significant differences were observed across days, Tuesdays seemed to yield the highest stress ratings. Interestingly, this finding is consistent with previous studies that considered self-reported emotional ratings of citizens in the UK (MacKerron & Mourato, 2012) and the intensity of smiles of those on a college campus (see Figure 10 and Hernandez et al., 2012).

![Figure 47: Average self-reported stress levels for different days of the week (Monday to Friday from left to right, respectively)](image-url)
8.4 Data Preparation

This section overviews the wearable data and the different signals that we used to capture the stress response.

8.4.1 Wearable Devices

Participants in the study carried seven devices during five regular days of work. Each of these devices contained multiple sensors that captured different types of information. The types of information can be grouped into the following categories:

**Physiological.** There were four different types of physiological signals captured during the study: electrocardiography (ECG) and respiratory motions (RESP) from the chest, and electrodermal activity (EDA) and skin temperature (ST) from the wrists. From the ECG and RESP signals, we extracted continuous heart rate (HR) and heart rate variability (HRV), and breathing rate (BR), respectively. These signals were automatically generated by the BioPatch device which leverages motion data to minimize potential artifacts. In the case of HRV, the BioPatch computes the standard deviation between R ECG peaks with their proprietary algorithm. The EDA signal was further decomposed into two different components: a phasic component which shows quick changes associated with stimulus-specific or nonspecific responses, and a tonic level which changes more slowly and can be observed in the absence of any particular discrete environmental event or external stimuli. The decomposition was made with the deconvolution approach proposed by Benedek & Kaernbach (2010), and offers the opportunity to better capture different parameters of the stress response. For instance, high levels of stress are usually associated with high tonic level and a large number of peaks (Boucsein, 2012). Figure 5 shows an example of EDA decomposition. While we captured EDA and TEMP from both wrists, this work uses the signals from the right wrist for most of the data collection days. However, due to sensor malfunction and detachment of electrodes, we used the signals from the left wrist for four days of one participant and one day of another participant. All but one of the participants were right-handed.
**Behavioral.** Most of the wearable devices of this study incorporated motion sensors such as accelerometers and/or gyroscopes. To capture behavior, this work extracts activity level using the accelerometers at four different body locations (chest, wrist, pocket and head). By considering different body locations, it offers the opportunity to capture different types of activities or behaviors. For instance, typing on the keyboard can be captured with the wrist sensors and subtle head gestures can be captured with the head sensors. Given motion data from one accelerometer, activity level was computed as follows: 1) an averaging filter was subtracted from each of the 3-axis components to remove trends and signal drifts, 2) the three components were aggregated with the square root summation of the squared components, and 3) an averaging filter of 20 seconds was applied to the final signal. Figure 48 shows an example of raw accelerometer data and estimated activity level. To minimize discomfort by the devices, the wrist-worn smartwatch was placed on the non-dominant hand.

**Contextual.** A subset of the devices collected information about the environment and work setting of the person. In particular, the smartphone captured ambient temperature, humidity, and atmospheric pressure, and the head-mounted device captured the amount of light in front of the device. An averaging filter of 10 seconds was applied to each of these signals to better reflect temporal dynamics. Additionally, at the end of each day, participants provided manual annotations about their posture (sitting, standing, others involving more activity), whether they were interacting with someone (no, remotely, in person), and the type of social interaction (with family members, with colleagues, with their boss) for the moments of the day when they were prompted to self-report their affective and work-related information. To minimize recall problems, participants were provided a tool to review the images captured by the wearable camera during the day. These features were incorporated as discrete features for part of the analysis.

Table 17 summarizes all the physiological, behavioral and contextual signals considered on this study.
Figure 48: Example of Glass accelerometer data and estimated activity level.

Table 17: Overview of the different signals and features used in the context of stress recognition (R: Reported).

<table>
<thead>
<tr>
<th>Category</th>
<th>Wearable signals</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physiology</td>
<td>EDA Tonic</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td>EDA Phasic</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td></td>
<td>Heart Rate</td>
<td>Area under the curve</td>
</tr>
<tr>
<td></td>
<td>Heart Rate Variability</td>
<td>Range of Values</td>
</tr>
<tr>
<td></td>
<td>Breathing Rate</td>
<td>Maximum Value</td>
</tr>
<tr>
<td></td>
<td>Skin Temperature</td>
<td>Minimum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Position of Min. Value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Position of Max. Value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Slope</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of Peaks</td>
</tr>
<tr>
<td>Behavior</td>
<td>Chest Activity</td>
<td>Avg. Amplitude of Peaks</td>
</tr>
<tr>
<td></td>
<td>Head Activity</td>
<td>Avg. Distance of Peaks</td>
</tr>
<tr>
<td></td>
<td>Wrist Activity</td>
<td>Number of Zero Crossings</td>
</tr>
<tr>
<td></td>
<td>Pocket Activity</td>
<td>Avg. Dist. of Zero Crossings</td>
</tr>
<tr>
<td>Context</td>
<td>Ambient Temperature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Atmospheric Pressure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Humidity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Light</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Posture (R)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social Interaction (R)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type of Interaction (R)</td>
<td></td>
</tr>
</tbody>
</table>

8.4.2 Physiology Preprocessing

Quick sensor movements may introduce noise in the physiological measures in the form of high frequency changes. This problem is very common in uncontrolled settings as these often involve gestures and body movements. To attenuate these artifacts, we applied a Hanning filter with a 1 second window to the EDA responses (before decomposition).
In the case of heart (HR) and breathing rate (BR) estimations, the BioPatch provided indices quantifying the quality of the signals, ranging from 0% (bad quality) to 100% (good quality). In our analysis, segments of data with a quality lower than 50% were excluded from the analysis. In the case of HRV, the final estimations were also visually inspected to ensure accurate assessments. Segments of HRV data containing abrupt amplitude decays were excluded from the analysis.

Finally, there were occasional quantization problems on the TEMP signals recorded by the Affectiva Q™ sensor. To minimize this problem, an averaging window of 10 seconds was applied.

8.4.3 Feature Extraction

During the study, we collected a total of 516 successfully submitted prompts. For each of these prompts, we time aligned the different signals and focused on the analysis of the five minute segment leading up to the prompt. Thus, the goal of this chapter is to infer self-reported stress levels from these 5-minute segments of multi-modal sensor data.

Before the previous signals can be used for classification, it is necessary to extract representative features. In this work, we explored the utility of the following features: mean of the signal, standard deviation, area under the curve, range of values, maximum and minimum values and their positions, slope, number of peaks, average amplitude of the detected peaks, average distance between the peaks, number of zero crossings after mean subtraction, and average distance between the crossings. To detect the peaks we used the FINDPEAKS function of MATLAB with different parameters for each of the wearable signals (see Table 18). These parameters were arbitrarily selected based on the properties of each signal (e.g., dynamic range) to ensure capturing relevant temporal changes. Future work will consider the automatic selection of these parameters based on different data statistics (e.g., mean, range of values).
Table 18: Peak detection parameters used for each of the wearable signals with the MATLAB FINDPEAK function

<table>
<thead>
<tr>
<th>Wearable signal</th>
<th>Minimum amplitude</th>
<th>Distance (seconds)</th>
<th>Minimum prominence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tonic &amp; Phasic EDA</td>
<td>0.01</td>
<td>3</td>
<td>0.01</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Heart Rate Variability</td>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Breathing Rate</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Skin Temperature</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>Activity Level</td>
<td>0.1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>Ambient Temperature</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>Atmospheric Pressure</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>Humidity</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>Ambient Light</td>
<td>0.1</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 49: Summary of features extracted from the wearable signals

