2020

Detecting Stressful Social Interactions Using Wearable Physiological and Inertial Sensors

Rummana Bari

TOWARDS AUTOMATED IDENTIFICATION OF SOURCES OF STRESS FROM WEARABLE PHYSIOLOGICAL AND INERTIAL SENSORS

by

Rummana Bari

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy (Ph.D.)

Major: Department of Electrical and Computer Engineering

The University of Memphis
August 2020

ACKNOWLEDGMENTS

I would like to express my special gratitude and thanks to my adviser Dr. Santosh Kumar who has been a great mentor through my PhD journey. He has motivated me to do critical thinking regarding research work, inspire me to improve writing skill, encouraged me to do collaboration with other researchers. I have learned how to maintain composure under work stress, respect others opinion and eagerness to learn from anybody and lots of other things that help to excel my future professional life.

I also want to mention, with gratitude, Dr. Eugene Buder, who worked closely with me towards this achievement. This whole dissertation came about with the novel couple study which would not have been possible without the help of Dr. Buder. He and his student Megan Battles helped in designing study protocol, collecting data in his lab, and annotating data for model development. I would also like to thank Dr. Madhusudhanan Balasubramanian and Dr. Peter Lau for agreeing to serve on my dissertation committee and providing valuable feedback. I would also like to thank fellow graduate students in MD2K lab who encouraged me to strive towards my goal. I would also like to thank Cheryl D Hayes and Lyndsey Rush for handling all the administrative work.

A special thanks to the lab partner and life partner, Md. Mahbubur Rahman, for being an integral part of my professional and social life. I am grateful to my daughter Mashiyat for becoming a source of energy in my life. I also cannot express how grateful I am to my mother, father, brother and sisters for all of the sacrifices that they have made for me.

Abstract

Stress is unavoidable in everyday life. Continuous, and repetitive stress can result in several health related short and long-term adverse consequences. Previous research found that most of the stress events occur due to interpersonal tension followed by work related stress. Enabling automated detection of stressful social interactions using wearable technology will help trigger just-in-time interventions which can help user cope with the stressful situation. In this dissertation, we show the feasibility of differentiating stressful social interactions from other stressors i.e., work and commute. However, collecting reliable ground truth stressor data in the natural environment is challenging. This dissertation addresses this challenge by designing a Day Reconstruction Method (DRM) based contextual stress visualization that highlights the continuous stress inferences from a wearable sensor with surrounding activities such as conversation, physical activity, and location on a timeline diagram. This dissertation proposed a respiration based conversation model to locate the interactions to support the visualization. Advantage of respiration signal is that it does not capture the content of conversation and hence, is more privacy preserving compared to audio. However, it requires wearing of chest worn sensor. This dissertation aims to determine stressful social interaction without wearing chest worn sensor or without requiring any conversation model which is privacy sensitive. Therefore, it focuses on detecting stressful social interactions directly from stress time-series only which can be captured using increasingly available wrist worn sensor. This dissertation propose a framework to systematically analyze the respiration data collected in natural environment. The analysis includes screening the low quality data, segmenting the respiration time-series by cycles and develop time-domain features. It proposes a Conditional Random Field, Context-Free Grammar (CRF-CFG) model to detect conversation from breathing patterns. This system is validated against audio ground-truth in the field with an accuracy of 71.7%. To detect stressful social interactions from stress time-series data, this dissertation introduces stress cycle concept to capture the cyclical patterns and identifies novel features from it. Furthermore, wrist worn accelerometry data in this study shows that hand gestures have a distinct pattern during stressful social interactions. The model presented in this dissertation augments accelerometry patterns with the stress cycle patterns for more accurate detection. Finally, the model is trained and validated using data collected from 38 participants in free-living conditions. The model can detect the stressful interactions with an F1-score of 0.83 using features obtained from just one stress cycle and enable the delivery of stress intervention within 3.9 minutes since the onset of a stressful social interaction.

TABLE OF CONTENTS

Content **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

PUBLICATIONS (Total Citations: 173)

1. **R. Bari**, M. Rahman, M. B. Parsons, E. H. Buder, and S. Kumar, "*Automated Detection of Stressful Conversations Using Wearable Physiological and Inertial Sensors*," under review in ACM UbiComp, 2020. (24 pages).

2. **R. Bari**, R.J. Adams, M. Rahman, M. B. Parsons, E. H. Buder, and S. Kumar, "*rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor*," ACM UbiComp, 2018. (26 pages) (Acceptance rate 22%).

3. M. Rahman, **R. Bari**, A. Ali, M. Sharmin, A. Raij, K. Hovsepian, S. Hossain, E. Ertin, A. Kennedy, D. Epstein, K. Preston, M. Jobes, S. Kedia, K. Ward, M. al'Absi, and S. Kumar, "*Are We There Yet? Feasibility of ContinuousStress Assessment via Wireless Physiological Sensors*," ACM BCB, 2014. (10 pages) (Acceptance rate 22%).

4. H. Sarker, M. Sharmin, A. Ali, M. Rahman, **R. Bari**, M. Hossain, and S. Kumar, "*Assessing the Availability of Users to Engage in Just-in-Time Intervention in the Natural Environment*," In Proceedings of ACM UbiComp, Seattle, WA, 2014. (12 pages) (Acceptance rate 25%).

5. A. Kennedy, D. Epstein, M. Jobes, K. Phillips, D. Agage, M. Tyburski, A. Ali, **R. Bari**, S. Hossain, K. Hovsepian, M. Rahman, E. Ertin, S. Kumar, and K. Preston, "*Continuous In-The-Field Measurement of Heart Rate: Correlates of Drug Use, Craving, Stress, and Mood in Polydrug Users*," In International Journal of Drug and Alcohol Dependence, 2015. (25 pages) (h-Index 110).

Chapter 1

Introduction

1.1 Motivation

Stress is unavoidable in our everyday life. There are numerous reasons for stress, such as difficulties in interpersonal relationships, long-standing pressures at work, an unsatisfying career, deadlines, test-taking, financial difficulties, health issues, care-giving, etc. The Yerkes—Dodson law of empirical relationship between arousal and performance states that humans perform at an optimal level under a certain amount of stress [1]. But continuous, repetitive, and excessive stress can result in emotional distress, headaches, back pain, elevated blood pressure, trouble sleeping, slower body recovery, and decreased mental performance, among several other short and long-term adverse consequences [2]. Therefore, it is important to better manage stress in daily life for better physical and mental health and well-being, relationship satisfaction, work performance, and an overall better quality of life.

Prior work has investigated and organized different types of stressors. For example, 1,031 participants were studied in [3]. They observed 4,000 stressful events from the daily life of these participants and organized the stressors in seven broad categories — interpersonal argument and tensions, work, home related stress, finances, health, networking, and miscellaneous. Among them, interpersonal argument and tensions occur most frequently (50% of the time) as people interact with partners, friends, family members, colleagues, and supervisors regularly. This is followed by work related stress (13.4% of the time) including work demand, overload, technical issues, and job security. Another work studied 225 graduate students and found that academic or professional demands, interpersonal demands, financial strains, and commuting were found to be the most common stressors [4].

As interactions with partner, family, friends, colleagues are a fundamental aspect of our daily life, stressful interaction is a major daily stressor for a large population.

Healthy interactions can provide happiness, social support, and cause fewer health issues [5, 6]. But, stressful interactions such as conflicts may lead to deleterious consequences to physical and psychological health (e.g., depression, anxiety, and substance abuse) and may affect the relationship quality, happiness, and overall life satisfaction [7, 8, 9, 10]. Moreover, stressful conversations at work can adversely impact productivity, job performance, and job satisfaction [11].

Therefore, it is important to understand the timing, frequency, and duration of stressful conversations to reduce their harmful effect in daily life. Sensor-based automated detection of stressful conversations from the natural environment can be used by researchers to investigate the antecedents, dynamics, and consequents of stressful conversations, potentially leading to novel therapies and interventions. Moreover, real-time detection of such conversations can be used to trigger just-in-time mobile interventions for deescalating a tense situation and for pacifying the users so that they can recover and cope better with the situation.

For decades, extensive research has been conducted on developing and implementing mindful stress management methods, such as deep breathing, yoga, meditation, biofeedback, guided visualizations [12], voice feedback to slow breathing pace, and guided body scans. They assist users in managing and manipulating their stress arousal. Initially, these mindful interventions were delivered face-to-face by coaches (including virtually), then delivered remotely over telephone, transitioning to delivery via text messages, to now being delivered on smart phone and smart watches. However, most of these mindful intervention methods requires active attention from the users and may interrupt the ongoing tasks. Recent work has demonstrated feasibility of mindless interventions which does not requires user's active attention. Researchers showed that it is possible to regulate user's emotions by providing false feedback of a slow heart rate via smart watch, or by using a voice modulation intervention that can change the emotional tone of users' own voices during test taking or while involved in an interpersonal conflict [13, 14, 15]. Other researchers have designed haptic interventions on car seats, helping users do deep breathing exercises while driving [16]. Therefore, users can still perform their tasks, and the technologies act in parallel in an unobtrusive way without interrupting them. Thus, the effectiveness of an intervention during stressful social interactions depends not only on the timing of interventions, but also the right choice of intervention. If we can able to detect a stressful interaction, a proper intervention can be delivered to minimize the intensity of the situation. For example, to deescalate an ongoing conflict, his/her own modulated voice can be delivered via an ear bud or can be provided with a haptic feedback using a wrist watch, which should be more appropriate and effective than suggesting yoga at that moment. Starting a yoga session during an on-going inter-personal interaction may interrupt the flow of that interaction. The critical missing component of just in interventions is finding the moment of stressful social interactions.

Fortunately, both conversational interactions [17, 18] and physiological response to stress [19] can be detected from wearable respiration sensors data. Combining these two inferences can potentially indicate the timing of stressful interactions or conversations. However, this method suffers from several challenges. First challenge is that it requires wearing of chest worn sensors to collect reliable respiration data. Second, it is unknown how to combine these two modalities. Moreover, social interactions or conversations can also be detected from audio signal which involves privacy concern in real life. Our goal is to determine stressful conversations without wearing chest worn sensor or without requiring any kind of conversation model either from respiration or audio. That motivates us to explore stressful interactions directly from stress time-series.

In this dissertation, we demonstrate the feasibility of detecting stressful social interactions or conversations from stress time-series data. In particular, we show that by analyzing the dynamics of stress time series, we can detect whether the current stress event is due to stressful conversations or other stressors such as commuting or work related stress. Automatic detection of stressful interactions or conversations from mobile sensors involves several challenges.

1.2 Challenges and Contributions

In this section, we present four technical challenges that arise in detecting stressful social interactions and our contributions to address each of them.

1. Challenges in obtaining ground truth labels to model stressful interactions.

The foremost challenge in developing stressful social interaction model from stress time-series is to get fine-grained labeling of the stress events i.e., timing and duration of the events. The traditional approach is to request users to proactively provide labels by manually keeping a dairy [20], retrospectively via an interview [3], or ecological momentary self-reports [4, 19]. However, these methods lack the temporal resolution and reliability needed to develop a sensor-based model successfully [21]. Alternatively, an observer can be assigned to follow each participant in their daily life. However, it involves significant expense, burden, and may still not capture several real-life scenarios in order to respect participants' privacy.

To collect ecologically valid data from the daily life of participants with unambiguous and temporally-precise labels, we designed and conducted a lab and a field study. Stressful conversations usually involve two (or more) parties, both of whose consent is generally needed, especially for capturing sensor data during stressful conversations and other real-life stressors. As cohabiting couples typically spend a lot of time together, we recruited couples to wear sensors and collect data concurrently. To detect the timing of stress events, we used a previously validated model to passively infer stress arousal from Electrocardiogram (ECG) and respiration signals that produce stress likelihood for every minute [19]. To find the start and end times of stress events from the continuous stress time-series data, we use the model presented in [22]. To overcome the imperfection of machine learning models for stress detection, we developed an automated stress visualization system utilizing Day Reconstruction Method (DRM) [23] concept to present the detected stress events to users with surrounding contexts (i.e., conversation, location, and physical activity), all derived from sensor data. The visualization helped users recall stressful events so they could confirm or refute a detected stress event and remember the reason for the confirmed stress events, providing us labels of stressors for each identified and confirmed stress event. As automated detection of conversations from audio or respiration data is limited to an F1 score of around 0.7 [18], we collected high-quality raw audio to verify the presence of conversations via human annotation. Finally, as collection of raw audio poses privacy concern and burden because the participants needed consent from anyone they talk to, we limited the data collection with each couple to one full day, similar to other studies that also recruited couples and collected wearable sensor and audio data from them [24, 25]. As one day of data consists of between 2 and 3 detected stress events [26], we get sufficient data from each couple for our modeling. To increase between-person and between-situation diversity in the data, we recruited 38 participants (19 cohabiting couples) in the field. Upon completing data collection, the participants were shown their own stress arousal data with other contextual information. They were asked to first verify the occurrence of detected stress events and then recall the reason of stress for each stress event detected correctly. After observing the visualization, they were able to recall 97 stressful events. Participants recalled several reasons for stress events such as meeting with a supervisor, having deadlines at work, job interviews, conflict with their partner, driving on a busy road, assignment deadlines, etc. We found majority of the stress events are due to interpersonal interactions.

To understand the nature of physiological response during stressful conversations, we conducted a lab study with 12 participants (6 cohabiting couples) that was structured to trigger stressful conversations among couples. The lab study ensures control of other potentially confounding events in the field that may affect physiology (i.e., physical activity), allowing us to discover the unique patterns of stress response in sensor data during stressful conversations.

2. Challenges in extracting features from respiration signal in field.

To enable stressful interaction or conversation detection model, we need dense and continuous stress time-series data in field. Also, to label stressful events, we provide conversation inference as a cue in the visualization system. Respiration signal can be used to infer psycho-physiological stress [19]. Another benefit of respiration sensing is that breathing kinematics can provide useful information about a person's speaking status. Conversation causes specific changes in breathing patterns in addition to generating sounds. Therefore, we choose to use respiration signal over audio to measure stress and conversation because this signal is less privacy sensitive.

We used a physiological sensor suite to collect respiration data continuously and passively in natural environment. Respiration data has traditionally been collected in controlled settings such as sleep labs and speech labs. But, the natural environment introduces numerous challenges to the screening, cleaning, and processing of this data. There are several challenges that prevent achieving good accuracy for detecting human states and behaviors at the cycle-level of granularity in respiration data collected in the field environment.

A first challenge is the accurate identification of breathing cycles, i.e., pinpoint several interesting points of a cycle such as onsets of inspiration and expiration that demarcate change in phases of breathing and are critical to accurate computation of features along both time and amplitude dimensions. Cycle identification is challenging due to voluntary control of breathing, the baseline shift in the respiratory data, daily activities, short breaths, end expiratory pauses or breath holds, and others. Second challenge is to handle the effect of activity and postures. Respiration signal can be easily influenced by movements of limbs and torso, changes in posture (i.e., sitting, supine), and physical activity (walking). To handle these challenges, we present a rigorous method for screening, cleaning respiration signals and developed moving average based algorithm for identifying respiration cycles captured in both lab and field settings. Among 1,934 respiration cycles collected in lab in presence of conversation, the proposed cycle identification method can identify 94.4% cycles correctly. Among 1,500 cycles collected in natural environments, the proposed method identified 96.34% cycles correctly in the presence of physical activities (walking) and in different postures (e.g., sitting and standing).

3. Challenges in developing conversation model from respiration signal. Developing respiration based conversation model involves several challenges. A first challenge is to get fine-grained labels for each cycle (speech and non-speech) which are necessary to train and validate a classifier. Most existing approaches for labeling data are inadequate for our study: a) requesting self-reports from the users is impractical, i.e., users cannot label each breath cycle when they are engaged in a natural conversation, b) having an observer annotate each cycle. Further, turn taking can occur swiftly, making it impossible to keep track of and synchronize the labels to the sensor data. A second challenge is segmenting the respiration signal into periods of conversation, which consists of both speech and non-speech cycles. For example, silence during a conversation may be due to all parties engaged in thinking or may mark the start of a new conversation episode. A third challenge is to generalize the conversation model built using controlled lab data to naturally occurring conversations in uncontrolled field environments, which may have different distributions of speech/non-speech durations. The final challenge is to validate the model in the field against a widely-used gold standard.

We present a Conditional Random Field, Context-Free Grammar (CRF-CFG) based conversation model to classify respiration cycles into speech or non-speech, and subsequently infer conversation episodes. Our model achieves 82.7% accuracy for speech/non-speech classification and it identifies conversation episodes with 95.9% accuracy on lab data using a leave-one-subject-out cross-validation. Finally, the system is validated against audio ground-truth in a field study with 38 participants. This model identifies conversation episodes with 71.7% accuracy on 254 hours of field data.These are comparable with conversation detection from high-quality audio recordings from the LENA device [27].

4. Challenges in developing stressful conversation model from stress time-series.

Next challenge is to to identify signatures that can distinguish stressful interactions or conversations within a stress time series. To the best of our knowledge, there exists no model to detect stressful interaction using physiological or inertial sensors. For better generalizability of the model, our goal is to discover features from stress time-series only. So that it can work irrespective of how stress is detected (e.g., from electrocardiogram, optical sensing on wrists, or sensing of electrodermal activity) and can be easily and widely deployable in tohe field.

In the lab data, we observe that the stress time-series follows a cyclical pattern that results from the interplay between the sympathetic and parasympathetic nervous system during a stress response, similar to that found in physiology during stress [28, 29]. To develop this model, we first develop a method to automatically identify this cyclical pattern or cycles in the stress time-series data. We use the cycle as a dynamic, natural window to segment the stress time series during a stress event. We then identify discriminative features from each stress cycle and train a machine learning model to determine whether a stress event is due to stressful conversations.

We show that using features from one stress cycle, the model can identify whether a stress is due to stressful interactions or conversation with an F1 score of 0.74. We also observe distinct patterns in hand gestures during stressful conversations. By augmenting the model with hand gesture features (derived from wrist-worn inertial sensors) within each stress cycle, the F1 score improves to 0.83. A stressful conversation usually consists of multiple stress cycles. Using all cycles improves the F1 score to 0.89, providing a trade-off between accuracy and how early since the start of a stressful conversation, an intervention can be delivered.

1.3 Dissertation Outline

In this dissertation, we develop machine learning methods and models to address the above challenges described above.

In Chapter 2, we review existing works for physiological response to different stressors, how physiology acts differently for different stressors in lab settings and existing conflict detection model in field.

In Chapter 3, we describe the procedure of data collection in lab and field settings. The lab data is mainly used to develop to respiration cycle identification algorithm and conversation model from the sequence of respiration cycles. Next we describe field study that is needed to model stressful interactions or conversations from stress time-series data.

In Chapter 4, we describe the algorithm to detect respiration cycles. We propose several metrics to evaluate the performance of the algorithm. We finally compare the performance with two existing models.

In Chapter 5, we present a CRF-CFG base conversation detection model developed using lab data. We then implement a lab to field generability model to improve the performance of this model in field. Finally, we compare the result with audio based conversation model.

In Chapter 6, we propose a stressful social interaction model using stress time-series data. We extract several novel features from the cyclical pattern of stress time-series. Later, we augment the performance of the model using wrist motion features. Finally, we describe the implication of this model for just-in-time intervention.

Chapter 2

Literature Review

We review the existing works for physiological responses to different sources of stress, detection of stress using mobile and wearable devices, and detection of social interactions and conflicts from physiological and audio signal.

2.1 Background on Physiological Response to Stressors

A stressor presents a challenge, opportunity, or threat to users. To help users prepare for stress response, their autonomous nervous system (ANS) activates their physiology that includes the cardio-respiratory system (i.e., heart and lungs), endocrine system (e.g., hormone secretion), and the thermoregulatory system (e.g., temperature and sweating). ANS comprises of the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS) [30]. The SNS elevates the physiology, preparing the body for a 'fight-or-flight response. To provide the needed energy, SNS stimulates several physiological parameters (e.g., heart rate, respiration rate, blood volume, body temperature, etc.). To limit any damage to the end organs, PNS acts as a counterbalance mechanism to restore calm and thus maintain homeostasis. Its strength is usually proportional to the increase caused by SNS, and it eventually brings the physiology back to a resting state.

The interplay of SNS and PNS can be illustrated by considering the impact on the cardiovascular system. In response to a stressor, the SNS increases the heart rate (HR). Once the threat is over, the PNS reduces HR, bringing it back to a resting state [31]. Heart rate variability (HRV) is a common measure to quantify the interaction of SNS and PNS. The HRV is defined as the variation in the beat-to-beat intervals. An increased/decreased HRV indicates increased activity of the PNS/SNS, respectively. Therefore, HRV is a simple measure to quantify the contributions of the PNS/SNS and has traditionally been used to estimate stress response. Heart rate variabilities (HRV) have been found to follow cyclical patterns in lab settings [28, 29]. De Geus, et al., showed that the heart rate increases when users face stressors [32]. For stressors, they used a tone avoidance task, a memory search task, and a cold pressor test. They found that the heart rate remains high as long as the stressor is present and goes back to the pre-stress level with the removal of stressors, resulting in a cyclical pattern.

The stress response can also be explained in terms of endocrine response to stress, i.e., salivary cortisol levels. In [33], authors investigated the cortisol level in 124 heterosexual dating couples during a conflict negotiation task. The cortisol was assessed at 7 points before and after the task, creating a trajectory of stress reactivity and recovery for each participant, resulting in a cyclical pattern.

The interplay of SNS and PNS can be distinct when presented with different stressors as the persistence of stress stimuli can differ. For example, during a cold pressor test, the initial stress response can be high due to shock from cold temperature, but physiology can gradually recover as the body gets used to the temperature difference. But, in a stressful conversation, there can be highly stressful moments, that may be followed by either further escalation or de-escalation, which can drive the activation of SNS and PNS differently than from a cold pressor test. In fact [34] showed that the stress responses to three different stressors (i.e., cognitive, emotional, and physical) are sufficiently distinct that they can be detected using a machine learning model. In another recent work, [35] showed that respiration pattern during stressful conversation is different than that during a stressor not involving conversation (i.e., cognitive). Both of these works used controlled lab experiments to show the distinction in stress response due to different stressors. We build upon these works to observe the physiological responses to real-life stressors occurring in daily life (in a field study) and develop a model that can successfully identify when a stress response is due to stressful conversations.

2.2 Stress Monitoring Using Wearable Sensors

There has been extensive work in detecting stress, first in the lab settings using ECG and respiration [36], gradually moving to ambulatory field environment (carrying

Holter Monitors in backpacks) [37], then to selected tasks in the field environment with wearable sensors [38], and finally to free-living environment with unobtrusively wearable wireless sensors [19, 39]. Recent works present stress detection from pulse plethysmograph (PPG) in conveniently-worn wrist devices [40, 41]. The focus of the machine learning models in these works was to develop a single model that can detect stress irrespective of the type of stressor. They mostly use diverse stressors in the lab settings, e.g., using a cold pressor as a physical stressor, mental arithmetic as a cognitive stressor, and public speaking as a social stressor. But, the goal for modeling has been to extract commonality in stress response captured by sensor data so that a single trained model can detect all stress events. Our goal here instead is to discover uniqueness in the stress responses due to different stressors.

There have been limited works in developing models to distinguish among different stressors. It was shown in [34] that stress responses during different stressors show discernible differences. Using lab data, they developed a Gaussian mixture model to cluster the physiological signals (consisting of heart rate, electrodermal response, and oxygen saturation) captured during cognitive stress (counting backward by sevens beginning from 2,485 and Stroop test), emotional stress (watching horror movie for 5 minutes), exercise (walking on a trade-mill for two minutes), and a resting state. They report an accuracy of 84% for the four class classifier, demonstrating the feasibility of developing models to distinguish among different stressors and rest state. As this work was limited to lab stressors, it did not show how well these patterns can distinguish among real-life stressors in field settings.

A recent work [35] collected respiration data in the lab settings, where they included a non-verbal relaxer (watched 10 minutes neutral movie), a verbal relaxer (talked in mother language for 5 minutes on a chosen topic), a verbal stressor (prepared and participated in an interview), and a non-verbal stressor (took part in a cognitive task). In order to improve the accuracy of stress detection, they developed a two-stage model. In the first stage, they detect whether a conversation is taking place, and depending on the outcome, they apply different stress models to detect whether the signals exhibit a stress response. They showed that using a two-stage classifier, they achieve 83% accuracy compared to 76% when using a one-layer classifier that does not detect conversations, demonstrating that stress response in respiration is different during stress events with or without a conversation. As their goal was to improve the detection of stress model similar to other works in stress detection, they did not address the issue of distinguishing verbal stressors from non-verbal stressors on their lab dataset. In Section 6.8, we construct a baseline model motivated by this work that uses an automated detection of conversation and automated detection of stress and combines both to detect stressful conversations. We find the best performance from such a model is limited to an F1 score of 0.6.

2.3 Detection of Conflicts using Audio and Physiological Data

Finally, [24] showed the feasibility of detecting whether an interpersonal conflict occurred in each hour (reporting accuracy of 69.2%) using wearable sensor and audio data for that hour from romantic couples who wore sensors for a day in field. As the focus of this work was to detect for each hour whether any conflict occurred or not, they did not present any model to distinguish among different stressors or find different sources of stress.

Our work builds upon, contributes to, and complements the above works by presenting new methods to identify distinguishing patterns in stress dynamics of an individual in daily life using mobile sensors, and demonstrating that it is feasible to detect stressful interactions from other daily stressors. In summary, to the best of our knowledge, our work is the first attempt at demonstrating that stressful conversations can be detected automatically from wearable physiological sensors in daily life, without the need for audio data.

Chapter 3

Study Design and Data Collection

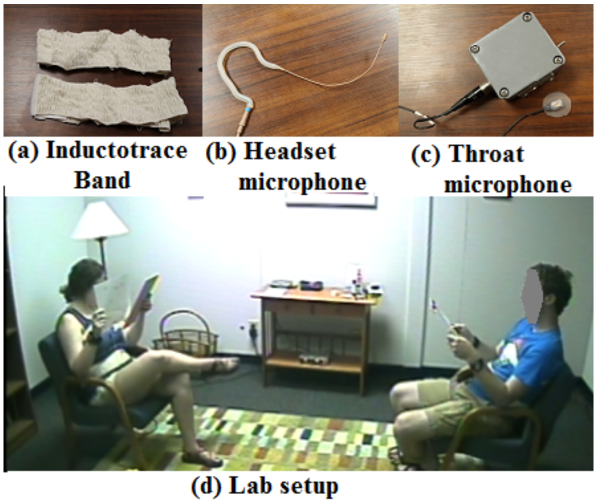
For development, training, and testing of stressful interaction or conversation model, we collected data in both lab and field settings. The project recruited couples living together to maximize the occurrence of interpersonal interactions, including stressful ones. The lab study was designed to capture elicited and fully observable interpersonal conflicts, whereas the field study captured conflicts naturally occurring in daily life. All studies were approved by the Institutional Review Board (IRB) at University of Memphis, and all participants provided written informed consent. We now describe details of both studies.

3.1 Lab Data Collection to Model Conversational Interaction

We designed the lab study to serve two purposes- (1) to develop a model to detect conversation from respiration data and (2) to collect clean confounder-free data (e.g., due to physical activity) during stressful conversations that can be used to find any distinguishing pattern in the stress time-series signal. The lab tasks were designed to create difficult communication situations and thus induce interpersonal conflicts. Conversational interaction data is collected in two settings - (1) in a true laboratory setting and (2) in natural environment. Interaction data collected in true laboratory setting is designed to collect conversation data in sitting position and to validate the performance of the chest band sensor with a hospital grade system. Later on, chest band sensor is used to collect data in field. Participants engaged in several vocal and conversational tasks in the University of Memphis Social Interaction Laboratory. Moreover, interaction data is collected in natural environment to enhance the generalizability of the model to detect conversation in presence of free-living activity since activity also affects respiration measurements.

3.1.1 Participants

In true laboratory setting, data is collected from 12 individuals (6 pairs of cohabiting couples) from students, full-time professionals and part-time employees at a


(a) Inductotrace Band (b) Headset microphone (c) Throat microphone

(d) Lab setup

Fig. 3.1: Lab equipment and lab setup.

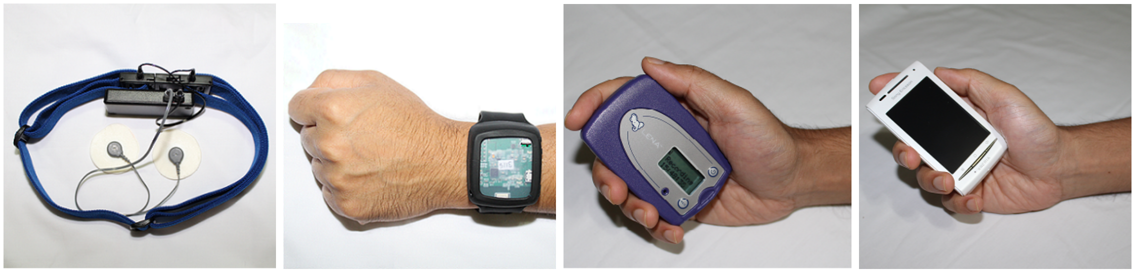


Fig. 3.2: (a) Chest band sensor (captures respiration, ECG, and accelerometer signal). (b) Samrt watch. (c) LENA audio recorder, and (d) Study phone (Sony Ericsson Xperia X10, Android Smart phone).

university. Participants included 7 women (mean age: 29.9 ± 7.4 years) and 5 men (mean age: 27.2 ± 2.9 years).

To acquire conversation data in presence of activities, labeled quiet breathing and speech breathing data were collected in presence of physical activity (i.e., walking) from 5 healthy adults (mean age: 30.9 ± 1.3 years) in natural environment, .

3.1.2 Devices

In the lab, respiratory activity was measured with two types of Respiratory Inductance Plethysmography (RIP) bands. The first one is a hospital grade Inductotrace band which quantifies changes in the rib cage and abdomen cross-sectional areas by means of two elastic transducer belts placed at the level of the armpits and the navel (see Figure 3.1a). Inductotrace bands were connected to a calibration unit (Inductotrace system, Ambulatory Monitoring Inc.) via a transducer oscillator. A Data Translation DT381 analog-to-digital (A-D) converter operated by TF32 software was used to convert this signal into digital form on a computer.

The Inductotrace system, however, is not suitable for collecting data in the field as it is bulky, requires a fixed setup, and is not wireless. To monitor respiratory behavior in the field, we use the AutoSense chest suite of sensors [42] that collects Electrocardiogram (ECG), respiration and 3-axis accelerometer signals (Figure 3.2a). In this experiment, we are able to compare the performance of the field instruments to well calibrated hospital-grade respiratory monitoring equipment to provide ground truth data and improve the potential of field sensors for modeling conversational behaviors in the field.

A headset microphone as shown in Figure 3.1b was placed in front of the participant's mouth and processed through an analog amplifier. Participants also wore a throat microphone (see Figure 3.1c), which captures the vibration of the throat that occurs during speaking and helps to isolate very low level speech that might otherwise be overlaid by airborne cross talk (PentaxMedical model 7184-9700). In this setting, we obtained video with both face and side views of the conversational partners. Figure 3.1d shows the whole lab setup where conversation partners were seated face-to-face, as captured using the side view video camera.

In natural setting, participants wore the chest band sensors (Figure 3.2a) and a wristband on their dominant wrist (Figure 3.2b). They wore a LENA audio recorder [27] to detect conversation events (as shown in Figure 3.2c). They were instructed to carry the recorder in a pouch placed around the abdomen to reduce occlusion of microphone and other audio artifacts. Each participant was provided with an Android smartphone shown in Figure 3.2d that receives and stores data from all wearable sensors.

3.1.3 Lab Study Protocol

The lab study tasks were designed to capture both regular and difficult communications and therefore possible interpersonal conflicts. Each couple took part in

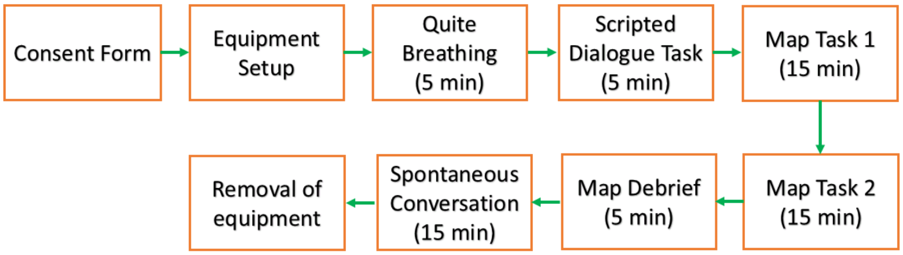


Fig. 3.3: The sequence of tasks in true laboratory setting.

several interaction tasks in a sitting position with limited or no movement. So the variation in stress likelihood probably corresponds to physiological arousal. To capture baseline measures, participants remained seated face-to-face in a comfortable chair silently for five minutes. Next, they were asked to read an interactive script that was created using previously recorded spontaneous conversation as a 'Scripted Dialogue' task. This lasted for approximately five minutes. The third phase of lab recording then utilized a task that involved recreating a map [43] which elicits goal-oriented conversation. Both participants were given maps that had been used in prior literature, one presenting a pre-printed route with a starting and finishing point for the Instruction Giver and the other presenting a map with only a starting point for the Instruction Receiver. The Instruction Follower attempted to recreate the Instruction Giver's pre-printed route based on verbal directions from the Instruction Giver. In the maps, several mismatches in the route between the two partners map were intentionally included to induce conflict between them. A (blocking) screen was placed between them for visual separation. They then switched roles and were given another set of maps to generate another conversation to complete the task (Map Task 2). The Map task lasted for approximately twenty minutes. After that, participants took part in a five minute debriefing conversation; as the nature of the map task tended to induce some conflict between partners which they were motivated to resolve. Finally, to obtain spontaneous natural dialogue, participants were encouraged to engage in continuous speech on their chosen topic for fifteen minutes.

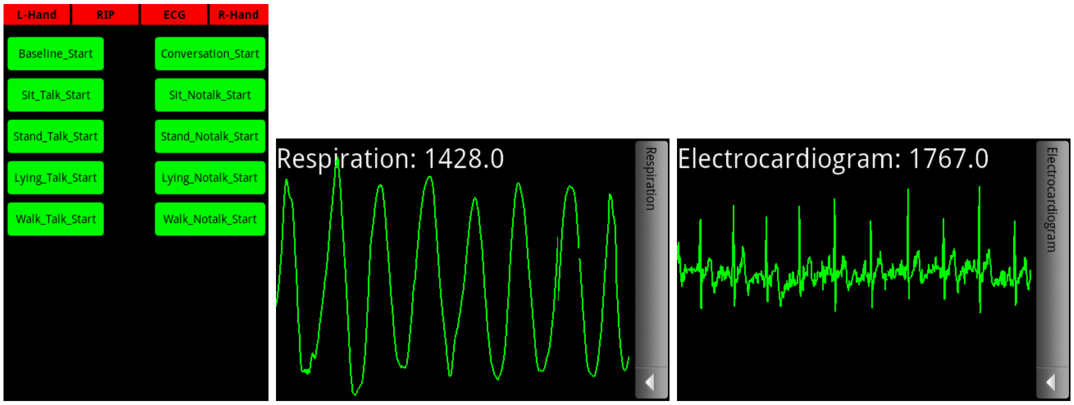


Fig. 3.4: Design of phone interface to collect conversation data in natural environment. Also participants could able to see the quality of collected signal (i.e., ECG, respiration) in the phone interface.