While some of the features aim to capture the temporal aspects of the responses (e.g., slope captures an overall increase or decrease of the response), other features aim to capture overall activation throughout the period (e.g., average number of peaks can be seen as an indicator of physiological arousal). Similar features have been successfully used in previous studies to recognize engagement of TV viewers and social interactions.
8.4.4 Feature Normalization

The goal of this work is to infer self-reported stress levels from wearable data. In particular, we focus on the data captured during the five minutes before the prompt. However, as discussed earlier, the range of values for each feature may vary for each day of data collection due to different sensor placements or different daily routines. To address this problem, we normalized each of the features with the wearable data collected throughout the day. In particular, we modified the range of each feature to be between zero and one for each day, and then we quantified how much each deviated from the daily averages. To do so, we first segmented all the wearable data into non-overlapping segments of five minutes and extracted the same features used for the analysis. Once all the features were computed, we calculated the daily averages and the maximum and minimum observed values for each feature and person. We then subtracted the average from the feature to be normalized and then rescaled it between zero and one with the maximum and minimum daily values. Besides minimizing daily and person variance, this process also provides the same weight to each of the features and minimizes daily differences (e.g., different electrodermal activity ranges each day). Finally, we subtracted the daily mean from each of the features, helping reflect deviations from the normally observed features.

8.4.5 Refining Stress Ratings and Outlier Correction

While most of the time participants responded in a way that was consistent with the definition of stress we asked them to use when reporting stress (negative emotions and high workload, section 4.4), there were a few outliers that did not follow the same trend. For instance, one of the participants reported high stress levels when experiencing positive emotions (red dots with positive valence on the top-left graph of Figure 46). These outliers could have happened due to a variety of reasons. For instance, the participant could have forgotten the definition we provided and/or s/he could have accidentally submitted the wrong answer. To minimize the presence of these outliers we followed a regression approach. In particular, we used the affective and work-related answers (valence, arousal,
demands, and resources) to fit a linear regression model with quadratic components that predicts the self-reported stress levels. Once the model is learned, we use it to obtain new stress levels from the original affective and work-related questions. The bottom graphs of Figure 46 show the new distribution of derived self-reported stress levels for each of the different affective and work-related questions. As can be seen, most of the outliers have been corrected. Moreover, this approach also provides continuous values for stress (instead of the five discrete values) which is more convenient for performing fine-grained analysis of the data. Finally, as it is based on the answers of five different questions instead of a single one, it minimizes the risks of accidental submissions or ambiguous answer to a specific question. Appendix D contains the regression coefficients. Note that we used quadratic components to better capture the relationship between arousal, valence and stress ("U" quadratic shape).

8.5 Experimental Setting

8.5.1 Class Labels

This work studies the predictive power of different types of wearable data to address the problem of stress recognition "in the wild." To study this, we tackle the problem as a binary classification task, in which self-reports of lower stress ratings were grouped into the "low stress" class and self-reports with higher ratings from the same person were grouped into the "high stress" class. This approach yielded a balanced class distribution which simplifies part of the challenges of real-life classification (e.g., biases towards more frequent classes). Data from each participant was used as a separate classification problem which helps better study individual differences. Furthermore, we excluded 10% of the self-reports whose ratings were around the average in terms of stress, further amplifying the differences between the two classes. Therefore, the final number of prompts used for the analysis was 460 which were equally split into two classes for each of the participants. The average number of samples per participant for the two classes combined was 30 (STD = 5.2), the minimum was 22 and the maximum was 40.
8.5.2 Classification

To perform classification, we used the LIBSVM library (Chang & Lin, 2001), which provides an efficient implementation of Support Vector Machines (SVMs; Boser, Guyon & Vapnik, 1992). We used a 10-fold-cross-validation protocol for testing and training of the algorithm. Therefore, we divided the sessions into 10 different groups and used 9 of them as a training set and the remaining one as the testing set. This process was iteratively repeated until stress labels were automatically generated for all the groups. During the training phase, the training set was divided into 10 different groups and followed the same iterative process to gather performance scores for different misclassification costs \((\log_2 C, \text{for } C = \{-10, -9, -8, ..., 25\})\) of a Linear SVMs with probabilistic estimates. Once the process was completed, we used the whole training set and the best misclassification cost to obtain the final classifier model, which then was used in the testing set. The whole process was repeated 20 times with a randomized selection of the groups to obtain an average recognition score. This number of iterations ensured convergence of the results.

8.5.3 Missing Data

Collecting good quality wearable data in-situ is a considerable challenge. During our study, we followed several strategies to address, or help avoid, problems. For instance, at the beginning of each day, researchers synchronized and charged the different devices and helped participants put the sensors on. Also, some of the sensors used sticky gelled electrodes to minimize the impact of motion artifacts. When considering the Electrodermal Activity readings, this approach dramatically improved the quality of data versus previous efforts which explored dry electrodes and conductive gel on the electrodes (Hernandez et al., 2014). Finally, wearable sensor data was also visually inspected at the end of each day to quickly detect any potential issues (e.g., battery problems, loose electrodes, broken sensors).

Despite these efforts, there were a few instances in which parts of the sensors malfunctioned. In particular, the logging application on the smartphone did not work for one day of data collection of one participant and the BioPatch was accidentally switched off by another participant on a different date. Moreover, participants sporadically removed
some of the devices during the study for several reasons (e.g., going to the restroom, charging sensors). As a result, there were several samples of our dataset for which part of the signals were missing or the quality was too poor to provide relevant stress value. To help quantify the problem of missing data in our dataset, Table 19 shows a summary of the percentage of missing samples for each of the participants, being the letters M and F indicative of their gender (male and female, respectively). Figure 50 shows a bar graph with the average missing values for each of the participants.

When considering the whole dataset, there was a total of around 7% of features missing. This is very small considering our experiment occurred in real-life settings and highlights the effectiveness of our methods towards minimizing missing data. As can be seen, both EDA features as well as TEMP contained zero missing values which was partly due to the effectiveness of the sticky electrodes of the QTM sensor and the fact that we could leverage the readings of a second device when one was not working. This was also the case for the contextual reports as participants were required to provide annotations as part of the study. Among all considered signals, BR was the one that yielded the most missing values, averaging 40% missing data across all participants. There were two main factors that influenced this. First, the specific form-factor of the BioPatch relied on impedance-based measurements which are not as accurate as traditional bands around the chest. Second, accurately measuring breathing rate during daily life is very challenging due to motion artifacts. For instance, frequent activities such as speaking can quickly interrupt or obscure respiratory motions. Furthermore, due to its low frequency movements, longer periods of still data are usually required. Future work will consider enhancing the impedance-based measurements with the other wearable motion sensors. The second worst signal was HRV, yielding missing data for around 20% of the samples. Similarly, HRV suffers from the same problems as BR (e.g., motion artifacts, requirement of long-periods of still data). However, the proprietary algorithms for HRV (discussed in section 8.4.1) already incorporate motion data to help minimize the amount of missing data.

Finally, when comparing the missing data of motion sensors across body locations, both the head and wrist locations yielded more missing data (around 7.5%) in comparison with the phone (2.5%) and the chest (1.7%). This is directly associated with the times when participants removed some of the sensors. Interestingly, there were significant gender
differences when comparing the amount of missing samples for the wrist. The female population provided less missing data. Moreover, the missing data across participants for the same sensor was significantly correlated with the conscientiousness dimension of the Big Five ($r = 0.54$, $p<0.01$), indicating that there is a relationship between people scoring higher on conscientiousness and how well they kept the wrist device charged and working.

While there are many approaches to deal with missing data (e.g., Ding & Simonoff, 2010), this work follows a personalized imputation approach. In particular, missing features of one person are replaced by the average of good quality features of the same person. This approach enabled us to not only to keep the same sample size when comparing different wearable signals, but also offers the opportunity to show the impact of missing data in terms of classification performance. Future efforts will consider the effects of alternative imputation methods as well as alternative methods that can handle missing data without imputation.

### 8.5.4 Performance

Two of the most common methods to evaluate performance of a classifier are the area under the Receiver Operating Characteristic (ROC) curve, and the area under the Precision/Recall (PR). In this work we use the average between the areas under the two curves obtained when considering different thresholds on the probability estimates provided by the SVMs. This metric was used as a reference to find the optimal misclassification cost of the SVMs during the training phases and is used in the following section to report classification performance. In particular, this metric ranges from 0 (worst performance) to 100 (maximum performance) and better captures overall discriminative power of a classifier (i.e., different thresholds on the probability estimates) instead of a static one like accuracy. Note that a classifier that only predicts one class or the most likely one will obtain a performance of 0 as none of the curves can be computed (i.e., probability estimates are always one).
Table 19: Missing data (%) for each of the participants and wearable signals. Values over 25% are highlighted in red bold

<table>
<thead>
<tr>
<th>Physiology</th>
<th>Behavior</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>M8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F6</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>F7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Max.</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 20: Classification scores (%) for each of the participants and wearable signals. Scores over 65% are highlighted in green bold. (HR: Heart Rate, HRV: Heart Rate Variability, BR: Breathing Rate, TEMP: Skin Temperature, Hum.: Humidity, Temp.: Ambient Temperature, Press.: Atmospheric Pressure)
Figure 50: Average percentage of missing samples and standard errors (green bar) across all participants for each of the signals.

Figure 51: Average classification scores (%) and standard errors (green bar) across all the participants for each of the signals considered individually. (HR: Heart Rate, HRV: Heart Rate Variability, BR: Breathing Rate, TEMP: Skin Temperature, Hum.: Humidity, Temp.: Ambient Temperature, Press.: Atmospheric Pressure)
8.6 Results

Table 20 shows the predictive values obtained for each of the participants using the different physiological, behavioral and contextual wearable signals. Scores above 65% are highlighted in green bold numbers to help identify the best performing signals. As can be seen, there is large variability across the table and only a limited subset of the signals worked well for each participant. On average, there were around three (STD = 1.7) individual signals per person which performed above 65%. The maximum was six signals (M1 and M3) and the minimum was 0 for F5 (i.e. no wearable signals yielded an accuracy above 65%). While the distribution of the results changed from person to person, they seemed to be similarly distributed across the different types of signals (physiological, behavioral and contextual), indicating that very different types of wearable data can help capture different aspects of the stress response. Figure 51 shows a bar graph with the average classification scores for each of the participants.

When considering the physiological signals (left block), both EDA features outperformed the other physiological signals. On the other hand, HR yielded the lowest average score (54%) even though it was the best physiological signal for M1 (75%). Between the two EDA features, the tonic component slightly outperformed the phasic EDA (59.2% and 58.6%, respectively). However, the latter was the best performing physiological signal for five participants (M2, M3, F1, F2 and F3), two more than the tonic component, which was the best signal for only three participants (M5, M7 and M8). While the sample size is relatively small, this finding suggests that phasic features may be a more generalizable predictor of stress across participants and genders.

To better understand the variability across the table, we consider the EDA features. In this case, both tonic and phasic components provided high recognition rates for some of the participants (e.g., M5, F2) but yielded poor performance on others (e.g., M6, F5). Figure 52 shows a representative EDA response (before decomposition) for one person from each group (M5 and M6). As can be seen, even though both participants reported experiencing stressful events during the day, their physiological responses are very different. The tonic and phasic components can be easily distinguished in one of the EDA signals, while the other signal is less responsive almost the whole day. Note that the scale
of each graph has been modified to amplify small changes. A signal with small amplitude dynamics makes it challenging to capture relevant physiological changes. There are several factors that can contribute to these differences. For instance, different people have different densities of electrodermal glands and the differences can be exaggerated when considering the upper wrist as a measurement location. Moreover, there are well-studied physiological traits that mediate the different types of responses such as EDA lability (Mundy-Castle & McKiever, 1953). Scoring high on this trait, for instance, is associated with more frequent occurrences of nonspecific EDA responses at rest and more peaked-looking responses. Similarly, scoring low on EDA lability (a.k.a., EDA stability) is associated with less frequent nonspecific EDA responses and more flat-looking responses for most of the day. In our case, the second type of signals yielded the worst recognition results. These differences provide further support that trait physiological features such as EDA lability could be used as potential indicators to select relevant features. For an example of this approach, see the exploration in section 3.1.2.

While the maximum recognition rates for HR and BR were above 65% (75% and 69%, respectively), they were only the best performing physiological modality for one participant each (M1 and F7, respectively). This indicates that they may not be a very generalizable indicator of stress, especially when used alone. In contrast, HRV was the best performing modality for three participants of the study (M4, M6 and F6). Interestingly, the results obtained for BR and HRV were inversely correlated for several participants. In other
words, for cases when HRV did not provide recognition value, BR performed well (e.g., M8, F4, F5, F7). On the other hand, for cases when BR did not provide recognition value, HRV performed well (e.g., M1, M3, M6, M7, F6). Moreover, there seemed to be significant gender differences when considering BR signals (Two-sample t-Test, \( t(13) = 2.23, p = 0.04 \)), with the female population yielding better recognition rates. While all these findings require a more thorough analysis with larger datasets, they seem to support that there are complementary signals (HRV and BR) and that certain parameters such as gender could be used to help select more adequate features.

When considering the behavioral signals (center columns), the level of head motion yielded the highest predictive value (59.1%), providing comparable results to the tonic component of EDA. Similarly, while the results of these two are correlated for some cases, there are other instances for which they seem complementary (e.g., M4, M6). It is important to note that the extent of motion of the head indirectly captures head gestures (e.g., head nods, shakes), facial expressions (e.g., frowning), and eye-behaviors (e.g., blinking) which have been studied in the context of emotion sensing (e.g., Scheirer et al., 1999; El Kaliouby & Robinson, 2005; Hernandez et al., 2013). Future work will consider extracting more complex features that can differentiate between these cues. The amount of wrist motion was the lowest performing behavioral signal. This is partly to be expected, as the range of wrist motions during daily life is considerably larger than the other locations. Therefore, more complex features that aim at recognizing specific activities may be more relevant (e.g., typing at the computer, walking). Despite the difference in performance, both wrist and head motions were found to be the most discriminative behavioral metrics for four of the participants each (M1, F2, F4, F6, and M2, M3, M4, and M7, respectively). On the other hand, the motion from the pocket was found to be the most discriminative behavioral signal for five of the participants (M5, M6, F3, F5, and F7) and the chest for only two (M8 and F1). This variability suggests that building generalizable models that can work on other people is challenging and that multiple channels of information should be considered.

When considering the contextual signals (right block), atmospheric pressure was the one that yielded the highest recognition rates (58.41%). This finding could provide partial support to previous work studying the relationship between rainy weather (usually
associated with lower atmospheric pressure) and mood (e.g., Barnston, 1988; Keller et al., 2005; Klimstra et al., 2011). However, it is important to note that the duration of the study was relatively short (five days) and, therefore, this finding may be dependent on the specific weather conditions when it was performed. In contrast, ambient temperature yielded the lowest average recognition value across all signals (including physiological and behavioral signals). The poor recognition value could be due to several factors. For instance, it may be the case that the features we used in this work (e.g., average, slope) may not be appropriate for this type of signal (which would be consistent with the poor performance of skin temperature). Moreover, participants spent most of the time of the data collection in the same building which temperature is globally controlled and remained relatively constant throughout the study.