To collect interaction data in natural setting, participants were given a phone interface with labels: Walk-Talk and Walk-NoTalk. They were asked to mark the timing of different activities i.e., walking and high level conversational state, i.e. talking or not, on the study phone interface by choosing the appropriate label (shown in Figure 3.4a). Also they could able to see the quality of the collected physiological data, ECG and respiration in the phone interface(shown in Figure 3.4b,c).

3.2 Field Study Design to Capture Interpersonal Interactions

To understand the nature of stress patterns during stressful conversations and collect ecologically valid sensor data with precise labels for model development, we designed and conducted a field study. The ‘Field’ study was designed to (1) capture interpersonal interactions including stressful ones and other stressors in real life and (2) evaluate the performance of the conversation model in the natural environment. Participants wore the sensors for a day during their awake hours.

3.2.1 Field Study Requirements

To facilitate successful model development for detecting stressful conversations, we sought a study design that satisfies the following requirements to produce the necessary sensor data and associated labels.

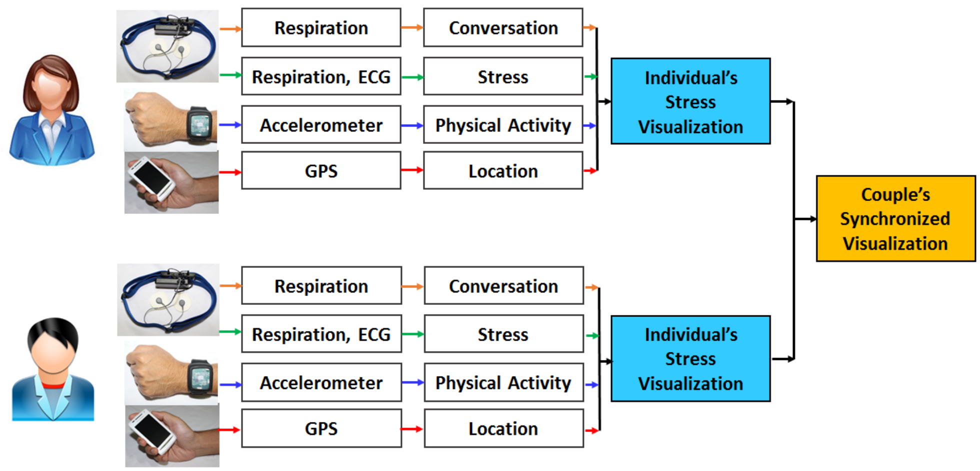


Fig. 3.5: Data collected in their natural environment. Both partners semantic locations, physical activity, conversation and stress data is inferred from the sensor data and feed to develop a visualization

1. **Ecologically Valid Sensor Data:** The study should capture physiological sensor data from the field environment during real-life stressors of different types. (Section 3.2.2)

2. **Stress Event Localization:** The start and end times of each stress event should be located precisely in the sensor data stream. (Section 3.2.3)

3. **Stressor Labels:** Each stress event should have an assigned label of reason, i.e., stressor. (Section 3.2.4)

4. **Resolving Ambiguity in Stressor Labels:** Each detected stress event, especially stressful conversations, should be independently confirmed so as to remove any ambiguity due to machine learning models or recall errors by the participants. (Section 3.2.6)

5. **Coverage of Stressful Conversations:** The study should have appropriate consent and sensor data available from both the conversing partners, including during stressful conversations. (Section 3.2.5)

In the following, we describe how the study we conducted satisfies each of these requirements.

3.2.2 Wearable Devices for Ecologically Valid Sensor Data

To capture reliable physiological data in the field, participants wore a chest band (Figure 3.2a) with Electrocardiogram (ECG) and respiration sensors [42]. To capture physical activity that can confound the inference of stress form physiological sensors and to provide physical activity context surrounding stress events, the chestband included 3-axis accelerometer signals. To capture hand gestures during conversations, the participants also wore a wristband consisting of a 3-axis accelerometer and a 3-axis gyroscope on their dominant hand (Figure 3.2b). To unambiguously verify the occurrence and timing of stressful conversations, they wore a LENA audio recorder [27] to capture high-quality audio (Figure 3.2c). They were instructed to carry the recorder in a pouch placed around the waist to reduce occlusion of the microphone and to increase the likelihood of capturing high quality audio.

To capture the location context, each participant was provided with an Android smartphone that collected GPS-traces (Figure 3.2d). For time synchronization among all sensor signals, the smartphone also received and stored data from all wearable sensors. Participants were asked to carry all the devices during their waking hours except during showers and contact sports, to maximize the opportunity to capture sensor data during stress events.

3.2.3 Stress Detection and Stress Event Localization

To meet the requirements of precisely locating the start and end times of stress events, we employed previously validated algorithms on the collected sensor data. We first use the cStress model [19] to obtain a stress state from each minute of ECG and respiration signals that represent the physiological response to a stressor. The model outputs a probability measure of stress scaled between 0 and 1, termed as ‘stress likelihood’ as shown in Figure 3.6. From ECG, the model computes the mean, median,

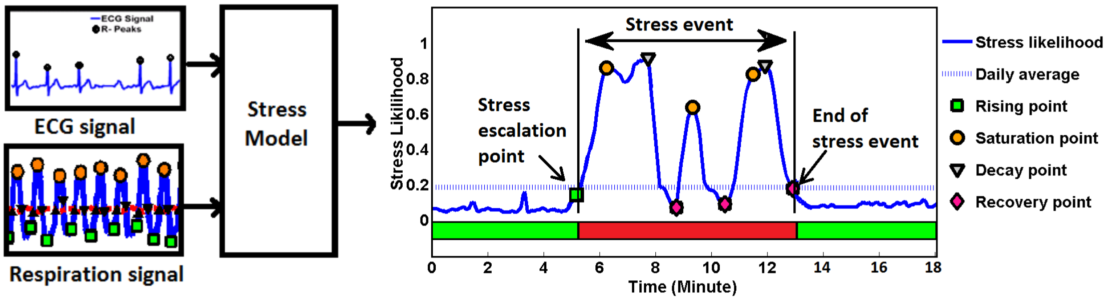


Fig. 3.6: Inferences of continuous stress likelihood using ECG and respiration signal.

20th, and 80th percentiles of heart rate, variance, and quartile deviation of HRV and energy of HRV in different frequency bands (0.10.2Hz, 0.20.3Hz, 0.30.4Hz). From respiration signals, it computes mean, median, 80th percentile, and quartile deviation from inhalation (I), exhalation (E) duration, ratio between I/E, stretch, and inspiration volume, computed in each breath cycle within a minute. In cross-subject validation using SVM on lab data, the cStress model classified stress and non-stress minutes with an F1 score of 0.81 in $(n = 21)$ participants who were subjected to three validated stressors public speaking, mental arithmetic, and cold-pressor tasks. When tested on a dataset from another group of participants $(n = 26)$ subjected to the same lab stress protocol, the model was able to classify stress and non-stress minutes with an F1 score of 0.9. The model was also evaluated against self-reports collected in the field. In the first study of $(n = 20)$ healthy adults who provided 1,060 self-reports in a 7-day study, the model reported an F1-score of 0.71 for the median participant. On a second field study with $(n = 38)$ polydrug users who wore the sensors for four weeks, the model reported a median F1 score of 0.72 [22]. In a third field study of $(n = 53)$ newly-abstinent smokers who wore the sensors for 4 days, the model reported a median F1 score of 0.65 [26].

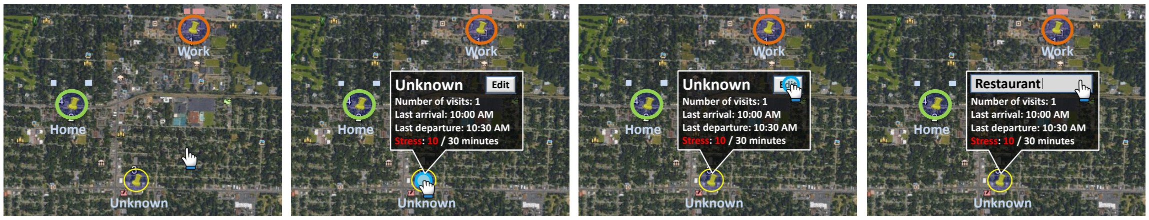
The cStress model only provides a stress likelihood for each minute, which does not indicate the start and end time of a stress event. To obtain stress events from the noisy and largely discontinuous (due to missing data or confounding from physical activity) time series of stress likelihoods, we apply the stress event detection model presented in [22]. This model first generates stress likelihood in minute-windows using the cStress model, but sliding every 5 seconds, to reduce the noise in the stress likelihood time series. Second, it excludes any data when participant may be recovering from physical activity (after accelerometer signals show no activity). Third, it uses k-nearest neighbor approach to impute any missing values of stress likelihood that is 'missing at random'. Fourth, it applies a moving average convergence divergence (MACD) method to find the cross over points that partition the continuous stress likelihood time-series into stress events, clearly marking the start and end times, as shown in Figure 3.6. Fifth, it excludes any windows that have more than 50% of stress likelihoods imputed. Finally, it applies a density threshold (to the area under the stress likelihood curve) to decide which windows are stressful events. In the field-collected data, between 2 and 4 stress events per day were detected [22].

3.2.4 Context Inferences and Visualization for Stressor Label Assignment

To obtain stressor labels for each of the detected stress events in the field, we wanted to assist the participants in recalling the surrounding contexts for the detected stress events so that they can confirm or refute the detected events and then recall the reasons for stress. To aid their recall, we detected several contexts such as location from GPS, conversation status from respiration signal, and activity status from accelerometers. This information was presented to the participants so that they could reconstruct those moments of stress events and recall the stressor responsible for that stress event. We first describe how we process the sensor data to obtain the surrounding contexts and then present the visualization.

Inferring Significant Locations Using Historical Map-Based Visualization:

Location is an important memory cue. When it is annotated with a time range, this information can help users to reconstruct their day and facilitate self-reflection [44]. Locations of interest are places where a user spends a significant amount of time. We adopted the spatio-temporal clustering algorithm proposed in [45] to infer significant



Unknown
Number of visits: 1
Last arrival: 10:00 AM
Last departure: 10:30 AM
Stress: 10 / 30 minutes

Unknown
Number of visits: 1
Last arrival: 10:00 AM
Last departure: 10:30 AM
Stress: 10 / 30 minutes

Restaurant
Number of visits: 1
Last arrival: 10:00 AM
Last departure: 10:30 AM
Stress: 10 / 30 minutes

Fig. 3.7: The circles represent significant locations visited by a user in a day. At a given location, the thickness of the circle corresponds to the duration of time spent and its color indicates the intensity of the average stress. Significant places can be labeled by the user. Clicking on a pushpin displays the frequency of visit to the location, start and end times of the last visit, and the duration of time that the user was stressed at that location. Users can edit and relabel the unknown locations, as shown in the picture.

locations, arrival time, departure time, duration of stay, and sequence and frequency of location visits throughout the day, all from GPS traces. A distance threshold of 100 meters and a time threshold of 10 minutes were used to find the spatio-temporal clusters.

We utilized a map-based visualization technique (as shown in Figure 3.7) developed in [46] to observe the location clusters on Google Earth. Labeling of the location clusters was semi-automated. The two most common location clusters, *home* and *work*, were automatically labeled based on the address provided by the participants at the beginning of the study. To label the remaining location clusters, the participants were asked to provide the semantic labels during the data review session. This helped resolve ambiguities for co-located places (e.g., grocery store and a restaurant). Distinct semantic locations thus obtained included: *own home*, *parent's home*, *others home*, *work*, *restaurant*, *store*, *grocery*, *religious place* (e.g., church, mosque), and *recreation center* (e.g., gymnasium).

Inferring Commute:

Driving episodes are detected from GPS-derived speed by applying a threshold for maximum gait speed of 2.533 meters/second [47]. A driving session is composed of driving segments separated by stops, e.g., due to a traffic light. The in-between stops usually are of short duration unless there is traffic congestion. The end of a driving session is defined as a stop (with speed equal to zero) for more than two minutes. Driving segments, separated by less than two minute stop, are considered to be part of the same driving episode [48].

Inferring Physical Activity:

For activity inference, we use the on-body accelerometer based activity detection approach presented in [49]. The pre-processing steps include filtering of raw data and removal of gravitational acceleration and drift from the filtered data. Finally, we compute the standard deviation of the magnitude of acceleration ($a_{mag} = \sqrt{a_x^2 + a_y^2 + a_z^2}$), which is independent of the orientation of the accelerometers. We use this measure to perform activity detection for each 10-second segment.

Inferring Conversation Episodes:

For detecting conversations from respiration data, we used the method proposed in [17]. This model extracts features in respiration cycles in each 30 second window, trains a machine learning model to produce speaking, listening, or quiet states, and then applies a Hidden Markov Model (HMM) to construct the conversation status for each 30 seconds window of respiration data. It achieves 87% accuracy in distinguishing conversation from non-conversation.

Contextualized Timeline Visualization to Assist in the Recall of Stressors

We developed a contextualized timeline visualization by building upon stress visualizations presented in [50]. In order to help the participants reconstruct the moments surrounding the stress event, we made several adaptations in the visualization, guided by the *Day Reconstruction Method (DRM)* [23].

We incorporated three design qualities for effective health data representation [51]. (1) the design must feel familiar to users, mirroring their own experience, (2) creating designs that leave space for users' own interpretation of their bodily data, and (3) the modalities used in the design do not contradict one another, but instead harmonize, helping users to make sense of the representation.

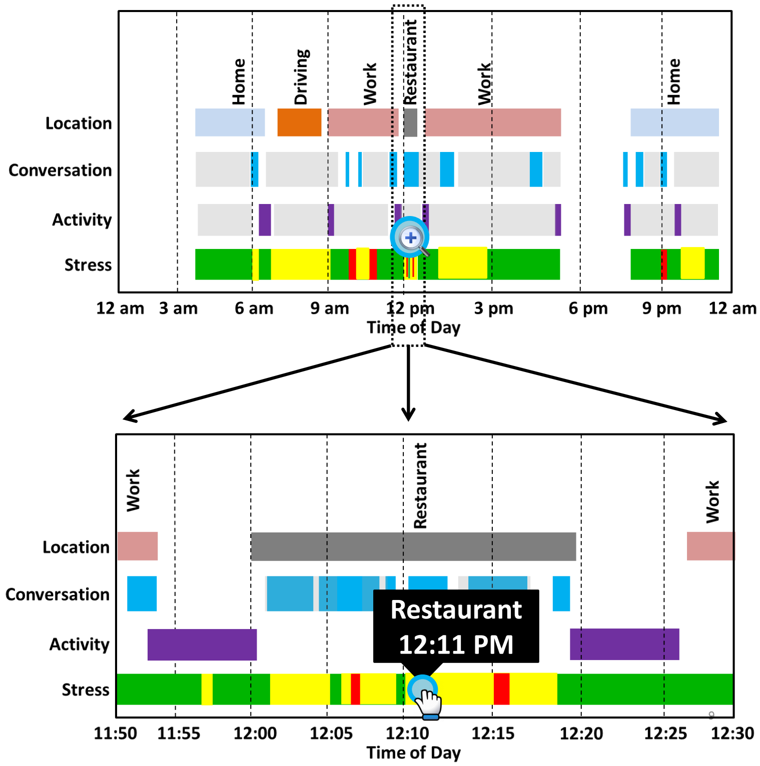
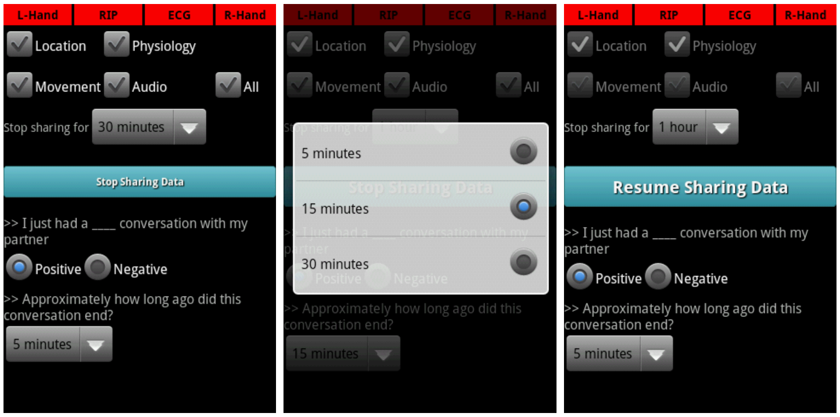


Fig. 3.8: Stress timeline visualization consists of four channels of inferences. Significant locations are marked with corresponding semantic location labels (e.g., *Home*). Dark color represents the presence of conversation (blue) and activity (purple) and grey color implies its absence. The bar display for Stress has three colors (Green = No stress, Yellow = Medium stress, and Red = High stress). Gaps in any of the channel indicate unavailable data. The interface has zoom-in (e.g., restaurant is zoomed-in in the lower figure) and zoom-out plus info-tip features (shown in black box with exact time in the lower zoomed-in part) to precisely pinpoint each stress events and corresponding contexts.