Finally, when considering the maximum value achieved by each of the signals across participants (bottom row), all of them were above 65% which indicates that each of the computed signals provides insightful recognition information for at least one of the participants. Moreover, when considering the best performing signal across all the different signals for each person (last column), the average recognition rate was 72%, which is comparable to our previous workplace study (section 3.1.2). Note, however, that these two results assume that the classifiers have prior knowledge about what signals are more descriptive for each person and this may not be readily available in real-life applications. Future work will consider combinations of features and the development of tools that can automatically learn what subsets of features are better for each person.

8.7 Discussion

The motivation for this study is to advance the measurement of workplace stress with wearable devices. To further explore this idea we collected multiple signals of wearable data from 15 participants during five regular days of work, and asked them to self-report affective and work-related information several times a day. While our first attempt was to build a one-size-fits-all method (trained and tested with data of different people), the variance across participants and real-life settings proved to be too large. Based on this, we followed an idiographic approach to better examine the predictive value of different wearable signals.
As we collected completely unsupervised real-life wearable data, a number of problems arose. Encountering these problems helped us study some of the main challenges of real-life stress measurement. In the context of self-reported stress measurement, this work demonstrates that asking several quick affective and work-related questions can help better capture stress. For instance, we requested participants to only provide high scores when experiencing negative stress and high work overload. However, there were some high stress self-reports with positive valence and/or low demands. By relying on the responses of other questions (e.g., valence, arousal, demands, resources), we were able to effectively correct some of these responses. Furthermore, we could advance the understanding of how the different affective and work-related components are connected.

Our work provided a quantification of the amount of missing data per signal, information that is critical to understanding what types of devices and signals are more appropriate. In our study, both BR and HRV yielded the highest amounts of missing data despite the use of sticky electrodes on the chest. In contrast, sticky electrodes on the Q™ sensor greatly helped to minimize the amount of motion artifacts and missing data in the EDA signals. We have also found an intriguing negative correlation between conscientiousness and the amount of missing data for the wrist, suggesting that personality may play an important role when handling and caring for the devices (e.g., charging batteries). Overall, applying a variety of approaches (e.g., daily interactions, sticky electrodes on the wrist, sensor redundancy) we were able to effectively minimize the problem of missing data.

The main focus of this work has been the evaluation of the predictive value of several types of wearable signals. In particular, we extracted physiological (e.g., EDA, HR, HRV), behavioral (e.g., activity level at the head) and contextual (e.g., environmental and social interaction reports) signals. Previously, these have been mostly studied in controlled laboratory settings. To our knowledge this is the first study to collect so many wearable data signals and compare them in the same longitudinal study. Our results demonstrate that different types of signals work for different people. When considering averages across participants, we have seen that electrodermal activity features such as tonic and phasic components yielded higher scores than other traditional stress metrics (e.g., HRV). Among the behavioral cues, head motions yielded similar performance and outperformed other
traditional locations for motion sensors (pocket, wrist and chest). Finally, among the contextual cues, we found atmospheric pressure to be the most descriptive. However, environmental signals such as this one may be more dependent on when the data were collected.

While we have started to explore some of the variance across participants and signals (e.g., EDA lability), there is still plenty of work to be done. This first analysis has focused on the recognition value of different wearable signals separately. However, much more value can be obtained when combining different types of data. To do so while avoiding overfitting it is important to implement feature selection protocols guided by some of the findings of this work. For instance, EDA lability or gender can be used to decide whether EDA features or BR are appropriate. While we have started exploring this space, a data collection with a larger number of participants and more longitudinal analysis can help draw stronger conclusions. By performing longer studies, we can increase the chances of capturing more genuinely stressful events during the study and evaluate whether the findings are generalizable over different periods of time. For instance, biologic rhythms such as the menstrual cycle can influence physiological baselines. By capturing data from more participants, we can continue to learn about why some signals work better for certain people than for others. Such analysis is critical to build a more generalizable system. Moreover, once the sample size is increased, we can start considering more sophisticated methods (e.g., RBF kernels, features selection) whilst still avoiding overfitting.

Overall, the results of this work are very encouraging and demonstrate that a single signal modality such as tonic EDA or head motions can yield an average recognition rate of 60% in real-life when previous information about the person is available. There are, however, many opportunities for improvement (e.g., multimodal approaches, feature selection), especially when considering the unique individual differences of each person’s data and traits.
8.8 Conclusions

This chapter describes a real-life stress analysis from wearable data considering multiple physiological, behavioral and contextual signals. To our knowledge this is the first study to use a head-mounted wearable device in the context of real-life stress recognition and to collect and compare a variety of real-life wearable signals in the context of stress. The results of this work demonstrate that real-life stress recognition from wearable data is possible and a significant increase in performance can be obtained when certain information is available (e.g., best feature set is known). There are, however, a considerable number of challenges that need to be addressed before we can have a one-size-fits-all solution. Future efforts will focus on collecting data for longer periods of time and for more people to help address some of these challenges. We are looking towards a future when wearable devices can help not only measure daily stress but also to prevent its negative health outcomes.
Chapter 9

Conclusions

This thesis has presented novel algorithms and methods to help advance the understanding and measurement of stress in real-life. To do so, we have studied some of the main challenges of stress measurement and leveraged the advances of existing wearable devices to help address them. This final chapter provides a general discussion of the work, a summary of the different contributions, potential areas of improvement and future work, and concluding remarks.

9.1 Discussion

This thesis has studied how different types of wearable devices can help advance three challenging areas of real-life stress measurement: 1) collection of ambulatory self-reports, 2) in-situ physiological measurement, and 3) automated stress recognition from wearable data. To further explore these, we performed several explorations, controlled experiments, and a real-life workplace study involving 15 individuals carrying seven wearable devices during five typical days of work. Based on these efforts, this section summarizes some of the main findings.

Gathering accurate self-reported stress levels in real-life is vital for improving our understanding of real-life stress. As reviewed in section 2.4, there are several challenges associated with eliciting self-report data in real-life settings. One of the challenges is delivering noticeable prompts without the task being too disruptive. To help tackle some of the challenges, we created a custom experience sampling tool that can be deployed on many different Android platforms (including: head and wrist-worn devices, smartphones
and tablets). Through a large user experience study we evaluated how different device form-factors can significantly affect the reporting process. In our real-life study, we found that using either wrist-worn or head-worn wearable devices led to significant reductions in the amount of missed interruptions in comparison with the traditional phone inside the pocket. Moreover, we have shown that using wrist-worn or head-worn devices can lead to a significant reduction in the time between triggering of prompts and the first interaction with the application, helping minimize some of the physiological, recall and cognitive biases associated with longer response times. However, both wrist and head-worn devices have their own limitations. In the case of wrist-worn devices, we have found that the limited screen size can affect some of the responses such as 2D-Grid based questions. In the case of the head-worn device, several participants experienced social and physical discomfort due to its novel form-factors and other limitations. Moreover, we have shown that novel interactions such as head gestures and finger swipes do not affect the distribution of the responses but can significantly slow down the reporting process. Since the goal of our study is to measure stress levels “in the wild,” the wrist-worn wearable device seemed to be the most adequate for our purposes.