We created a stacked timeline visualization shown in Figure 3.8 for individual users. We used horizontal and vertical placement along with color coding as our visual encoding channel as these channels are most effective in supporting the comparison of multiple data streams [52]. In the timeline, the horizontal axis shows the time of day, and vertical axes is divided into four channels that represent four inferences (location, conversation, activity, and stress likelihood). We use hue as the color component to code different levels of stress — green represents no stress, yellow stands for medium, and red indicates high levels of stress likelihood (based on perceived stress categories reported in [53]). Deeper shades of color for conversation and activity time series show the



☑ Location ☑ Physiology
☑ Movement ☑ Audio ☑ All
Stop sharing for 30 minutes
>> I just had a ___ conversation with my partner
● Positive ○ Negative
>> Approximately how long ago did this conversation end?
5 minutes

○ 5 minutes
● 15 minutes
○ 30 minutes

☑ Location ☑ Physiology
☑ Movement ☑ Audio ☑ All
Stop sharing for 1 hour
>> I just had a ___ conversation with my partner
● Positive ○ Negative
>> Approximately how long ago did this conversation end?
5 minutes

Fig. 3.9: Phone interface for the field data collection. User can select the type of data they don't want to share selecting the radio button and duration from the drop down menu.

occurrence of conversation and physical movement, respectively, and grey color indicates the absence of conversation or absence of movement. Significant locations are marked with corresponding labels. If a transition between locations takes place using a motorized vehicle, then the transition is labeled as commuting. For all the four data streams, the presence of a gap implies missing data for that time period. Aligning all data streams using the same timeline facilitates understanding of the role of different contexts such as location or conversation on stress events.

It is difficult to pinpoint a stressful event when the data is on the scale of several hours (e.g., over 12 hours of data was collected per day). Therefore, we use interaction to provide users the ability to zoom in and out at different temporal resolutions. By providing details-on-demand, we allow users to view precise stress likelihood levels and associated contexts (e.g., location, conversation, and physical activity status). To help them in recalling a specific event, we use tool-tip texts displayed at the time of occurrence of each event.

3.2.5 Participant Selection and Protocol To Capture Real-Life Stress Events

We recruited couples to wear sensors and collect data concurrently to maximize the coverage of stressful conversations. The field study included 38 individuals (19 pairs of cohabiting couples). Field study participants included 20 women (mean age: 28.53 ±

4.89 years) and 18 men (mean age: 28.92 ± 2.10 years). Eighteen participants were Caucasian and the rest were Asian. Twenty participants (10 pairs) participated during weekdays and the rest participated during weekends.

The field study consisted of three phases — (1) an enrollment session, (2) free-living data collection, and (3) a data review session to label detected stress events using the visualization. During the enrollment session, participants gave consent and completed a demographic questionnaire, a dyadic adjustment scale [54], and a pre-study questionnaire. Participants were shown an example visualization generated from previously collected sample data. This was designed to help them understand how the field data collected would help them understand their own stress patterns and identify daily stressors for potential stress management in daily life. This orientation was also designed to motivate the participants for careful data collection when they were in free-living condition.

Afterward, participants were shown how to wear the sensors and monitor the status of sensor data collection. They then proceeded to collect sensor data in the field. After completing at least 24 hours with the sensors since the start of the data collection, both partners came back to the lab next day to review stress visualizations generated from their own data and annotate the automatically detected stress events captured in the field. Each individual was compensated at a rate of $2.50 an hour for up to 12 hours for field session data. The maximum amount of compensation each individual could earn for the field session was $30 (12 hours x $2.50/hour). Also, each individual received $10 for the data review session. Thus, each individual earned $60 for participating in the study.

Because the field study involved collection of continuous audio, location, and physiological data from the participants, they were given an option to pause data collection during their private moments. They could proactively pause data collection using the “Stop” button in the smartphone software (see Figure 3.9) during data collection in the field. Also, they were given the option to retroactively delete data during private

Table 3.1: Summary of stress events captured in participant's daily life.

Stressors	Number of stressful events	Average event duration (Minute)	What's going on during stressful events
Stressful Conversations	53	22.68 (3.83)	Conversations with partner, friends, colleagues, supervisor
Commute	30	12.74 (2.28)	Time pressure, other driver's behavior, construction on road
Work	14	18.23 (3.54)	Deadline, answering work related email/text

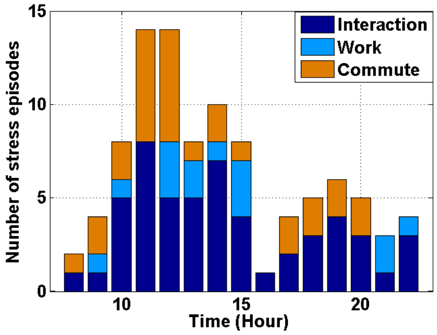


Fig. 3.10: Distribution of stress events throughout the whole day.

moments during the data review session. The data collection was limited to 24 hours to reduce privacy concerns associated with the raw recording of audio data in the natural environment; participants were instructed to get verbal consent from conversation partner(s) other than their romantic partner before recording audio conversation involving them. If any partner(s) declined the request, participants were instructed to stop recording the audio.

3.2.6 Stressor Labels Collected and Confirmed

To resolve any ambiguity in stress event detection due to the usage of machine learning models from sensor data, including the elimination of any false detection, the participants were asked to confirm each stress event in the visualization of their data. To further confirm the stress events and to contextualize it, several follow up questions were asked such as *"what's going on?"*, *" where were they?"*,*"who were they with?"*.

Participants were asked to rate the usability of the visualization interface on a 5-point Likert scale. We asked them if the interface was *"Easy to understand"*, if they felt that *"Visualization helped understand both risks and benefits"*, and finally, if *"they thought that most people would learn to use the visualization quickly"*. All the participants either agreed (6 out of 38) or strongly agreed (32 out of 38) that the visualization was easy to understand. Thirty out of 38 agreed or strongly agreed that most people would learn to use these visual representations quickly. We also asked each of them an open ended question: "What things did you *Like* and *Dislike* in the study". Twenty seven participants responded to this question, and 20 mentioned that they liked the stress visualization system. For example, C4F commented, *"[I] Liked visualization of the day, disliked wearing all the sensors"*.

Participants recalled several reasons for stress events (i.e., stressors) such as meeting with a supervisor, having deadlines at work, job interviews, conflict with their partner, driving on a busy road, assignment deadlines, etc. For the 12 events, they either disagreed with the visualization output or could not remember whether the stress event occurred. In addition, we asked all the participants whether they recalled any stress event that happened during the study that was not identified by the system (false negative). Two participants (out of 38) reported three such false negative events (over 38 person days of data collection). These three stress events missed by the sensors were not included in our model training or testing as the start and end times of these events could not be determined precisely.

To resolve any ambiguity in the start and end of stressful conversations, we verified the occurrence of conversations by listening to the raw audio. We find that each stress event attributed to stressful conversations were correctly labeled. It may be because of our contextualized visualization that showed the participants whether they were having conversations at the time of a detected stress event and where they were, e.g., at home or office.

Participants were able to recall 97 stressful events, during which sensor data was available and not confounded by physical activity and hence usable for sensor-based stress inference. We find that all such detected stress events belong to three major categories — stressful conversations, commute, and work-related stress. Table 3.1 shows the number of stress events in each category, the average duration of stressful events, and what's happening during these moments. In our data set, we find that 53 stressful events were due to conversations with partner, friends, parents, colleagues, supervisors, etc., accounting for almost 54% of all stress events. We also found 30 stressful events during commute and 14 events due to work. Any stress event that involved a conversation whether at home, work, or anywhere else, is included in the category of stressful conversation. The same would be the case for work-related stressor, unless it involved a conversation, in which case it belongs to the stressful conversation category. We note that the percentage of stress events in each category matched with the percentage reported in [3]. The distribution of stress events in our dataset in these three categories is shown in Figure 3.10.

Participants reported several stress events that did not belong to the above three categories. For example, they mentioned household chores (8), stress during shopping or grocery (5), and miscellaneous (15) stress events that included feeling sick, another family member is sick, worrying about the partner, water leaking inside house, cleaning the house, etc. We were unable to use these stress events in our modeling because the sensor data collected during these events were confounded by physical activity. Hence, they were excluded from our modeling.

Chapter 4

Processing of Respiration Signal

To develop stressful social interaction model from stress time-series, we need a a stress model that outputs a continuous stress probability for the whole day. We also need a conversation model to develop the visualization system. In this work, we aim to use respiration signal to develop those models. To achieve these goals, first we need to identify each breathing cycle from the respiration time-series. Next we need to compute features from each breathing cycle that work for both models. In this chapter, we present a rigorous method for screening, cleaning respiration signals and improved algorithms for identifying respiration cycles captured in field setting.

4.1 Data Screening and Processing to Locate Each Breath Cycle

Data collected in field environment are subjected to various sources of artifacts, losses, and degradation in quality. The first challenge is, accurate identification of a breathing cycle i.e., pinpoint several interesting points of a cycle such as the onset of an inspiration, the onset of an expiration. Second challenge is to handle the effect of activity and postures. Respiration signal can be easily influenced by movements of limbs and torso, changes in posture (i.e., sitting, supine), and physical activity (walking). To support the physiological need for various activities, inhalation and exhalation duration and magnitude of the signal may change significantly (see Figure 4.1). Rigorous data

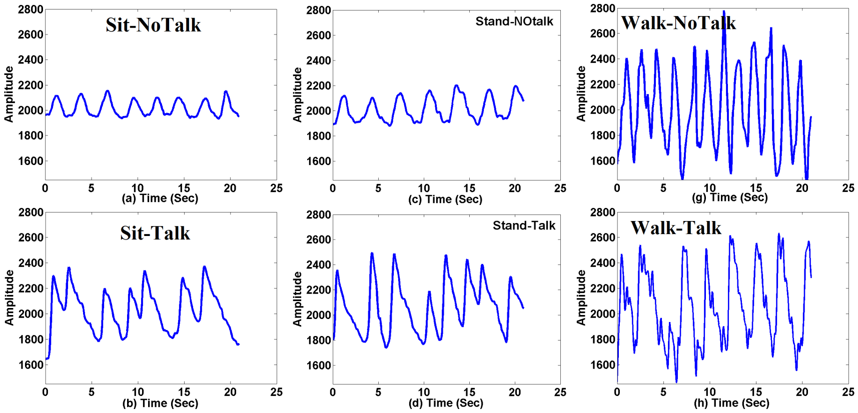


Fig. 4.1: Effect of postures, physical activity and vocalization on breathing cycles.

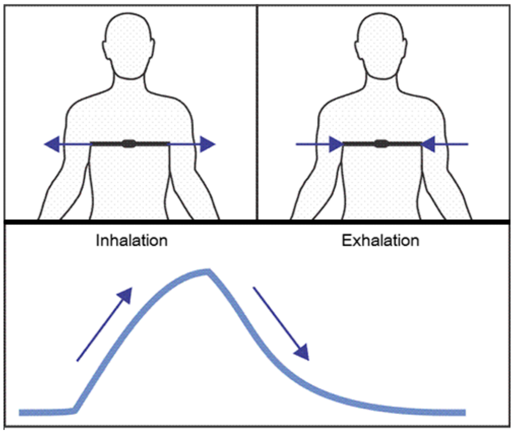


Fig. 4.2: Effect of postures, physical activity and vocalization on breathing cycles.

processing is essential to obtain usable results from the data collected in the field. Here, We describe a series of methods to screen, clean and process to locate each breathing cycle.

4.2 Background of Respiration Signal Morphology

Rib bones in combination with diaphragm help us breath air in and out from our lung. During inhalation, external intercostal muscles (tiny muscles located in between each rib) nearest the sternum contract and lift the rib cage up and out to make more room for the lung. As we exhale, the internal intercostal muscles contract and allow the weight of the ribs to move back down. On the other hand, the diaphragm can operate as a voluntary muscle or involuntary muscle, thus allowing us to hold our breath or slow our breathing if we wish to. When the diaphragm contracts, it moves down towards the stomach. This creates a vacuum in the cavity containing the lungs. This vacuum causes the lungs to expand and pull air down and in. When we breath out and the diaphragm relaxes and moves up again.

Respiratory Inductive Plethysmograph (RIP) sensor around the chest captures the above phenomenon and generates breathing signal shown in Figure 4.2. Waveforms of the breathing signals varies based on the current context and underlying activities of the user. For example, breathing, while user is sitting quietly, looks similar to sinusoidal wave. However, duration of the waveform decreases due to physical activity. Speech breathing

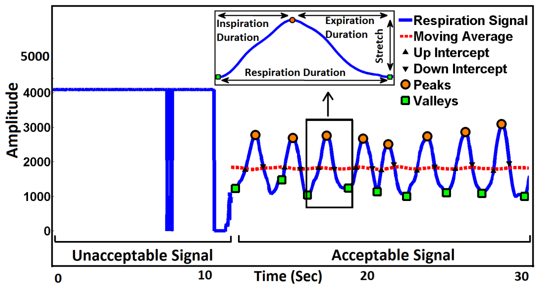


Fig. 4.3: Unacceptable signal looks flat and saturated at the top, whereas legitimate signal follows sinusoidal pattern. Each cycle is segmented using moving average based peak- valley detection.

becomes more similar to saw-tooth shape [55]. Respiration also varies across demographics such as, age and weight, due to weather conditions,etc.

4.3 Quality Screening of Respiration Signal

Breathing dynamics can be captured using respiratory inductive plethysmograph (RIP) by tracking the rhythmic motion of ribcage during breathing. Thus respiration signals are largely affected by physical movement and positioning of the chest band, we mark the signal acceptable as long as the signal follows rhythmic pattern. Another challenge is, slipping of the band from its expected location which sometimes results in a low amplitude signal, still considered acceptable if it retains the characteristic morphology of a respiration signal. Detaching of sensor from body results in a low variation which is considered unacceptable.

4.4 Cycle Identification

The first stage in detecting conversation from respiratory waveforms is the automated detection of individual breath cycles. Manual data labeling of each respiration cycle is hard and time-consuming, especially for 12 hours of respiration data per participant which consists of on average 10,000 breath cycles. Hence, an automatic method is needed to identify breath cycles without human intervention. In this section, we describe a method to identify respiration cycles automatically in both the lab and field settings.

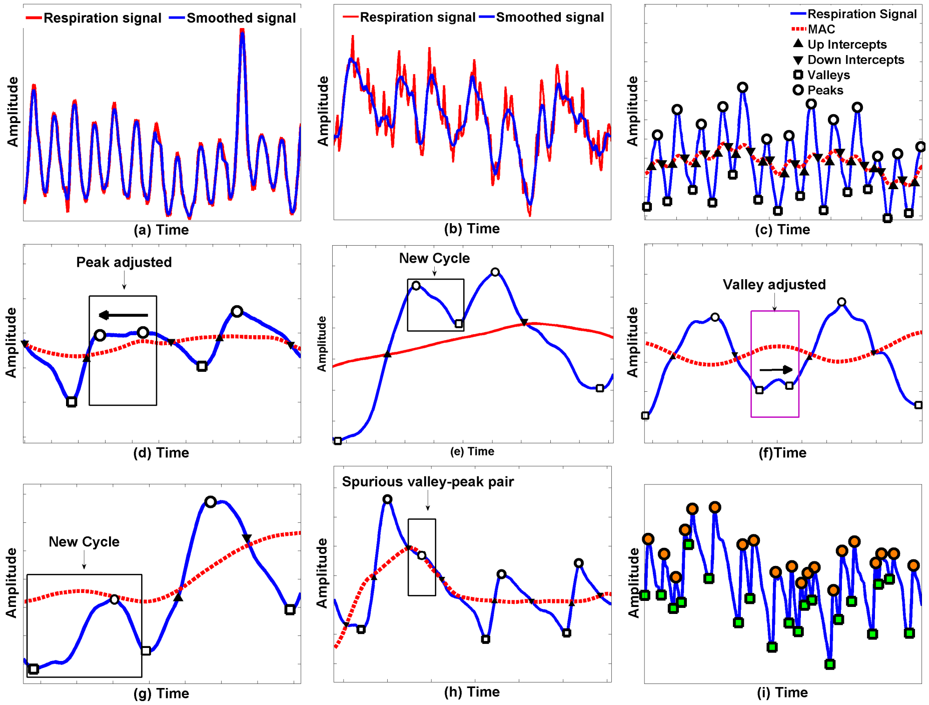


Fig. 4.4: (a) Raw and smoothed signal during sitting. (b) Raw and smoothed signal during walking. (c) The moving average curve (MAC) closely follows the trend in the respiratory signal. Peaks and valleys are respectively determined by the maximum and minimum between pairs of alternating up intercepts and down intercepts. (d) There is a breath hold near the peak region which results in a wrong peak position. The peak is automatically shifted towards the left to a point where majority of inspiration has completed. (e) A new cycle is found above MAC as it satisfies all properties of a breathing cycle. (f) Taking a minimum results in a wrong valley due to the presence of an end expiratory pause. The valley is automatically shifted towards the right to a point where signal starts rising monotonically. (g) A new cycle is detected below MAC as it satisfies all properties of a breathing cycle. (h) Spurious valley-peak pairs are automatically removed if they are too close. (i) Final peaks and valleys identified by the algorithm.

4.4.1 Cycle Identification Algorithm

Step 1: Signal Smoothing. The first step is to smooth the raw signal using a moving average filter of M points. Let x be a respiration signal with M number of samples in the moving average, and y the smoothed signal. Larger values of M flatten the fluctuations in the signal. Respiration signals exhibit fewer bumps or small oscillations while the wearer is sitting or standing (see Figure 4.4a) as compared to walking. During walking, the body shakes or hands move back and forth for each step, causing visible bumps in the respiration signal as depicted in Figure 4.4b. Larger values of M reduce the impact of bumps in walking cycles and reduce the number of spurious cycles detected by the algorithm. If M is chosen to be too large, we risk over-smoothing and losing sharpness around points of interest (e.g., peaks and valleys).

We chose a value for M that balances the proportion of correctly identified cycles against the amplitude reduction due to smoothing. We iteratively tuned the value of M by applying the algorithm on field data. The most appropriate value of M was found to be 5 (250 ms) for sitting and standing signals, and 11 (515 ms) for walking. The equation for smoothing respiratory raw signals appears in Equation 4.1.

$$y(t) = \frac{1}{M} \sum_{j=\frac{-(M-1)}{2}}^{\frac{(M-1)}{2}} x(t+j) \tag{4.1}$$

Step 2: Moving Average Centerline (MAC). The next step is to compute a moving average centerline (MAC) curve using Equation 4.2, where y is the smoothed respiratory signal, L its duration, t is time, and $\overline{y(t)}|_{t-T}^{t+T}$ the average value of y during $[t1, t2]$. The MAC appears as a center line (shown as red dotted line in Figure 4.4c) that intercepts each breathing cycle twice, once in the inspiration phase and then in the expiration phase. T is the average cycle duration. The average cycle duration is 2.94 seconds.

$$MAC(t) = \overline{y(t)}|_{t-T}^{t+T}, \;\; if \;\; T < t \leq L - T \tag{4.2}$$

After visual inspection we found that, in cases of large baseline drift in field data, $T = 3$ seconds setting takes time to cope with the drift and results in missed cycles. We visually confirmed that $T = 2$ seconds is fast enough to keep track with the signal drift and intercepts more cycles in baseline shifted region. However, in the cases of regular/quiet breathing cycles, we found the $T = 2$ and $T = 3$ result in nearly the same performance and chose $T = 2$ for the window width.

Step 3: Intercept Identification. Next, we identify the points where the MAC curve intercepts the smoothed signal. The following equations are used to find the up intercepts where the MAC crosses the inspiration branch. Similarly, down intercepts are the points where the MAC curve crosses the expiration branch of the signal. Ideally, there should be exactly one up intercept and one down intercept for each breath cycle as shown in Figure 4.4c.

$I_{up} = y(t-1) \leq MAC(t) \leq y(t)$

$I_{dn} = y(t-1) \geq MAC(t) \geq y(t)$

Step 4: Intercept Screening. To avoid spurious intercepts, if there are more than two consecutive intercepts with the same label, only the last one is kept. The resultant sequence becomes: $I_{dn}(1) < I_{up}(1) < I_{dn}(2) < I_{up}(2) ... < I_{dn}(m) < I_{up}(m)$ where m is the number of up (down) intercepts.

Step 5: Peak (Expiration onset) Detection. The peak or onset of expiration of a breathing cycle is determined by finding the maximum between consecutive up and down intercepts using the formula,

$peak(i) = \max(y(I_{up}(i)) : y(I_{dn}(i+1)))$,

where $i = 1, 2, ..., p$ and $p =$ number of peaks. In cases of a regular breathing signal (as Figure 4.4c), taking a maximum provides the location of exact peak position. However, breathing signals may not always be so rhythmic (e.g. during speaking), thus the maximum value may not represent the actual peak position. If there exists one or more notches in the peak region as seen in Figures 4.4d and 4.4e, two things can happen — either the peak needs to be adjusted to its actual position or another cycle must be considered. In the first case where a peak needs to be adjusted, the maximum point among all the notches is considered as a candidate peak. We consider the maximum value as a peak if 70% of inspiration of that cycle is done up to that point. The value 70% was tuned from the annotated data collected in the lab.

However, if the MAC line fails to intersect small cycles at the top as shown in Figure 4.4e, there is a possibility that there exists another cycle within the detected cycle, thus shifting the peak to left may not suffice. To address this issue, we look for a portion within a cycle that looks like a breathing cycle, i.e., it has ascending and descending trends resembling inspiration and expiration phases. Then, we split the cycle into two. We detect the points of interest in the two newly formed cycles. If both cycles' inspiration and expiration durations are greater than 0.4 seconds [56,57], and total cycle duration lies within the range of 0.8 seconds to 12.5 seconds [19,57], we consider both cycles as valid cycles. If any of the newly formed cycles fail to meet these criteria, we assume there is only one cycle and the position of the peak is adjusted if required.

Step 6: Cycle's Start and End Point Detection (Valleys). In general, a valley is the minimum point between a down intercept and the following up intercept for a regular semi-sinusoidal breathing cycle. However, if a cycle has an expiratory pause, the minimum point may not represent the actual valley. Therefore, we consider the minimum as a candidate valley. From this candidate valley to the next up intercept, we compute all the slopes. By examining the slopes, we determine the point from where the signal monotonically rises towards the next peak and consider that as the actual valley (see Figure 4.4f).

However, the MAC curve may not intersect a cycle if the amplitude changes dramatically. For example, if the baseline shifts abruptly or there lies a small cycle adjacent to a larger one, a moving average can't cope with the change so quickly and may not intersect, as depicted in Figure 4.4g. Similarly, as described above, we look for a portion within a cycle that looks like a breathing cycle and detect the interesting points of the new cycle. If all the durations satisfy the standard durations [19,56,57], we consider both cycles as valid cycles.

Step 7: Peak-Valley Screening. When searching for peaks and valleys, only those where time intervals of more than 0.4 seconds [57] exist, from a peak to the next valley or from a valley to the next peak, assuming that the minimum breathing period is around 0.8s. Otherwise, the peaks and valleys are considered to be spurious are removed as shown in Figure 4.4h. Second, if an inspiration or expiration amplitude is too small, 10%

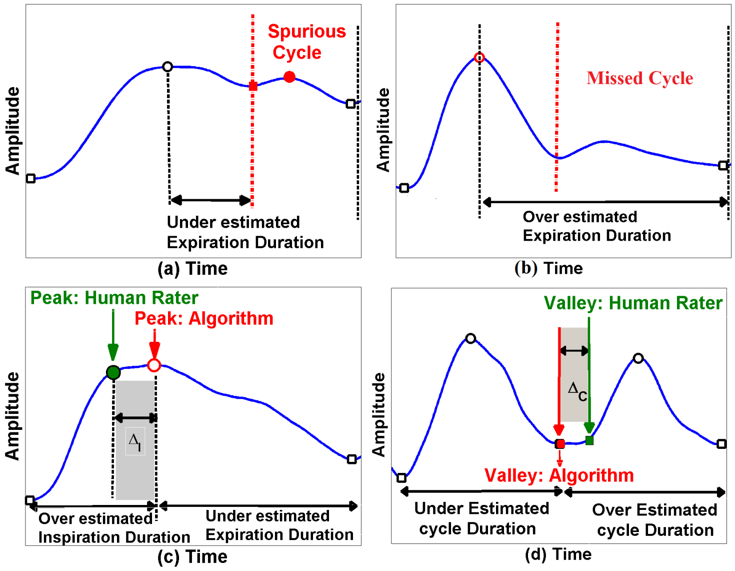

Amplitude
Spurious Cycle
Under estimated Expiration Duration
(a) Time
Missed Cycle
Over estimated Expiration Duration
(b) Time
Peak: Human Rater
Peak: Algorithm
Δ_I
Over estimated Inspiration Duration
Under estimated Expiration Duration
(c) Time
Valley: Human Rater
Δ_C
Valley: Algorithm
Under Estimated cycle Duration
Over Estimated cycle Duration
(d) Time

Fig. 4.5: Example of (a) Spurious cycle in the expiration region resulting in splitting of a true cycle into two. (b) A missing cycle resulting in one long duration cycle. (c) Mislocated peaks, (d) Mislocated valleys.

of the mean cycle amplitude, the associated cycle is not considered to be of good quality and is screened out.

4.4.2 Evaluation Metric

It is usual to compute the number of correctly identified peaks and valleys. They suffice when only the respiration rate is to be computed. However, they do not indicate the accuracy in features related to respiration rhythm (e.g., inhalation, exhalation) that are needed in inferences of speaking or smoking events from respiration signal. This is because even if the number of peaks and valleys are identified correctly, their respective locations in the signal waveform may introduce errors in the resultant features. For accurate inferences, the locations of peaks and valleys along both time and amplitude dimensions are important. Therefore, we use the following metrics.

1. ***Spurious cycle rate.*** A spurious cycle can affect the inspiration/expiration duration depending on where it is detected (see Figure 4.5a).

Spurious cycles Rate: Percentage of cycles that are spuriously detected with respect to the total number of actual cycles (N). N is the number of actual cycles annotated by human rater.

$Error(\%)$ =Number of spurious cycles$/N * 100$

2. ***Missed cycle rate.*** Missing of one or more cycles results in elongated cycle duration as shown in Figure 4.5b.

   Missed cycles Rate: Percentage of cycles that are missed with respect to total number of actual cycles (N). $Error(\%)$ =Number of missed cycles$/N * 100$

3. ***Error in Inspiration duration due to Mislocated Peaks.*** Mislocated Peaks introduce error in the corresponding cycle's inspiration and expiration duration although cycle duration may still be correct (see Figure 4.5c). Thus, a cycle's inspiration duration may decrease (increase) and that cycle's expiration duration may increase (decrease) depending on the peak position. This error can't be captured using the respiration duration. This absolute duration error is measured in seconds and defined as *Error in Inspiration duration (*Δ_I*)*

4. ***Error in Cycle duration due to Mislocated valleys.*** Incorrect positioning of a valley affects both the current and the next cycle duration as shown in Figure 4.5d which either underestimate or overestimate the durations of neighboring cycles. A mislocated valley decreases (or increases) the current cycle's duration and increases (or decreases) the next cycle's duration. This absolute duration error is measured in seconds and defined as *Error in Cycle duration (*Δ_C*).*

4.4.3 Algorithm Evaluation and Performance Comparison

We implemented two other widely used methods to compare with the performance of our algorithm. The first one is a threshold based method [17] where the threshold is set by taking the average of the signal for every 30 second window. The second one is a

Table 4.1: Performance comparison of the current method with the state-of-the-art cycle identification methods with lab data (with 1,938 respiration cycles). Paired t-test shows significant reduction in inspiration duration error with respect to the existing methods and the base method (p-value < 0.001). The cycle duration error is significantly higher in the Threshold method, compared with other methods.

Methods	Spurious cycles	missed cycles	Error in Inspiration duration (second)	Error in Cycle duration (second)
Threshold based	1.5%	61.7%	0.81 ± 0.02	6.59 ± 0.04
Maxima-Minima	6.6%	4.0%	0.42 ± 0.01	0.45 ± 0.41
Base Method	2.1%	12.2%	0.44 ± 0.02	0.68 ± 0.06
Current Method	3.1%	5.6%	0.29 ± 0.01	0.43 ± 0.04

change point detection method described in [58]. We also present the performance evaluation of the semi-automatic method [59], which we call the 'base method'.

Evaluation on Lab Data.

We compare the performance of the current method on lab data (1,938 marked respiration cycles) with the base method [59] as well as two other methods i.e., the threshold based and Maxima-Minima based methods. The results are presented in Table 4.1. In comparison with the base method, percentage of missed cycles reduces from 12.2% to 5.6 % though spurious cycles increase by 1% in the current method. The Maxima-Minima based method detects extra 6.6% as spurious cycles and misses 4% cycles. The original threshold based method [17] was developed using filtered respiration signals. This might be one reason for so many missed cycles i.e., 61.7% using our unfiltered respiration signals.

Paired t-tests show significant reduction in inspiration duration error (p-value < 0.001) with respect to the base method and the existing methods. However, in the case of cycle duration, error has significantly dropped with respect to the base method and the threshold based method (p-value < 0.001), but no significant difference is found with Maxima-Minima based method.

Evaluation on Data from a Natural Setting.

To measure the performance with field data, we applied all the methods on data that includes several postures and activities, such as sitting, standing, walking and conversation. Two human raters annotated these data independently and inter-rater agreement between them was > 0.81.

Evaluation on real-life data shows that among 1,500 respiration cycles (around 2 hours) that occurred in the presence of physical activity, overall, the current method accurately identified 96.34% cycles, missed 3.66% cycles and identified extra 1.9% cycles as spurious (Table 4.2). Overall performance of the Maxima-Minima method revealed that it could identify 99.64% cycles accurately and detect an extra 16.71% cycles as spurious. The base method identified 89.83% cycles correctly while it missed 10.16% cycles and no spurious cycles were found. Table 4.2 shows that most spurious cycles were found during walking for both the Maxima-Minima method and the current method. Spurious rate was higher during walking because of the presence of bumps in the respiration cycle as shown in Figure 4.4b.

Table 4.3 shows that the performance of cycle detection methods vary in presence of conversation. Maxima-Minima method located 99.22% true cycles with 35.95% spurious cycles. the base method detected 82.63% cycles correctly with a miss of 17.37%. However, our current method identified 94.84% cycles correctly with a miss of 5.16% and 4.17% spurious cycles.

4.5 Related Work

In this section, we discuss the traditional approaches to process and identify breathing cycles. The simplest procedure for detecting breaths is a threshold level detector [17, 60, 61, 62]. In this approach, a breath is detected when the waveform passes through a predetermined threshold level in a given direction (i.e., up or down). The difficulty in this approach is finding an appropriate threshold that works across diverse participants and diverse contexts e.g., conversation, physical activity. Using too small of a

Table 4.2: Performance evaluation of breathing cycle identification methods in presence of physical activity and postures. Here, spur.= spurious.

Methods	Walking (%)			Sitting (%)			Standing (%)			Overall (%)		
	True cycle	Miss cycle	Spur. cycle	True cycle	Miss cycle	Spur. cycle	True cycle	Miss cycle	Spur. cycle	True cycle	Miss cycle	Spur. cycle
Threshold based	69.03	30.97	0.79	71.99	28.01	0.69	75.96	24.04	0.00	72.15	27.85	0.54
Maxima-minima	98.99	1.01	40.55	100	0.00	6.73	99.74	0.26	7.99	99.64	0.36	16.71
Base method	85.64	14.36	0.00	94.10	5.90	0.00	87.37	12.63	0.00	89.83	10.16	0.00
Current method	97.14	2.86	4.68	97.17	2.83	0.83	94.20	5.80	0.79	96.34	3.66	1.90

Table 4.3: Performance evaluation of breathing cycle identification methods in presence of conversation collected in field.

Methods	Conversation (%)			Non-conversation (%)		
	True cycles	Missed cycles	Spurious cycles	True cycles	Missed cycles	Spurious cycles
Threshold based	72.36	27.64	1.42	72.03	27.97	0.00
Maxima-minima	99.22	0.78	35.95	99.89	0.11	5.46
Base method	82.63	17.37	0.00	94.02	5.98	0.00
Current method	94.84	5.16	4.17	97.21	2.79	0.58

threshold may create spurious peaks whereas too large of a threshold may lead to missed peaks. Moreover, body orientation may shift the signal baseline. To allow for changes in mean level, a moving baseline can be used, but even then sudden mean level changes will still result in missed breath detection.

Another popular technique to find respiration cycles is to use a change-point detection algorithm (i.e., track local maxima and minima) [58, 63, 64]. However, there can be a large number of change points even within a cycle, especially in the presence of activity (e.g., walking,). Hence, more sophisticated methods are needed to discard excess peaks.

A semi-automatic method was developed for peak and valley detection in free-breathing respiratory waveforms in [59]. Breath cycles are identified by locating the intercepts of a moving average with the inspiration and expiration branches of the signal and finally manual adjustments are applied. Because manual selection is not practical for a dataset containing a large number of respiration cycles, a computerized method is desirable. Another semi-automatic method for detecting breathing cycles is proposed in [65], which also needs user intervention to make a decision either to: keep, adjust/move, delete or add points of interest.

None of the above mentioned methods are validated in natural environments to identify breath cycles in different situations e.g., in the presence of physical activity or conversation. We build upon the method proposed in [59]. We make several improvements to clean, screen, and detect breath cycles accurately in the natural environment. Our current method shows the feasibility of identifying breath cycles in both lab and field data, and to locate points of interest within a cycle, e.g., peak, start and end of a cycle. Among 1,934 respiration cycles collected in lab in presence of conversation, the proposed cycle identification method can identify 94.4% cycles correctly. Among 1,500 cycles collected in natural environments, the proposed method identified 96.34% cycles correctly in the presence of physical activities (walking) and in different postures (e.g., sitting and standing). In the presence of conversation, this method correctly identifies 94.84% of cycles collected in the field environment.

4.6 Conclusions

Reliable detection of respiration cycles results in accurate feature calculation. In the following chapters, these features are used to model behaviors, such as conversation and stress inferences from respiration signal.

Chapter 5

Conversation Detection from Respiration Signal

In this chapter, we explore the potential for detecting conversations from respiratory measurements. As respiratory cycle is a unit of speech breathing, cycle-based classification is the finest granularity for speech modeling from respiration data. Each respiration cycle dynamically varies in duration. Hence, cycle-based dynamic windowing is an appropriate approach for the respiration based speech modeling as presented in the current model. To generate labels, speech/non-speech cycles were carefully marked based on audio, video, and hospital grade respiratory inductive plethysmograph bands with synchronized channels in the lab setting and by using audio processing from LENA and confirmation from human raters in the field. The details of data collection is described in section 3.1. Here, we describe the development of machine learning model to detect conversation in field.

5.1 Speech Detection Using Conditional Random Field-Context Free Grammar

Given a sequence of respiration cycles, we now turn to the problem of labeling each cycle as corresponding to speech or not and segmenting these cycles into period of conversation. We achieve this using a Conditional Random Field Context Free Grammar (CRF-CFG) model. In this section, we begin by reviewing the CRF-CFG model [66] and then describe how we apply it to speech detection and conversation episode segmentation. The CRF-CFG model was first used in mHealth to extract heart-beat signal morphology (QRS complex) in ECG time-series data [67]. To the best of our knowledge, ours is the first work to apply CRF-CFG model for detecting conversation episodes on respiration time-series data. We begin by reviewing the conditional random field (CRF) model [68] and context free grammars (CFGs) and then describe how a CRF can be used to parameterize a distribution over parse trees. Finally, we present the CFG used for speech detection and conversation episode segmentation. In section 5.5, we present experiments validating this model on the lab and field data described in previous sections.