Another vital aspect towards measuring stress is the development of methods that can comfortably measure changes associated with stress such as physiological signals. Traditional methods for accurately measuring physiological parameters (e.g., heart and breathing rates) in real-life usually require sticky electrodes or bands around the chest. These approaches can be cumbersome (e.g., skin irritations) and burden the user (e.g., replacing electrodes). Moreover, the costs associated with these types of devices are usually high, limiting the potential scalability of such approaches. To help minimize these challenges we leveraged wearable motion sensors, which are amongst the most pervasive, low-cost and energy-efficient sensors available in the market. While motions sensors have been mostly used to capture and understand behavioral and contextual motions (e.g., steps, activities), this work demonstrates that they can also capture subtle and periodic motions associated with the beating of the heart and contraction/dilation of the chest resulting from respiration when the user is relatively “still.” To study this, we performed three laboratory studies focusing on three well differentiated devices (head-mounted, wrist-worn and smartphone devices), and created two novel methods (one for heart rate and another for
breathing rate estimation) and modified them to appropriately capture and analyze the subtle motions of each body location. Among some of the main findings, we have shown that three types of devices can capture relevant physiological parameters whilst participants stay in relatively sedentary positions (standing, lying down, and sitting down before and after physical exercise) and found the heart and breathing rate estimations were as accurate as the ones provided by FDA-cleared devices, even when the motion sensors were far from the chest. In particular, the head-worn wearable device provided the most accurate heart rate assessments (mean absolute error of 0.82 beats per minute) followed by the wrist-worn device (1.39 beats per minute) and then the phone (2.19 beats per minute). In the context of breathing rate estimation, the wrist-worn device outperformed the others (0.38 breaths per minute), followed by the phone (1.13 breaths per minute), and then the head-worn device (1.16 breaths per minute). We have also compared different motion-based sensors and their combinations, and found that gyroscopes (previously unexplored in this context) outperformed and complemented sensors such as accelerometers. Moreover, we have shown that other motion-sensitive sensors such as wearable cameras can also indirectly capture the vital signs of the wearer. Finally, we have evaluated the generalization of the previous findings in the context of our large real-life study and found comparable results for relatively “still” moments of the day. In particular, we found in a group of 15 office workers that the head-worn wearable device could make heart rate assessments around 20.6% of the day, followed by the phone (6%), and then the wrist-worn device (2.7%). These findings demonstrate that motion-based wearable sensors can provide opportunistic real-life physiological measurements during daily life while being more comfortable and lower-cost than other traditional sensors.

A final step towards the measurement of stress in real-life is to connect the previous two approaches to leverage the different benefits from each. To explore this, we leveraged supervised machine learning methods to automatically infer self-reported stress levels from wearable data collected in our real-life workplace study. In particular, we identified and extracted several physiological, behavioral and contextual metrics relevant in the context of stress and examined their predictive value for each of the participants. To do so, we first analyzed the surveys and self-reported questions we used to capture stress. While the surveys and ratings were correlated in the directions we expected (e.g., higher self-reported
stress levels associated with negative valence and decreased resources), we still found occasional cases when the different answers were incongruent (e.g., higher stress-level with positive valence). To partially address this, we leveraged other relevant affective (arousal and valence) and work-related (resources and demands) self-reports to derive refined stress ratings. We also quantified the amount of missing data for each of the different types of signals and found that heart rate variability and breathing rates were significantly more challenging to be accurately measured in real-life, despite using the sticky pre-gelled electrodes attached to the chest in order to improve the signal-to-noise ratios. In the context of stress recognition, we observed large variability across participants of our study and found that what worked best for one person did not necessarily work well for another person. When considering the average stress recognition results across all the participants (using a combination of area under the ROC and PR curves), we found that electrodermal features (both tonic and phasic components) were the best performing physiological metrics, the activity level of the head was the best performing behavioral signal (versus chest, wrist and pocket), and atmospheric pressure was the best performing contextual signal. While the work presented here is only a first step towards a more thorough analysis, the findings of this study will help inform what signals and wearable devices may be more appropriate in the future. In our case, wrist-worn and head-worn devices provided similar average performance across participants when considering a single modality and person-dependent classification models. There are, however, many more opportunities to extend this work and better understand real-life stress. Future directions are explained in the following section.

Overall, this work has systematically demonstrated that different wearable devices offer a different set of advantages that can help partially tackle some of the challenges associated with real-life stress measurement, moving us one step closer towards that goal. While mainly focused on stress, the findings of this work are also relevant in other research areas. In the case of self-reported data collection, experience sampling is widely used not only to capture emotion reports but also other factors such as behavior and personality traits. Our findings can be used to help capture accurate information and help minimize the burdens associated with experience sampling during daily life. In the case of physiological sensing, the methods we have developed are useful not only in the context of stress
measurement but also to help enhance health monitoring and care delivery. For instance, resting heart rate has been shown to be associated with different cardiovascular risks (Cook et al., 2006). By using the proposed methods, we could comfortably and passively capture this type of parameter to better track chronic conditions and potentially prevent other negative health outcomes. Finally, in the case of stress recognition from wearable signals, the signal processing techniques and analysis we employed can be used in many different contexts too, especially to help improve the understanding of workplace behavior and study key aspects such as productivity, decision making, and workplace engagement. We are looking forward to a future when wearable devices can accurately capture and analyze information from the wearer not only to improve their wellbeing but also to enhance their daily life.

9.2 Summary of Contributions

This section summarizes the main contributions of this thesis for each of the main topic areas:

**Self-report data collection:**
- Design and development of an experience sampling tool that prompts two different types of questions (2D-Grid, 5-point Likert scale) and can be used on different types of platforms (phone, wrist-worn and head-worn wearable devices).
- Qualitative and quantitative comparisons of experience sampling data in real-life with wrist and head-worn wearable devices as well as the traditional phone in the pocket.

**Physiological sensing:**
- Development of novel multimodal methods for heart and breathing rate estimation from motion signals, and their adaptations to three different body locations (head, wrist and pocket). While previous work has considered other body locations (e.g., chest, below feet, back, ear), this is the first work to consider the head (above the eye), wrist and pocket locations.
• Evaluation of the performance of the previous methods in three separate laboratory studies (12 participants each) and a real-life workplace environment (15 participants, five days each).
• Systematic comparison across different types of motion sensitive sensors (accelerometer, gyroscope and camera) in the context of physiological parameter estimation. While accelerometers have been the preferred choice in previous work, to the best of our knowledge this is the first work to demonstrate that both gyroscopes and wearable cameras can be used in this context and that they can give accurate estimates.
• Comparison of wrist motion-based sensors with traditional light-based methods in the context of heart rate estimation.

Stress recognition from wearable signals:
• Improvement of self-reported stress levels by leveraging other affective and work-related self-reports and regression.
• Comparison of multiple wearable signals captured in a real-life experiment in the context of stress measurement. To the best of our knowledge, this is the first work to demonstrate that a head-worn motion sensor can be used in real-life settings to measure stress. Furthermore, this is the first work to demonstrate the collection and analysis of such a large and varied number of wearable signals in the same study.

9.3 Limitations and Future Work

As the title of this thesis suggests, the work presented here presents a step towards real-life stress measurement using wearable devices. There are of course many ways in which this work can be extended and improved upon. This section describes some of the most promising lines of future research.