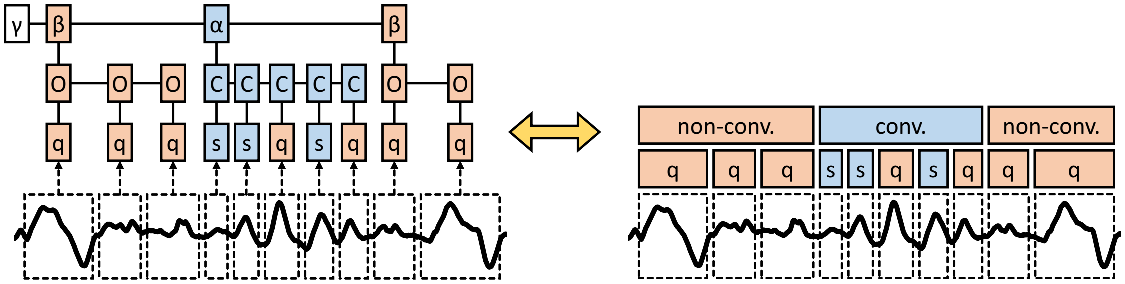


Fig. 5.1: An example parse (left) using the grammar described in equation 5.2. Also shown is the mapping from the parse to a labeled segmentation (right) where q and s stand for quiet and speaking respectively.

5.1.1 Conditional Random Fields

Conditional random fields (CRFs) are a sub-class of probabilistic graphical models [69] that encode correlations between label variables. A CRF defines a conditional distribution over a set of L label variables $\mathbf{Y} = \{Y_1, ..., Y_L\}$ given a corresponding set of M feature variables $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_M\}$. We assume each feature variable $\mathbf{X}_i \in \mathbb{R}^D$ is a D dimensional real vector and label variable Y_i take values in a set $\mathcal{Y}_i$; however, there may be additional constraints on the set of possible joint configurations, denoted by $\mathbf{Y}$. Throughout this work, we will use upper-case to refer to random variables (e.g., $\mathbf{Y}$) and lower case to refer to particular assignments to those variables (e.g., $\mathbf{y}$).

A general log-linear CRF is defined through a linear energy function that takes the form of a weighted sum of K feature functions f_k involving values of $\mathbf{Y}$ and $\mathbf{X}$:

$$E_\theta(\mathbf{y}, \mathbf{x}) = -\sum_{k=1}^{K} \theta_k f_k(\mathbf{y}, \mathbf{x})$$

These feature functions are typically sparse in the sense that they involve few label and feature variables. The set of label and feature variables referenced in function f_k is referred to as its scope S_k. If S_k contains at most two variables for all k, then the model is referred to as a pair-wise CRF, and it can be represented using a graph $\mathcal{G}$ where an undirected edge connects each pair of variables that share a scope. If the graph $\mathcal{G}$ is a tree, then the resulting CRF is referred to as a tree-structured CRF.

The joint probability $P_\theta(\mathbf{y}|\mathbf{x})$ of a setting of the label variables $\mathbf{y} = [y_1, ..., y_L]$ conditioned on the observed feature variables $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_L]$ is given below. $Z_{\mathbf{W}}(\mathbf{x})$ is referred to as the *partition function* and is the normalization term of the probability distribution.

$$P_\theta(\mathbf{y}|\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{y}, \mathbf{x}))}{\sum_{\mathbf{y} \in \mathcal{Y}^L} \exp(-E_\theta(\mathbf{y}, \mathbf{x}))} \tag{5.1}$$

The parameters of a CRF can be estimated using either maximum likelihood estimation (MLE) or max-margin learning [70]. Importantly, the inference routines required to learn the parameters for a tree-structured CRF can be computed exactly in time linear in the number of variables in the model using the belief propagation algorithm [69]. Chain-structured CRFs are an important special case of tree-structured CRFs. The main weakness of chain-structured models is that they cannot model long-range dependencies. In the next section we describe the context free grammar conditional random field model which remedies this problem.

5.1.2 Context Free Grammars

A context free grammar (CFG) is defined by a set of production rules $\mathcal{R}$ that map from a set of non-terminal symbols $\mathcal{I}$ to strings of terminal and non-terminal symbols. We call the set of terminal symbols $\mathcal{V}$. Beginning with a special "start" symbol, these rules can be recursively applied until only terminal symbols remain. A sequence of such recursive applications produces a tree structure referred to as a parse tree. Given a grammar G, the set of strings of terminal symbols that can be produced in this way is referred to as the language defined by this G. Each production rule can be written as $A \rightarrow BC$ or $A \rightarrow a$ where capital letters denote non-terminal symbols and lower-case letters denote terminal symbols[1]. Formally, a grammar is defined as the tuple $G = (\mathcal{I}, \mathcal{V}, \mathcal{R}, \alpha)$ where $\mathcal{I}$ is the set of non-terminal symbols, $\mathcal{V}$ is the set of terminal symbols, $\mathcal{R}$ is the set of production rules, and $\alpha \in \mathcal{I}$ is the "start" symbol. For example,

[1] We assume a slightly relaxed form equivalent to Chomsky normal form.

consider a simple CFG with $\mathcal{I} = \{\gamma, A, B\}$, $\mathcal{V} = \{a, b\}$ and the production rules $\gamma \rightarrow AB$, $A \rightarrow aA$, $A \rightarrow a$, $B \rightarrow bB$, $B \rightarrow b$.[2] The recursive application of these rules produces strings that contain any number of a's followed by any number of b's.

The problem of parsing a string is the problem of identifying the parse tree used to generate the string. In the simple example described above, every string in the language has a unique valid parse, but this is not the case in general. In cases where multiple trees are possible, a weight can associate each rule with a large weight indicating that a rule is more likely to be observed. Then parsing becomes the problem of finding the parse tree with the maximum weight. Finally, a weighted CFG can be interpreted as defining an unnormalized distribution over parse trees given the input string where the maximum weighted parse tree is the most probable parse tree under this distribution. The conditional random field context free grammar (CRF-CFG) model presented in the next section further conditions weighted CFG on features of the input sequence.

5.1.3 The CRF-CFG Model

The conditional random field context free grammar (CRF-CFG) model is a CRF model that defines a distribution over parse trees given a grammar $G = (\mathcal{I}, \mathcal{V}, \mathcal{R}, \gamma)$ and a length L feature sequence $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_L]$ [66]. The set of all parse trees is represented by a set of binary random variables $\mathbf{Y} = \{y_{A,BC,i,j,l} \mid A \rightarrow BC \in \mathcal{R}, 1 \leq i \leq j < l \leq L\}$. $y_{A,BC,i,j,l}$ takes the value 1 if and only if the parse contains the sub-tree rooted at A covering positions i through l, A's left child is B covering positions i through j, and A's right child is C covering positions j through l. Otherwise, $y_{A,BC,i,j,l}$ takes the value 0.

As in all CRFs, the CRF-CFG model is defined by a set of feature functions. In this case, there are a set of K^r scalar feature functions for every production rule $r \in \mathcal{R}$: $f_k^r(y_{r,i,j,l}, i, j, l, \mathbf{x})$ for $k = 1, ..., K^r$. $f_k^r(y_{r,i,j,l}, i, j, l, \mathbf{x})$ takes the value 0 if $y_{r,i,j,l} = 0$

[2] For brevity, we will write production rules using "|" to denote multiple possible productions from the same non-terminal symbol. Using this notation, we can write the example grammar as $A \rightarrow aA|A$ and $B \rightarrow bB|B$.

otherwise it may be any function of the input sequence $\mathbf{x}$ and the indices of the production rule i, j, and l which leads to tremendous flexibility.

Finally, the probability of a parse tree $\mathbf{y}$ given an input sequence $\mathbf{x}$ is given by

$$P_\theta(\mathbf{y}, \mathbf{x}) \propto 1_{y \in \mathcal{Y}} \exp \left(\sum_{r \in \mathcal{R}} \sum_{i \leq j < l} \sum_{k=1}^{K^r} \theta_k^r f_k^r(y_{r,i,j,l}, i, j, l, \mathbf{x}) \right),$$

where 1 is the indicator function and $\mathcal{Y}$ is the set of all valid parse trees. While this model is substantially richer and more complex than the linear chain CRF, it has the important property that the maximum probability parse can still be computed in polynomial time given a setting of the weights θ. Specifically, the maximum probability parse can be computed in $\mathcal{O}(L^3)$ time using the inside-outside dynamic programming algorithm originally developed for the weighted CFG model [71].

5.1.4 Context-Free Grammars for segmentation

In the speech detection task, we are interested in jointly labeling the sequence of respiration cycles as corresponding to speech or not and segmenting the cycles into contiguous, non-overlapping segments of conversation and non-conversation activities. In this section, we use the CFG formalism to describe the set of all such segmentations and labellings of a sequence and then use the CRF-CFG model to induce a distribution over these segmentations given features available from the sensor data. The complete speech detection grammar is described below and an example parse is shown in Figure 5.1.

$$
\begin{aligned}
&\gamma \rightarrow \alpha \mid \beta \\
&\alpha \rightarrow C\beta \mid C \\
&\beta \rightarrow O\alpha \mid O \\
&O \rightarrow sO \mid qO \mid s \mid q \\
&C \rightarrow sC \mid qC \mid s \mid q
\end{aligned}
\tag{5.2}
$$

In this case, the set of terminals is $\mathcal{V} = \{s, q\}$ which indicate whether a respiration cycle contains speaking (s) or not (q). The symbols C and O are structural symbols that indicate whether we are currently in a conversation or other state respectively. The α and β symbols represent the roots of conversation and non-conversation segments respectively.

There are a few noteworthy structural characteristics of this grammar. First, speaking symbols are allowed in both conversation and non-conversation segments to allow for short duration speaking events outside of conversations. Second, the sequence labels and segmentation interact only through the weights on the terminal producing rules such as $O \rightarrow sO$, which means that the probability of a cycle label **conditioned on the segment it is in**, is independent of all other cycle labels in the segment. One possible extension to this model is to allow for Markov type interactions between labels within a segment, but we leave this for future work. It is further worth noting, that while the number of parameters in a CRF-CFG model scales linearly with the number of production rules in the grammar, the proposed grammar is relatively small and adds minimal model complexity relative to structure. Finally, because this model only provides a single layer of segmentation, marginal and MAP inference can be performed in $\mathcal{O}(L^2)$.

We estimate the parameters of this model using loss-augmented max-margin learning [70, 72]. For the augmentation loss, we use the Hamming loss between the true and predicted sequence labels.

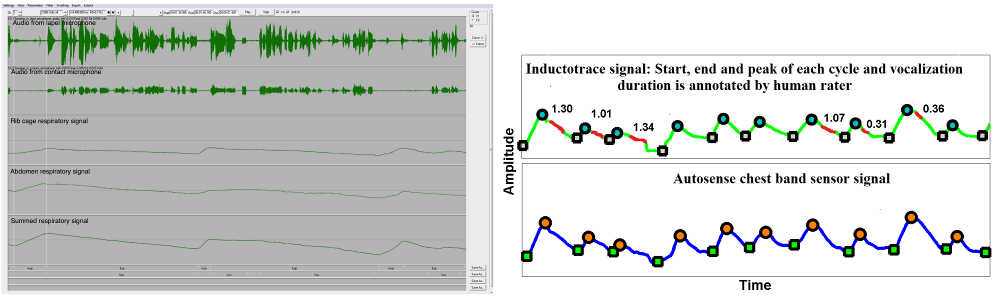


Fig. 5.2: (a) A snippet of AACT screen which was used to label respiration data from inductotrace band. The screen contains five different time synchronized signals. The video is also synchronized. From the top, the signals are from — headset microphone, contact microphone, ribcage inductotrace band, abdomen inductotrace band and summed ribcage and abdomen signal. All the signals were utilized to label each respiration cycle as well as the duration of vocalization occurring within each cycle. (b) The top panel shows the ribcage inductotrace signal with the annotated labels, cycle start and end position, peak position etc. The vocalization location is indicated by the red color in the signal and duration of vocalization is written on top of it within the speech cycles. The bottom signal is the AutoSense chest band respiration signal, which is synchronized with the inductotrace signal. The ground truth annotation of the inductotrace signal serves as a reference to label AutoSense signal.

5.2 Data Labeling

For development, training, and testing of the conversation model, we need to label each respiration cycle as speech and non-speech and as well as the conversation episodes.

5.2.1 Lab Data Labeling

To get fine granularity labeling of the data collected in lab, we utilized the information from headset microphones, throat microphones and video to precisely mark the speech status of each cycle. We trained four coders to label the Inductotrace signal using the Action Analysis Coding and Training software (AACT; Delgado and Milenkovic, 2017), which gave the coders access to the time-synchronized audio and video recordings as well as the respiratory signals. This multi-modal analysis environment allowed both rib cage and abdominal signals as well as their sum to be inspected in synchrony with audio to certify when speech related exhalation was occurring, and often when non-speech exhalations and inhalations occurred as well. Furthermore, synchronized video recordings of the lab conversations also allowed coders to observe when respiratory signals were affected by motion. A snippet of AACT screen is shown in

Figure 5.2a. All displays and sound signals were considered when marking the onsets and offsets of inspiration, expiration, and utterances produced by each conversation partner. After a training period, coders labeled respiratory and audio data for the same four sessions. Inter-rater reliability was assessed: all reliability kappas were significant and greater than 0.8. Coders were then assigned to label individual sessions for the rest of the dataset. This training was conducted by a speech scientist with 30+ years of experience examining conversational speech and 15+ years of experience examining respiratory kinematics during conversation.

Next, AutoSense chest band sensor data, which was worn simultaneously with the Inductotrace bands, was labeled. As these two systems are independent, participants were told to take three quick breaths before each task, afterwards, to sync the signals from both types of bands. First, we aligned the Inductotrace signal and the AutoSense respiration signal as shown in Figure 5.2b. The top panel in this figure shows the Inductotrace sum signal plotted with manually labeled start and end time for each cycle. The manual marking of the Inductotrace signal serves as a reference to label the AutoSense chest band signal.

5.2.2 Field Data Labeling

In the field, we collected respiration and audio data from 38 participants to evaluate the lab-to-field generalizability of the proposed *rConverse* model. On average, we collected 12 hours of audio data/day from each participant (sampling rate 16 KHz). Among the 38 participants, audio data was lost from 5 participants due to file corruption. Additionally, respiration data from 1 participant was of poor quality. We were able to analyze data from the remaining 32 participants.

Labeling field conversation data from the audio stream presented several challenges. First, since our dataset contains around half million respiration cycles and each cycle varies in fine-grained time-granularity (milliseconds to seconds), it is not practical to annotate each respiration cycle as containing speech or not. Therefore, we focus on marking start and end of conversations. To label the time-series for conversation, we used audio from LENA as an indicator of the presence of conversation and corrected false positives generated by LENA using the raw audio signal.

Second, there is a time drift (up to 1 minute) between the audio device and the respiration sensor and it is difficult to build in explicit synchronization actions as in the lab due to intermittent data loss from exercise of privacy control by the participants. Third, the large volume of audio data (over 200 hours) requires extensive time and effort for human raters to annotate, especially to mark each turn-taking in the conversation. Rapid turn-taking inside the conversation aggravates this challenge. Fourth, it is difficult to mark the start and end boundaries of a conversation episode when both conversing parties are silent (e.g., thinking) in a conversation.

Therefore, when annotating the beginnings and endings of conversations, we assumed that a pause of greater than one minute constituted the start of a new conversation. We labeled 254 hours of audio data, on average 8 hours per participant.

5.3 Feature Extraction and Selection

In the previous section, it was assumed that input signal had been discretized into a sequence of respiration cycles, and that features had been extracted from each cycle to form a feature sequence $\mathbf{x}$. In this section, we present the feature extraction methods used to derive features from each respiration cycle. Further, we present a series of feature selection strategies to minimize covariate shift between the lab and field domains.

5.3.1 Feature Extraction and Normalization

We compute the duration, amplitude, area and several other features for the inspiration, expiration and respiration segments of each cycle as depicted in Figure 5.3

Duration features. These features measure the duration for the segments of each cycle: inspiration, expiration and respiration phase. *Inspiration duration (T_I).* The process of actively drawing air into the lungs is defined as inspiration. Inspiration time is measured as the time between the beginning and end of inspiration phase as indicated by an upward

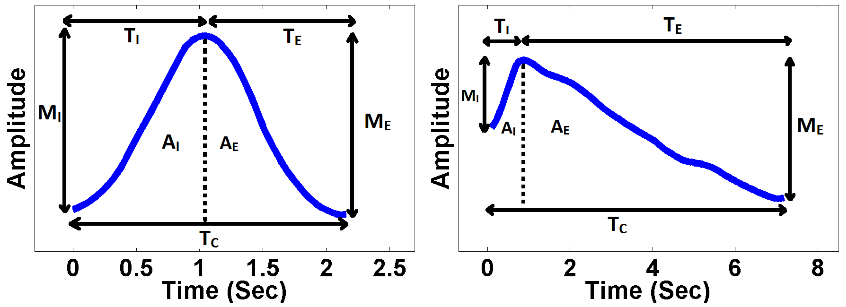


Fig. 5.3: Features of interest in a theoretical quiet and speech cycle. T_I=Inspiration duration, T_E= Expiration duration, T_C= Respiration Cycle duration, M_I= Inspiration magnitude, M_E= Expiration magnitude, A_I= Inspiration area, A_E= Expiration area.

slope from left to right in the respiration signal. *Expiration duration (T_E).* Expiration is normally a passive process where air leaves the lungs. Expiration time is defined as the time from the end of inspiration to the beginning of inspiration of the next cycle. *Cycle duration (T_C).* The time it takes to complete a breathing cycle, calculated as $(T_I + T_E)$.

Magnitude features. The amplitude of a cycle varies for different activities, postures and conversation shown in Figure 5.3.

Inspiration magnitude (M_I). is defined as the vertical distance between the maximum and minimum of each inspiration phase. *Expiration magnitude (M_E)* is defined as the vertical distance between the maximum and minimum of each expiration phase. *Magnitude Difference* is defined as the difference between inspiration magnitude and expiration magnitude. During quiet breathing, difference of magnitude is small compared to speech breathing cycles. *Stretch* is defined as the vertical distance between the maximum and minimum point within a cycle.

Area features. The change in air volume during the inhalation and exhalation stages is reflected with these features. *Inspiration area (A_I)* is defined as the area under the curve between the beginning of inspiration to the end of inspiration phase for each cycle. *Expiration area (A_E)* is defined as the area under the curve from the end of inspiration phase of a cycle to the start of inspiration phase of the next cycle. Mean inspiratory flow rate (A_I+A_E)/T_I or drive is defined as a ratio of cycle area to inspiration duration.

Flow rate features. We measure the instantaneous flow rate for both inhaling and exhaling phases. *Inspiratory Flow rate (V_I)* is described as the time requires to inhale the amount of air during the inspiration phase. *Expiratory Flow Rate (V_E)* is described as the time requires to exhale the amount of air during the exhalation phase.

Ratio features. We use several ratio features. Ratio of inspiration to expiration duration, area and flow rate is presented as IE_T, IE_A, IE_V respectively. Fractional inspiratory time or effective timing ratio is defined as a ratio of T_I to T_{tot}.

Power in Frequency Bands. We calculate the spectral power in several frequency bands, 0.01-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz, 0.6-0.8 Hz and 0.8-1 Hz. We further measure the LF to HF spectral power (LF/HF) ratio where spectral power is calculated in the low frequency band between 0.05 Hz and 0.15 Hz (LF) and high frequency band from 0.15 Hz to 0.5 Hz (HF).

Breath-by-Breath Correlation. From the lab data, we see that the correlation between two neighboring cycles is high when both of them are non-speaking cycles. Otherwise, correlation is mostly low when adjacent cycles are either speaking-speaking or speaking-quiet. Thus we measure the cross-correlation of a cycle with its previous cycle and with the next cycle and using them as features.

Other Features. We also calculate the energy, entropy and skewness of each cycles.

Additionally, we apply a simple non-linear transformation to these features by finding five equal sized percentile bins for each feature and compute the distance from the center of each percentile bin to the input feature value. Finally, we z-normalize all feature values.

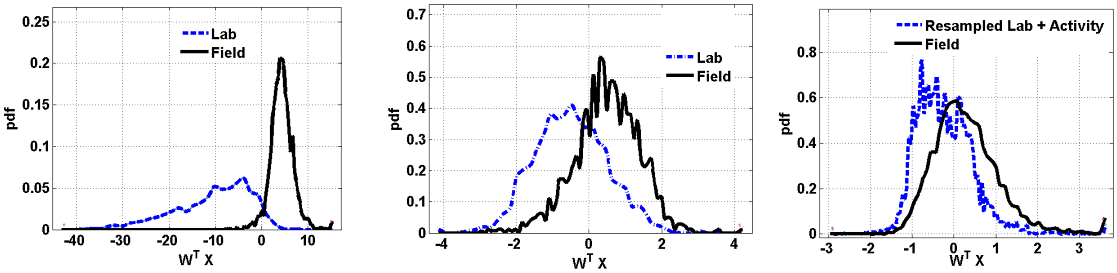


Fig. 5.4: (a) Covariate shift between lab and field feature distributions is $95.6 \pm 0.1\%$ with all features. (b) After applying feature selection method, covariate shift is reduced to $76.1 \pm 0.4\%$. (c) Adding activity data with the resampled lab data has further reduced the covariate shift to $63.4 \pm 0.02\%$.

5.3.2 Feature Selection to Reduce Covariate Shift for Lab to Field Generalization

Covariate shift refers to a significant difference between the lab and field feature distributions. This difference can result in decreased generalization performance of models trained on lab data to a field setting. While several methods exist to address covariate shift in the independent classification setting (e.g. [73]), these methods do not generalize to the structured prediction setting where objective functions do not decompose over individual variables. Instead, we propose a feature selection method to select cycle level features that balance class discrimination against domain discrimination. We did this by training the importance weighted logistic regression model and selected 20 features with the highest absolute weights in the resulting model.

Specifically, [73] used the following importance weighted logistic regression model:

$$\underset{x}{\operatorname{argmin}} \sum_{i=1}^{N} \delta(y_i, x_i) \log(1 + \exp(-y_i(w^T x_i + w_0))) + \lambda ||w||^2 \qquad (5.3)$$

where λ controls regularization strength and the importance weights $\delta(y_i, x_i)$ are given by a second, unweighted, logistic regression model trained to discriminate the lab

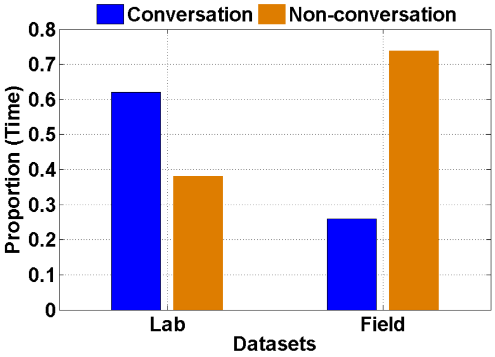


Fig. 5.5: Proportion of time spent on conversation and non-conversation tasks in lab and field respectively.

and field data. Let $Q(x_i)$ be the output from a logistic regression model trained to discriminate the lab data from the field data. Then,

$$\delta_i(y_i, x_i) = 1/(1 - Q(x_i)) \tag{5.4}$$

The regularization parameter was tuned over a logarithmic grid using leave-one-subject-out cross-validation on the training set.

We tested the effectiveness of this method by training a logistic regression model to discriminate the lab and field datasets and evaluating the accuracy of this model. Using the raw features, a logistic regression model can discriminate the lab and field data with an accuracy of 95.6%. After applying feature selection, this accuracy goes down to 76.1% indicating that the covariate shift was substantially reduced. To demonstrate this visually, we took the feature weights learned by a logistic regression model trained to discriminate lab and field data and plotted the distribution of weighted sums of feature vectors. Figure 5.4a shows this distribution for all features and Figure 5.4b shows this distribution for selected features.

5.3.3 Resampled Lab Data - Handling Prior Probability Shift

The way participants spent time within conversations in lab environment may not be representative of their behavior in the field. Figure 5.5 shows the amount of time participants spend in conversation activities in the lab and field. A smaller fraction of time is spent in conversation in the field (about 26%, which is about 3 hours out of 12 hours), while the training data collection protocol significantly over-represents the proportion of time spent in conversation (about 62%) in lab. To address the issue of prior probability shift, the non-conversation data in lab is resampled to match with the conversation distribution in field. On average, 3 hours of conversation per day in the collected dataset may seem high. Several factors can help explain the large quantity of conversation in field: 1) cohabiting couples were recruited to maximize conversational interaction; 2) most of the couples conducted their field recordings on weekends when they were spending most of their time together; 3) these participants were aware that we are seeking conversational interaction so they may have produced even more than typical (few participants mentioned this in their exit interviews).

5.3.4 Conversation in Presence of Activity

Data collected in lab typically exercises a very limited number of contexts relative to field environment. Physical activity is a common phenomenon which is absent in data collected in lab settings. This factor can lead to significant differences in between lab and field feature distributions [73], which can be accounted for by covariate shifts.

To see the effect of activity, the training- Field data collected in presence of physical activity (i.e., walking), is combined with the resampled lab data. The activity enriched data with resampled lab data adds significant variability and the covariate shift of the resultant dataset reduces to 63.3% (Figure 5.4c).

5.4 Empirical Protocols

In this section we describe the details of data preparation, training protocols, and evaluation metrics.

5.4.1 Tasks

There are two tasks of interest in the speech detection problem: Cycle level speech labeling (**Task 1**) and conversation episode detection (**Task 2**). Cycle level speech labeling entails labeling each individual respiration cycle as corresponding to speech or not. Conversation episode detection entails segmenting each sequence of respiration cycles into contiguous periods of conversation and non-conversation activities.

5.4.2 Data Preparation

As described above, labeled respiration data was collected from 12 subjects in the lab. We dropped the data from 1 participant due to poor data quality. In order to create a single, long session for each subject, we concatenated the data for each subject in a random order. The resulting dataset contains 11 separate respiration waveforms which we process using the feature extraction methods described above to create a training set with 11 unique labeled feature sequences.

5.4.3 Baseline Models and Hyper-parameter Selection

We compare our the CRF-CFG model against two common baselines: Logistic Regression (LR) and a linear-chain conditional random field model (CRF-LC). All models are trained using max-margin learning and all models include ℓ_2 regularization on the parameters [70]. For all models, the regularization strength parameter, λ, was tuned over a logarithmic grid, $\{10^{-1}, 10^{0}, ..., 10^{5}\}$, using leave-one-subject-out cross-validation on the training set. We selected the value of λ that maximized cycle level accuracy averaged across all folds and then trained a final model on all of the training data using this λ value.

5.4.4 Evaluation Metrics

Evaluation on Lab Data: We assessed the performance of all models on Task 1 (cycle labeling) using standard classification metrics such as accuracy, precision, recall, and F1 score. To evaluate conversation episode detection performance (Task 2), we compare the predicted segmentation with the true segmentation by projecting each segmentation onto the input sequence and calculating the performance metrics on the resulting binary sequences.

Evaluation on Field Data: We compare the performance of our model for detecting conversation with that from audio data by the speech classifier of the LENA foundation. To account for the time drift of up to one minute between respiration

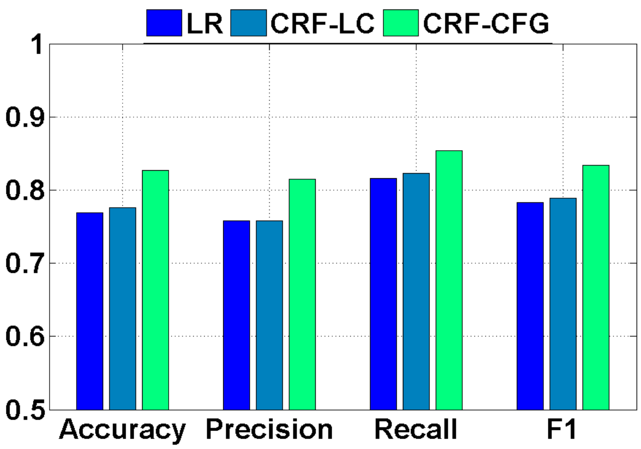


Fig. 5.6: Cycle labeling performance of different models on training data. LR: Logistic Regression, LC-CRF: Linear Chain CRF, CRF-CFG: CRF with Context Free Grammar.

Table 5.1: Confusion Matrix for cycle labeling on training lab data with CRF-CFG model using leave-one-subject-out validation; Cycle labeling Accuracy=82.7%, Precision=81.5%, Recall=85.4%, F1=0.83, and False Positive Rate=20.1%.

		Classified by Model		
		Speech	non-speech	Total
Actual	Speech	833 (85.4%)	142 (14.6%)	975
	Non-speech	189 (20.1%)	753 (79.9%)	942
	Total	1022	895	1917

time-series and the audio time-series, we segment both the time-series into one minute windows. If both ground truth annotated conversation and model detected conversation is present in any one minute window, we consider that window to be a true positive (TP). Similarly, we calculate true negatives (TN), false positives (FP), and false negatives (FN). Finally, we compute the accuracy, precision, recall, F1-score, and false positive rates (FPR).

5.5 Results

5.5.1 Experiment 1: Comparison Against Baseline Models

To evaluate the CRF-CFG model against the classification baselines, we performed a leave-one-subject-out evaluation using the lab data for which we have detailed respiration cycle level labels. The leave-one-subject-out prediction results for Task 1 (cycle labeling) for each model averaged across subjects is shown in Figure 5.6.

The accuracy, precision, recall and F1-score of CRF-CFG model for cycle labeling using lab data is 82.7%, 81.5%, 85.4%, and 0.83, respectively. Table 5.1 contains the confusion matrix of the cross-subject validation for CRF-CFG model. Whereas, accuracy of LR and CRF-LC models are 76.9% and 77.6% respectively. The fact that improvement of CRF-LC over LR indicates that there are reasonable correlations between adjacent respiration cycles; however, the CRF-CFG model improves further over CRF-LC, indicating that the Markov assumption may not hold in this context. That is, a cycle labeling benefits from knowing whether it is in a conversation and not just what its neighbors labels are. The accuracy, precision, recall, and F1-score of CRF-CFG model for Task 2 (episode detection) on the lab data is 95.9%, 91.28%, 96.0%, and 0.94 respectively.

5.5.2 Experiment 2: Conversation Detection in the Field

In order to test the various feature selection and data augmentation methods proposed in Section 5.3 we perform an ablation study, adding in each proposed augmentation one at a time. Then, using all augmentation methods, we compare the performance of the CRF-CFG model against both human annotated ground truth and LENA model on the task of conversation episode detection (Task 2).

Performance using lab data trained on all features

The lab data model trained with all features can identify the conversation episodes in field with an accuracy of 52.03% (Figure 5.7). The precision and recall is 43.02% and 97.02%, respectively.

Performance using lab data trained on selected features

Deploying the lab model trained with selected features that reduce covariate shift from lab to field data, the conversation episode detection accuracy in field is 60.8%, precision is 58.6% and recall is 98.01% (Figure 5.7) while the false positive rate is 87.5%. Thus, feature selection method has improved the accuracy by 8.8% in field. The F1 score is 0.72 for this model.

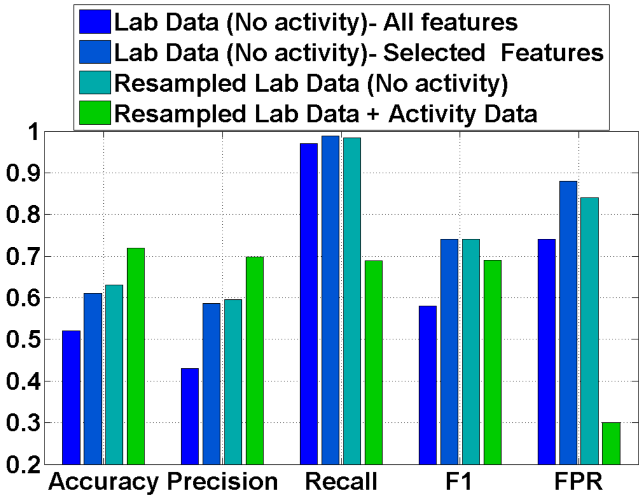


Fig. 5.7: Model performance comparison to detect conversation episodes on field data. First bar indicates the performance of model trained on lab data with all features. Second bar indicates the performance of model trained on lab data with selected features after covariate shift reduction. Third bar indicates the performance of model trained on resampled lab data with selected features. Fourth bar indicates the performance of model trained on activity enriched resampled lab data with selected features. The fourth model shows better performance (higher accuracy, lower false positive rate) over other models to detect conversation episodes on field data.

However, in comparison with the performance with lab data, conversation episode detection accuracy drops from 95.9% (see Figure 5.6) to 58.6% on the field data using this model. Still there is a large gap of performance between lab and field.

Performance using resampled lab data trained on selected features

The resampled lab data model can identify the conversation episodes with an accuracy of 62.5% in field. The precision and recall are 59.6% and 98.4%, respectively. The false positive rate has been reduced to 84.4%. Thus, data resampling has improved the accuracy by 2% and reduced the FPR by 3.1% in field.

Performance using resampled lab data and activity data trained on selected features

The accuracy of the model using activity enriched data with resampled lab data is 71.7% and false positive rate is 30.03% in the field. The precision, recall and F1 score is 69.8%, 68.9% and 0.69. Thus accuracy is increased by 8.5% and FPR is reduced by 54.4%.

Table 5.2: Performance comparison between CRF-CFG model and LENA model that includes state-of-the-art algorithm to detect human speech on audio data.

Models	Accuracy (%)	Precision(%)	Recall(%)	F1-score	FPR(%)
CRG-CFG model	71.7	69.8	68.9	0.69	30.0
LENA model	71.9	73.4	66.5	0.69	26.6

Performance Comparison with Audio-based Conversation Model (LENA)

We compare the model performance with audio recorder (LENA) that also detects human speech and distinguishes human vocalization from electronic sounds (e.g., TV). Final model (Resampled lab with activity included) predictions and LENA predictions are compared with human annotated ground-truth on field data for performance comparison.

Accuracy to detect conversation by CRF-CFG model and the audio based model is similar (around 72% as shown in Table 5.2). We note that the audio recording used in this study capture high quality audio and it was not subject to occlusion, unlike audio capture on smartphones that may subsample or be occluded due to being in pocket or purse.

5.6 Related Work

Conversation modeling, based on acoustic data captured with smartphone microphones [74] or with wearable microphones [75] has been a fertile area of research for decades. Advanced research has been done in audio sensing not only to distinguish conversation episodes from ambient sound or music [74], but also to model various characteristics of a conversation, including turn-taking behavior [76], group size estimation [77], and speaker identification [78, 79]. Furthermore, acoustic researchers have also addressed speakers' emotions [80] and stress levels [81]; and developed socio-therapy applications [76] for children with autism.

In this work, we explore the potential for detecting conversations from respiratory measurements that can be useful when respiration data is collected in context of health related research (e.g., smoking cessation, asthma) or self-monitoring (e.g., biofeedback). A model for detecting conversations from respiration can be applied to such data collected to infer conversation episodes which play an important role in stress management, smoking lapse, depression, etc. An advantage of respiration based models is that they are more specific to the speaker and less privacy sensitive [17].