The first part of this thesis examines different aspects of experience sampling in a real-life study in which participants carried seven wearable devices during five days of work. To facilitate comparison across devices within each participant, each participant had to carry all the wearable devices for most of the day. While this enabled us to capture a large real-life dataset, carrying so many devices during daily life indirectly influenced some of
the work environment and behavior of participants (e.g., productivity, social interactions). Moreover, some contextual variables of the experiment (e.g., negative press associated with one of the devices, limited use of the devices, and previous experience with some of the devices) are also important when considering the implications of our findings. In our analysis we studied relevant factors such as the response time, the amount of missed interruptions, and social stigma. There are, however, many other relevant factors that are worth exploring. For instance, when and how to trigger the prompts based on context and available devices can be very helpful to minimize overall disruption. In our study, for instance, many participants disliked receiving notifications when interacting with other people. To address this, researchers could use different types of wearable data (e.g., acoustic signals, the number of faces in front of a wearable camera, vibratory motions of the body) to postpone non-critical notifications to more appropriate moments in time. Similarly, there are certain work conditions that require the use of both hands (e.g., soldering, laboratory assays), for which head gestures are more preferable than finger touch interactions. In contrast, touch interactions may be more suitable in other settings as they offer more intimate ways of interaction. This could be important in public settings when there are other people are around (e.g., classroom, meetings). Future work should consider more types of questions and interactions than the ones considered in this research (Grid/Likert scale questions and finger/head gestures interactions, respectively). Finally, this work has combined behavioral data (e.g., response times) and qualitative feedback (e.g., subjective ratings) in order to analyze the differences across devices. However, there are alternate methods that could help provide some additional insights. For instance, physiological signals could be used to help quantify the stress levels elicited by each of the devices in certain conditions (e.g., head-mounted during social interactions). However, as shown in this thesis, physiologically-based approaches introduce a new set of challenges (e.g., individual differences, daily baselines).

In the context of physiological measurement with motion-based sensors, this work has demonstrated that we can accurately estimate heart and breathing rates with motion sensors from different peripheral body locations. While these physiological metrics are a good representation of traditional cardiac and respiratory activity, there are other physiological parameters that may be more appropriate in the context of stress measurement. For
instance, lower heart rate variability has been largely associated with increased stress levels. However, estimating heart rate variability requires very precise detection of continuous heart beats which can be quite challenging when using wearable motion sensors. Another physiological change that has been widely associated with stress is increased blood pressure. While this signal cannot be easily measured from motion sensors alone, recent work (e.g., Winokur, 2014; Mukkamala et al., 2015; Kim et al., 2015) has demonstrated that the combination of light-based (e.g., blood volume pulse) and motion-based (e.g., ballistocardiography) approaches could help capture variables that are highly correlated with blood pressure. In the context of estimation methods, the algorithms presented in this work assume that the person is relatively stationary, limiting the number of potential assessments during the day. Future work will consider the development of more sophisticated methods that can detect different types of motion and selectively amplify the relevant ones (e.g., small amplitude and periodic signals).

This thesis has considered wearable motion sensors located on three different body locations that offer a good representation of the current landscape of existing wearable devices. In the future, however, we expect to have more varied form-factors (e.g., jewelry, ear rings, T-shirt buttons) that could provide some additional benefits. Therefore, it is very important to systematically quantify and model how cardiac and respiratory motions propagate throughout the body. Recent work by Wiens & Inan (2015) is making significant steps towards that goal. Finally, these types of motions also offer the opportunity to passively access personal information. For instance, we have recently shown (Hernandez, McDuff & Picard, 2015) that the same motions used to make heart rate assessments can be further analyzed to provide uniquely identifying information of each person as well as provide information about their posture and their gender. In the future, it is critical that a large effort is devoted towards studying the different privacy questions that this type of research presents such as “How can wearable data be appropriately anonymized?”, “How can wearable users be better informed about how their data are being used?”, and “How do we update privacy policies accordingly?”

Finally, this work has compared a large number of wearable signals in the context of stress recognition. The next step towards extending our analysis is to consider a multimodal approach that combines different types of data (e.g., physiological, contextual and
behavioral) to help better capture the different changes. To achieve that, it is important to develop feature selection methods that not only consider the individual differences but also the similarities across individuals. Successfully developing such methods with our current dataset is challenging, as the sample size is limited in comparison with the large individual differences and variability observed during everyday life. Thus, future efforts will consider gathering wearable data from more people and for longer periods of time. As part of the analysis, we have considered several types of wearable signals in the context of stress recognition. However, there are many others that could be also considered (e.g., roll and pitch of the motion sensors, faces detected in front of the wearable cameras). Moreover, head-mounted devices offer the opportunity of capturing other relevant signals such as eye blinking rate (e.g., Ishimaru et al., 2014). We could not include this method for detecting eye blinking rate in our analysis due to the limited battery of the device. In addition to wearable data, there are other types of data that are worthwhile to consider. For instance, we have shown that there is a subtle correlation between the emotional state of people and days of the week (Tuesdays are the day that yielded higher self-reported stress levels as well as less intense smiles on two separate studies), suggesting that adding the day of the week to the analysis could help improve performance. Additional sources of information (e.g., sleep quality) offer the opportunity to complement wearable data and improve recognition performance. Finally, our study has used general feature descriptors (e.g., mean, slope, number of peaks) in order to be able to consider many types of wearable signals. In the future, we will consider the design of more specialized features that can help capture the unique aspects of each signal.

Overall, there are many opportunities to advance this work and continue exploring the potential of wearable devices. To finally succeed in building stress recognition methods, it is important to leverage the expertise accumulated in other fields such as psychophysiology and psychology, which have been studying stress for over a century. Moreover, there are other relevant research areas that face similar challenges to those described in this work. For instance voice analysis and speech recognition also suffer from large individual differences and noisy signals. The challenges of analyzing highly variable signals have been partially addressed with domain adaptation methods that learn the individual traits of each person by leveraging the data from other people.
Studying stress measurement while writing a Ph.D. thesis is both an educational and paradoxical experience. The ultimate goal of the work is to help minimize unnecessary stress but the process towards achieving it can be quite stressful as well. After devoting much of my recent years to thinking about stress, it has become very apparent that stress (in most of its forms) is a fundamental mechanism that helps us perform our daily activities. Stress helps us not only to circumvent life-threatening situations, but also to keep motivated and persist towards achieving long-term goals. In a way, stress is like a "superpower" that can enhance our bodies and give us extraordinary capabilities when most needed. However, its repeated and long-term experience can also be very taxing and damaging to our bodies, leading to serious negative health outcomes. I strongly believe that accurate workplace stress measurement during daily life is possible and that wearable devices will play a major role towards achieving it. There are, however, several obstacles that will need to be addressed along the way. This thesis has focused on a subset of those and has demonstrated that different types of wearable devices can help improve the gathering of self-reports during the daily activity, provide comfortable and low-cost physiological measurements, and be used to automatically recognize stress levels during daily life, moving us one step closer towards wearable stress measurement.
Appendix

A. Tutorial Provided to Participants

Daily Prompts

During the day you will be asked four questions about your work and emotions. It is important that you provide the answers that most accurately reflect the previous 5 minutes (not the instantaneous moment).

How were you feeling during the previous 5 minutes?

Very energetic/low energetic axis. Energy or arousal is the physiological and psychological state of being awake or reactive to stimuli. Examples of emotions with high energy are angry or excited. Examples of emotions with low energy are relaxed or depressed. Note that high energy can be either positive or negative.

Very pleasant/very unpleasant axis. Pleasantness is the intrinsic attractiveness (positive valence) or aversiveness (negative valence) of an event, object, or situation. Examples of emotions with positive valence are happy or excited. Examples of emotions with negative
valence are angry or depressed. Note that both negative and positive valence can have either higher or low energy.

**What situation best reflects your previous 5 minutes?**

![Diagram of high/low demands and resources]

**High/low demands.** Demands are those physical, psychological, social or organizational aspects of the job that require sustained physical and/or psychological (cognitive and emotional) effort or skills and are therefore associated with certain physiological and/or psychological costs. Examples of high demands are high work pressure, unfavorable physical environment, and emotionally demanding interactions with other people. Note that job demands may not necessarily be negative.

**High/low resources:** Resources are those physical, psychological, social, or organizational aspects of the job that are either/or: functional in achieving work goals, reduce job demands and the associated physiological and psychological costs, stimulate personal growth, learning, and development. Examples of high resources are organizational support, performance feedback, good material, job autonomy, positive climate, etc.

**How stressed were you feeling during the previous 5 minutes?**

![Not at all to Extremely stressed scale]

**Not at all/extremely stressed.** Stress can be defined as a reaction from a calm state to an excited state for the purpose of preserving the integrity of the organism. While there are different types of stress, we focus on the negative emotional feeling associated with work
overload. Examples of potentially stressful situations are giving a speech, submitting a paper or disagreements with the boss. The middle point of the scale would be “Moderately.”

**Was this prompt disruptive?**

Not at all/extremely disruptive. A disruption is a major disturbance, something that changes your plans or interrupts some activity, event or process. Examples of disruptive interruptions could happen when interacting with others and/or interrupting the work flow. The middle point of the scale would be “Moderately.”

**Delaying Prompts**

If you are unable to respond to a specific prompt due to the circumstances (e.g., meeting with the boss), you can either:

- **Ignore.** Once the prompt appears, it will send subtle reminders every 30 seconds. If the prompt is not responded to within 3 minutes, it will be automatically closed. Note that consistently ignoring the prompts may result in a reduction or cancellation of the study compensation.

- **Postpone.** Once the prompt appears, you can choose to delay it for 5 minutes. To postpone it, you will need to perform one of the following actions depending on the device:
  - Swipe down with two fingers
  - Push the side button
  - Push the volume down button
Removing Sensors

If you consider that wearing the sensors could compromise your privacy and/or the privacy of others (e.g., going to the restroom), you can either temporarily remove some of the sensors and/or cover them. The only sensors that can be temporarily removed are the following:

Every time the sensors are removed, you will need to connect them to the USB charger that the experimenter provided. When ready to wear the sensors again, please remember to put them on the same body locations that the experimenter instructed. Please do not remove the rest of the sensors without supervision of the experimenter as it may negatively impact the quality of the data.

Note that failure to wear the sensors most of the work day and/or failure to consistently wear them on the same body locations may result in a reduction or cancellation of the study compensation.
B. Surveys of the Main Study

This section contains all the survey questions asked during the study. The input format for each of the questions is indicated below. * indicates the question is in a 5-Likert scale format. Possible answers or end points are indicated inside parentheses unless specified otherwise.

Big Five Questionnaire

I see myself as someone who...

- ...Is talkative
- ...Tends to find fault with others
- ...Does a thorough job
- ...Is depressed, blue
- ...Is original, comes up with new ideas
- ...Is reserved
- ...Is helpful and unselfish with others
- ...Can be somewhat careless
- ...Is relaxed, handles stress well
- ...Is curious about many different things
- ...Is full of energy
- ...Starts quarrels with others
- ...Is a reliable worker
- ...Can be tense
- ...Is ingenious, a deep thinker
- ...Generates a lot of enthusiasm
- ...Has a forgiving nature
- ...Tends to be disorganized
- ...Worries a lot
- ...Has an active imagination
• ...Tends to be quiet
• ...Is generally trusting
• ...Tends to be lazy
• ...Is emotionally stable, not easily upset
• ...Is inventive
• ...Has an assertive personality
• ...Can be cold and aloof
• ...Perseveres until the task is finished
• ...Can be moody
• ...Values artistic, aesthetic experiences
• ...Is sometimes shy, inhibited
• ...Is considerate and kind to almost everyone
• ...Does things efficiently
• ...Remains calm in tense situations
• ...Prefers work that is routine
• ...Is outgoing, sociable
• ...Is sometimes rude to others
• ...Makes plans and follows through with them
• ...Gets nervous easily
• ...Likes to reflect, play with ideas
• ...Has few artistic interests
• ...Likes to cooperate with others
• ...Is easily distracted
• ...Is sophisticated in art, music, or literature

* All of the questions are in a 5-Likert scale from 1 to 5 with end points “Strongly disagree” and “Strongly agree,” respectively.
Toronto Alexithymia Scale

- I am often confused about what emotion I am feeling.
- It is difficult for me to find the right words for my feelings.
- I have physical sensations that even doctors do not understand.
- I am able to describe my feelings easily.
- I prefer to analyze problems rather than just describe them.
- When I am upset, I do not know if I am sad, frightened, or angry.
- I am often puzzled by sensations in my body.
- I prefer to just let things happen rather than to understand why they turned out that way.
- I have feelings that I cannot quite identify.
- Being in touch with emotions is essential.
- I find it hard to describe how I feel about people.
- People tell me to describe my feelings more.
- I do not know what's going on inside me.
- I often do not know why I am angry.
- I prefer talking to people about their daily activities rather than their feelings.
- I prefer to watch "light" entertainment shows rather than psychological dramas.
- It is difficult for me to reveal my innermost feelings, even to close friends.
- I can feel close to someone, even in moments of silence.
- I find examination of my feelings useful in solving personal problems.
- Looking for hidden meanings in movies or plays distracts from their enjoyment.

* All of the questions are in a 5-Likert scale from 1 to 5 with end points “Strongly disagree” and “Strongly agree,” respectively.
Perceived Stress Scale

At the beginning of the study, participants had to respond to the 10-item Perceived Stress Scale. The questions were as follows:

- In the last month, how often have you felt confident about your ability to handle your personal problems?
- In the last month, how often have you felt that things were going your way?
- In the last month, how often have you found that you could not cope with all the things that you had to do?
- In the last month, how often have you been able to control irritations in your life?
- In the last month, how often have you felt that you were on top of things?
- In the last month, how often have you been angered because of things that were outside of your control?
- In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

* All of the questions are in a 5-Likert scale from 0 to 4 with end points “Never” and “Very often,” respectively.

Prompt Questions

For each of the prompts, participants had to respond to the following questions:

- How were you feeling during the previous 5 minutes? (2D-Grid)
  Horizontal axis: (“Very Unpleasant” and “Very Pleasant”)
  Vertical axis: (“Low Energy” and “Very Energetic”)
- What situation best reflects your previous 5 minutes? (2D-Grid)
  Horizontal axis: (“Low Resources” and “High Resources”)
  Vertical axis: (“Low Demands” and “High Demands”)
- How stressed were you feeling during the previous 5 minutes?
  * (“Not at all” and “Extremely”)
- Was this prompt disruptive?
  * (“Not at all” and “Extremely”)
End of the Day Questions

For each of the prompts during the day, participants had to respond to the following questions:

- What set of activities best describe what you were doing during the previous 5 minutes?
  Multiple selection ("Reading", "Typing", "Speaking", "E-mailing", "Walking around", "Exercising", "I do not remember" and "Other")

- Were you in a social situation during the previous 5 minutes?
  Multiple selection ("Yes, in person", "Yes, remotely", "No, I do not remember" and "Other")

- If you were in a social situation, which set of the following best describe your relationship with that person?
  Multiple selection ("Boss", "Friend", "Colleague", "Not applicable" and "Other")

- Were you listening to music during the previous 5 minutes?
  Selection ("Yes", "No" and "I do not remember")

- What body postures were you mostly having during the previous 5 minutes?
  Multiple selection ("Sitting", "Standing", "Lying", "I do not remember" and "Other")

At the end of each day, participants had to respond to the following questions:

- Were there significantly stressful events that were not reported during the day? If yes, please provide the approximate times (for instance, 10:00 - I had a public presentation, 15:30 - I had problems with my computer).
  Text

- Did you drink caffeine during the day? If yes, please provide the approximate times (e.g., 10:00, 15:00, 17:00).
  Text

- Did you experience any problems with the sensors? (e.g., I forgot to charge them, the battery died, they never asked me any question, I had to remove the sensors).
  Text
Overall, how pleasantly were you feeling today?
* ("Low unpleasant" and "Very pleasant")

Overall, how energetic were you feeling today?
* ("Low energy" and "Very energetic")

Overall, how would you rate the amount of demands of today?
* ("Very low" and "Very high")

Overall, how would you rate the amount of resources of today?
* ("Very low" and "Very high")

Overall, how stressed were your feeling today?
* ("Not at all" and "Extremely")

During the day of today, how often have you felt that you were unable to control the important things in your life?
* ("Never" and "Very often")

During the day of today, how often have you felt confident about your ability to handle your personal problems?
* ("Never" and "Very often")

During the day of today, how often have you felt that things were going your way?
* ("Never" and "Very often")

During the day of today, how often have you felt difficulties were piling up so high that you could not overcome them?
* ("Never" and "Very often")

Qualitative Questions about Devices

At the end of the study, participants had to respond to the following questions about the head-mounted and wrist-worn wearable devices as well as the smartphone:

- How often would you wear the device during daily life?
  * ("Never" and "All the time")

- Did wearing the device affect your social interactions?
  * ("Never" and "All the time")

- How would you rate the quality of your reports with this device?

200
At the end of the study, participants had to respond to the following questions about the wearable camera:

- How often would you wear the device during daily life?
  * ("Never" and "All the time")
- Did wearing the device affect your social interactions?
  * ("Never" and "All the time")
- Did wearing the device increase your stress levels?
  * ("Not at all" and "Significantly more")
- How did you find wearing the device during your workday?
  * ("Very uncomfortable" and "Very comfortable")
- How concerned were you feeling about the device capturing private information/moments?
  * ("Not at all" and "Extremely")
- Did the first-person images help you review your daily activities at the end of the date?
  * ("Not at all" and "Extremely")
- Did you learn something unexpected from the images? If yes, please briefly describe what.
  
Text
Questions about the Prompts

At the end of the study, participants had to respond to the following questions about the prompts:

- How difficult did you find rating your level of energy during your work day?
  * (“Very challenging” and “Very easy”)
- How difficult did you find rating your level of pleasantness during your work day?
  * (“Very challenging” and “Very easy”)
- How difficult did you find rating your level of demands during your work day?
  * (“Very challenging” and “Very easy”)
- How difficult did you find rating your level of resources during your work day?
  * (“Very challenging” and “Very easy”)
- How difficult did you find rating your level of stress during your work day?
  * (“Very challenging” and “Very easy”)
- How difficult did you find rating the level of disruptiveness of the prompts?
  * (“Very challenging” and “Very easy”)
- How did you feel about the number of prompts during the study?
  * (“Too few” and “Too many”)

Other Questions about Participants

- Do you have any cardiac condition for which you are being treated?
  Selection (“Yes” and “No”)
- Do you have any respiratory condition for which you are being treated?
  Selection (“Yes” and “No”)
- Do you have any musculoskeletal condition for which you are being treated?
  Selection (“Yes” and “No”)
- Gender
  Selection (“Male” and “Female”)
- When were you born?
  Date
• Race/Ethnicity
  Selection ("American Indian or Alaska Native", "Hawaiian or Other Pacific Islander", "Asian or Asian American", "Black or African American", "Hispanic or Latino", "Non-Hispanic White" and "Other")
• Height
  Number
• Weight
  Number
• Handedness
• Selection ("Left-handed", "Right-handed" and "Ambidextrous")
• Do you usually wear glasses?
  * ("Not at all" and "Most of the time")
• Do you usually wear eye contacts?
  * ("Not at all" and "Most of the time")
• Do you usually wear a watch?
  * ("Not at all" and "Most of the time")
• Where do you usually carry your phone?
  Multiple selection ("Front pocket of trousers/skirt", "Back pocket of trousers/skirt", "Front pocket shirt, Bag/Purse" and "Other")
• Please briefly describe the type of work you do and your work environment (few sentences) (e.g., I write code and I am mostly sit in my office, I work in a public space with 3 more people, I do not usually use the computer, I usually go from meeting to meeting)
  Text
C. Disclosure Signage

Participants of the study were instructed to use the signage of Figure 53 to inform other office workers about the cameras and the study.

Anonymized images inside this office may be captured for research purposes

In case of questions please contact:
Javier Hernandez (javierhr@mit.edu)

Figure 53: Disclosure signage used to inform other people about the cameras and the study
D. Self-reported Regression Coefficients

Affective and work-related self-reports were used to predict self-reported stress levels with a linear regression with quadratic components.

The linear regression model is as follows:

\[ S \sim 1 + A \times V + A \times D + A \times R + V \times D + V \times R + D \times R + A^2 + V^2 + R^2 + D^2 \]

where \( S \) = Stress, \( A \) = Arousal, \( V \) = Valence, \( D \) = Demands, and \( R \) = Resources. The estimated coefficients are shown on Table 21.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>tStat</th>
<th>pValue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.295748</td>
<td>0.478639</td>
<td>4.796404</td>
<td>2.13E-06</td>
</tr>
<tr>
<td>Arousal</td>
<td>0.003806</td>
<td>0.012596</td>
<td>0.30213</td>
<td>0.762678</td>
</tr>
<tr>
<td>Valence</td>
<td>-0.01324</td>
<td>0.012174</td>
<td>-1.08728</td>
<td>0.277435</td>
</tr>
<tr>
<td>Demands</td>
<td>-0.0121</td>
<td>0.012107</td>
<td>-0.99943</td>
<td>0.31807</td>
</tr>
<tr>
<td>Resources</td>
<td>0.038001</td>
<td>0.013052</td>
<td>2.911464</td>
<td>0.003758</td>
</tr>
<tr>
<td>Arousal x Valence</td>
<td>-0.00026</td>
<td>0.000123</td>
<td>-2.11766</td>
<td>0.034696</td>
</tr>
<tr>
<td>Arousal x Resources</td>
<td>2.99E-07</td>
<td>0.000115</td>
<td>0.002599</td>
<td>0.997928</td>
</tr>
<tr>
<td>Valence x Demands</td>
<td>0.000113</td>
<td>0.000122</td>
<td>0.923798</td>
<td>0.356036</td>
</tr>
<tr>
<td>Valence x Resources</td>
<td>7.28E-05</td>
<td>0.000122</td>
<td>0.598492</td>
<td>0.549782</td>
</tr>
<tr>
<td>Demands x Resources</td>
<td>6.29E-05</td>
<td>0.000145</td>
<td>0.433529</td>
<td>0.664817</td>
</tr>
<tr>
<td>Arousal^2</td>
<td>-0.00028</td>
<td>0.000109</td>
<td>-2.57036</td>
<td>0.010447</td>
</tr>
<tr>
<td>Valence^2</td>
<td>7.07E-05</td>
<td>0.000114</td>
<td>0.621481</td>
<td>0.534566</td>
</tr>
<tr>
<td>Demands^2</td>
<td>-2.93E-05</td>
<td>0.0001</td>
<td>-0.29224</td>
<td>0.770221</td>
</tr>
<tr>
<td>Resources^2</td>
<td>0.000416</td>
<td>9.95E-05</td>
<td>4.177311</td>
<td>3.48E-05</td>
</tr>
</tbody>
</table>
References