Respiration-based conversation modeling is, however, underexplored, perhaps due to the lack of reliable respiration signals collected in field setting. The emergence of connected wearable and contactless smart technologies have made it feasible to capture respiration data reliably and comfortably in everyday life.

Two common methods for continuous respiration rate monitoring in clinical settings are impedance pneumography and capnography, which require the use of a nasal probe [82]. These methods are expensive and intrusive, and therefore not useful for daily use. In order to minimize the discomfort, researchers developed pressure-based bed sensors [83, 84] for long-term and continuous respiration monitoring while users are lying down.

Several methods have been developed to measure respiration continuously in indoor settings (e.g., home, office) while users are mobile and not confined to a bed or any furniture [85, 86]. For example, Adib et al., developed a radar based, contactless Vital-Radio [85] to track respiration rhythm while the user is 8m away from the sensor, co-located with multiple other subjects, regardless of whether she is sleeping, watching TV, or typing on her laptop. In order to make the contactless respiration measurement infrastructureless and cost-effective, researchers have developed several methods based on commodity sensors, such as camera [87] and WiFi [88]. The basic idea of such systems is to measure displacements of the chest of human subjects during breathing. These methods can capture breathing depth, location, orientation, and respiration rate from a distance, making them viable for long-term respiration monitoring in indoor settings.

Wearable wireless sensors make the respiration signal continuously available in mobile settings. Commercial releases and research prototypes of wearable chestband [42, 89] and smart garments [90] have been developed to continuously measure respiration 24/7. They are either piezoelectric-based or inductance-based sensors to reliably capture respiration rhythms in natural settings. These straps are sometimes reported to be uncomfortable for the wearers.

Recently developed wearable devices enable respiration data to be captured more easily and comfortably in our daily lives. For example, commercially available accelerometer-based small devices (clipped with clothing) such as Spire or Prana [91] help users capture breathing information and visualize on a smartphone to aid in breathing regulation. The Philips Health Watch [92], an FDA[3] approved commercial product, makes respiration rate accessible from a comfortable, easy-to-wear smartwatch. A popular consumer device, Apple Watch, introduced the Breathe app in WatchOS3, and the Fitbit Charge 2 added a guided breathing tool called ‘relax’. The increasing number of devices and associated smartphone apps that feature respiration data capture and usage demonstrates that respiration data is becoming more accessible and can be collected unobtrusively in user’s natural environment.

We note, however, that capturing accurate respiration waveforms today still requires wearing a belt around the chest that may not be comfortable for long-term wearing. But, despite such constraints, chest-worn respiration sensors are being used to collect over 10,000 person days (over 100,000 hours) of data from over 1,000 participants at five sites across the US[4]. We have used a similar chestband sensor to collect reliable respiration data continuously in wide variety of field settings. Although our model has been developed on waveforms collected from a respiration belt worn around the chest in natural settings, they can be suitably adapted for other emerging respiration sensing modalties.

The closest work to ours is mConverse [17] that captured respiratory measurements from a chestband sensor to infer conversation events. However, as described in Section 1, this early model could only operate on 30-second windows. For training and validation, each 30-second window of respiration data was labeled based on a

[3] US Food and Drug Administration. https://www.fda.gov/

[4] See https://md2k.org/studies for a list of these deployments.

majority of speech or non-speech duration within the window as marked by a human observer. Consequently, this work either overestimated or underestimated speech and non-speech durations in a conversation.

Because respiratory cycle is a unit of speech breathing, cycle-based classification is the finest granularity for speech modeling from respiration data. Each respiration cycle dynamically varies in duration. Hence, cycle-based dynamic windowing is an appropriate approach for the respiration based speech modeling as presented in the current model. To generate labels, speech/non-speech cycles were carefully marked based on audio, video, and hospital grade respiratory inductive plethysmograph bands with synchronized channels in the lab setting and by using audio processing from LENA and confirmation from human raters in the field. Moreover, we present a CRF-CFG model which both classifies cycles into speech and non-speech, and further segments cycles into conversation episodes. This model is evaluated against gold-standard acoustic data collected in the natural environment.

On the modeling side, segmentation based models have been successfully used for a wide variety of activity recognition tasks [67, 93, 94, 95]. For example, Tang et al., [93] and Sung et al., [94] use conditional segmentation models for labeling and segmenting activities in video streams. Adams et al., [95] use a hierarchical segmentation model to label and segment smoking activities in respiration data. Most closely related to our approach, [67] use a CRF-CFG model for ECG morphology extraction. In this work, we develop a grammar for a CRF-CFG model to detect conversation episodes, which has different characteristics than prior works on ECG morphology or smoking, demonstrating wider applicability for the CRF-CFG approach.

5.7 Conclusions

This work presented a conversation episode identification model from respiration signals by classifying each breathing cycle into speech and non-speech. Audio captured in the field is used to validate the models. For these classification, we describe several intuitive time domain features from respiration which are different from the traditional features. These features can be of interest in detection of other daily behaviors such as laughing, singing, eating, drinking, etc. Previously, detection of momentary behaviors from respiration data collected in the field setting hadn't been realized. This work can contribute a comprehensive approach to processing of respiration data in the field setting and lead to momentary detection of various daily behaviors from respiration data and enhance the growing utility of respiration sensing.

Numerous real-life applications can be pursued using this method. First, turn taking can be observed in group conversations and analyzed to improve turn taking in meetings. The speaking turn has been defined as an uninterrupted series of speech segments from a single speaker. Second, back-channels can also be studied. They are unplanned, small vocalization, produced by the listener to give short feedback to the speaker. Some studies say, the cycle that contains back-channel is not a speech breath, because he/she has no intention to take floor. In future, we would analyze the effects of back-channels using our models. Third, using this method in real-life can help enhance the scientific studies of social interactions and help individuals reflect upon and improve their social interactions. Its usage together with processing of audio data captured on the microphone can help further characterize the content of conversation.

Chapter 6

Stressful Social Interaction Detection

To develop a model to distinguish stressful interaction or conversation from other daily stressors such work and commuting related stress, we designed the field study to collect interaction data in natural environment that is describe in Chapter 3. To handle the challenge of collecting labeled stress events, we designed a Day Reconstruction based visualization that helps participants to recall the stress events from the stress likelihood time-series. In this chapter, we describe how we extract distinguishing patterns from the stress event along with the wrist motion sensor data to detect stressful conversations. First, we describe our proposed method to identify cyclical pattern in a stress event followed by wrist motion patterns. Second, we describe feature extraction from the stress and wrist motion data to train a machine learning model for detecting stressful conversations. Finally, we evaluate our models and discuss implications of the models.

6.1 Key Ideas and Overall Approach

Input to the model is a continuous stress likelihood time-series, with one data point every few seconds. The time-series is annotated with the start and end times of stressful events. Assuming that each of the stressful events is attributed to one source of stress, the goal is to determine whether each event is due to stressful conversations or interaction.

6.1.1 Key Ideas

Our model development is based on three key ideas. First, we notice that stress time-series signal during stress event is episodic and often periodic, exhibiting peaks and troughs that can be used to naturally segment the data. Second, we identify several novel features from these cycles. Third, we observe that the pattern of hand gestures when stress occurs due to personal interactions is distinct in nature, as compared to when stress is due to work or commute. With increasing adoption of smartwatches and fitness trackers, it is increasingly feasible to capture hand movement patterns continuously. We also note that with recent improvements in optical sensing in smartwatches, stress may also be detected from smartwatches [40, 41], making for a complete sensor suite on which our model can be implemented.

6.1.2 Overall Approach

Our model development consists of the following major steps.

1. **Cyclical Pattern Identification:** Cyclical patterns in stress time-series are different than that in regular physiological signals such as respiration cycles. Respiration cycle is well defined by inhalation and exhalation phases associated with each breath but stress cycles do not have any such naturally defined phases. Therefore, existing methods for detecting peaks and troughs are not directly applicable to stress cyclical pattern identification. We propose a new method to detect cycles in the stress likelihood timeseries and characterize portions of interest from which distinguishable features can be computed.

2. **Intra-cycle Feature Extraction:** Unlike respiration, there is no natural phenomenon of inspiratory and expiratory time. Therefore, we need to discover new features that can characterize and interpret a stress cycle.

3. **Inter-cycle Feature Extraction:** To capture any patterns that span multiple stress cycles within a stress event, potentially covering all stress cycles within a stressful event, we compute features spanning multiple stress cycles.

4. **Wrist Motion Features:** Wrist motion sensors data have been researched extensively for activity and posture detection. We compute these features within each stress cycle to determine their utility in capturing the distinct signatures of hand gestures observed during stress events, to improve the accuracy of detecting stressful conversations.

6.2 Observation and Characterization of Stress Likelihoods Within Stress Events

We expect the physiological response during a stress event to exhibit a cyclical pattern. To investigate whether we observe a cyclical pattern during stressful

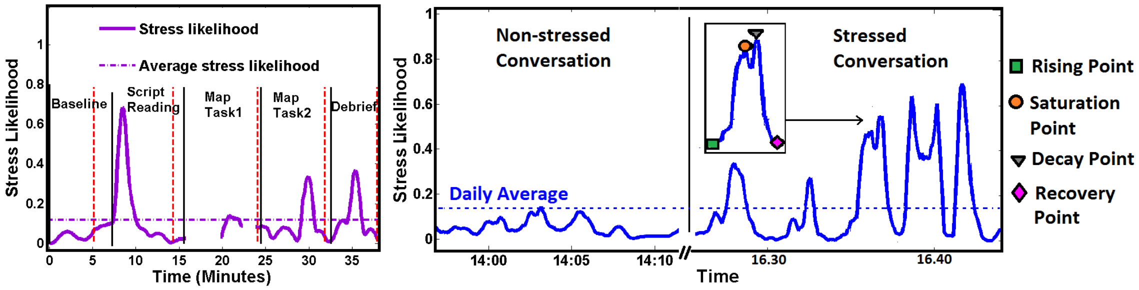


Fig. 6.1: (a) Cyclical pattern of stress likelihood observed in lab data. The vertical solid black and dotted red lines depict start and end times of each tasks. Stress cycles are visible during scripted reading, Map Task2 and debrief session. In between 15 and 25 minutes, we observe a portion of missing data. (b) Variation in stress likelihood pattern during stressful and non-stressful conversations in field. The horizontal blue dotted line shows the daily average of stress likelihood. During stressed conversation, we observe numerous high arousal stress cycles.

conversations, we analyzed the physiological data collected during the lab study, where stressful conversations took place and the physiological data was mostly free of any confounders. As described in Section 3.2.3, we apply the cStress model on physiological data to convert the physiological sensor data into stress likelihoods (in sliding minute-windows, starting every 5 seconds) as shown in Figure 3.6. We also mark the start and end of stress events.

We observe that the cyclical patterns previously observed in the physiological response during stress tasks (due to the interplay between SNS and PNS) is also observed in the stress likelihood time series within a stress event. The activation of SNS results in the elevation of physiological arousal which is captured by an increase in the stress likelihood produced by the cStress model. We define this point as stress ‘Rising point’ where stress arousal starts to elevate from its pre-stress condition, i.e., an average of daily stress likelihood as shown in Figure 3.6. Concurrently, each time SNS activates, the PNS gets activated as well to provide the corresponding counterbalance so as to keep the physiology in homeostasis balance. When the influence of PNS exceeds that of SNS, then it reaches a ‘Saturation point’, after which the stress arousal starts to decay, indicated by the ‘Decay point’ when the effect of stressor starts to mitigate. Finally, it reaches the pre-stress value or below the daily average of stress likelihood denoted by the ‘Recovery point'. We define this structure as a 'stress cycle', where stress cycle begins at a 'Rising point' and ends at a 'Recovery point'.

The cycle repeats if the current episode continues to produce new stress triggers (e.g., conflicting words spoken by the conversation partner). A stress event may consist of one or more stress cycles depending on the repetition of stress triggers within a stress event. In Figure 3.6, the depicted stress event consists of three stress cycles.

We illustrate the cyclical patterns in the stress likelihood time-series data during lab tasks in Figure 6.1a. It shows that stress likelihood was low during the baseline session. Stress likelihood rises during the scripted dialogue task as the individual was waiting for his/her turns, and they were focusing on their performance to make the dialogues look more natural. As the nature of the map tasks tended to induce some informational conflict between partners, we see high arousal stress cycles during Map Task 2 and during the debrief session when they were trying to resolve their conflict. Stress arousal in Map Task 1 is not as visible due to missing data.

We observe a similar cyclical pattern during stress events in the field data. Figure 6.1b depicts the stress arousal of a participant in the field during two separate conversational interactions at two different times. The first interaction (left portion) was a non-stressful conversation , where stress likelihood remains below the daily average. The second interaction (right portion) presents a stressful conversation, where we observe several stress cycles that rise above the daily average of stress likelihood. This particular stressful conversation consists of five stress cycles. In the following section, we describe how we identify stress cycles automatically from the stress time-series data.

6.3 Stress Cycle Identification Algorithm

As we assume stress cycle is the smallest unit inside a stressful event, this cycle is used to segment the day long stress time-series. To identify each stress cycle with all four interesting points — stress rising, saturation, decay and recovery point, we propose a moving average based method. For that, we build upon the cycle identification model used

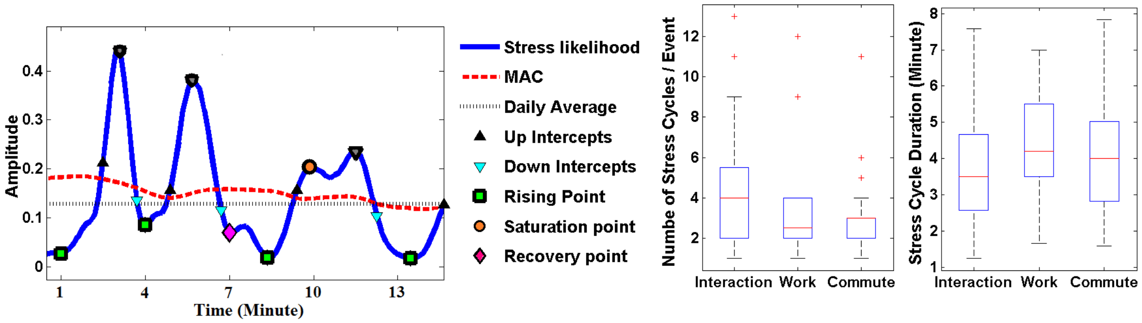


Fig. 6.2: (a) Moving average based cycle identification method to identify all four interesting points- (1) stress rising point; (2) stress saturation point; (3) stress decay point; and (4) stress recovery points are detected. (b) Summary of detected stress cycles for interaction, work and commute related stress. Median numbers of stress cycles/event are 4, 2.5 and 3 minutes for interaction, work and commute, respectively, where number of cycles/event is significantly higher for interaction related stress compared to other two stressors. Median stress cycle duration's are 3.5, 4.2 and 4.0 minutes for interaction, work and commute related stress, respectively. Significant difference is found is found between interaction and interaction related stress cycle duration.

to detect other physiological phenomena such as breathing cycle [18]. The method developed for breathing cycle identification will not be directly applicable for stress cycle identification. Breathing signal follows some specific structure with inspiration and expiration phases driven by the physiological phenomenon. On the other hand, stress cycle is guided mostly by the stressful situation and may not have any specific rules. Therefore, we have modified the algorithm to identify stress saturation and decay point.

First, we smooth the stress likelihood time-series using a 15 seconds moving average to remove spikes. Then another moving average centerline (MAC) curve is computed using a moving average of 2 minutes. The MAC appears as a center line (shown as red dotted line in Figure 6.2a) that intercepts each stress cycle twice, once in the rising trend and then in the falling trend. Next, we identify the up and down intercepts where the MAC curve intercepts the rising and falling branch of smoothed stress time-series respectively. The 'rising point' is the rightmost local minimum that lies below daily average found between consecutive down and up intercept pair. From this point, signal rises monotonically towards saturation point.

The ‘saturation point’ lies between the up intercept and the following peak of that cycle where rising trend reaches the peak. This point is the leftmost local maximum and must be above the up intercept and MAC curve line.

‘Decay point’ lies between the saturation point and the following down intercept when signal starts monotonically decreasing. This point is detected as the rightmost local maximum and must lie above the following down intercept and the MAC line. The falling trend reaches to ‘recovery point’ when it decreases to first local minimum value below daily average of stress likelihood.

Performance Evaluation: We annotated total 160 stress cycles from several stress events including interaction, work and commute. We use the following metrics to evaluate the performance of the algorithm — percentage of actual cycle detected, percentage of extra or spurious cycle found, and error in cycle duration due to mislocated rising and/or recovery point. Two coders independently labelled all the interesting points of a cycle. Inter-rater reliability was around 0.9 between the coders. The algorithm could identify 96% cycles accurately and detected 3% cycles as extra or spurious. The mean absolute error in identifying cycle start or rising point is 8.86 seconds. The mean absolute error in identifying cycle end or recovery point is 6.9 seconds. Therefore, mean error in cycle duration is 8.16 seconds. The rationale for calculating error in cycle duration is even if a rising point and a recovery point are identified correctly, their respective temporal position in the signal may introduce error in the resultant duration.

After applying this algorithm on all stressful events, we find the average number of stress cycles are 4.42, 3.6, and 2.9 for stressful interaction, work and commute, respectively as depicted in Figure 6.2b. Number of cycles per event for interaction related stress is significantly higher compared to both work and commute related stress cycles at 5% significance level (using t-test). But, no significance difference is found between work and commute related number of stress cycles. Average stress cycle duration is 3.7, 4.8, and 4.02 minutes for interaction, work and commute, respectively depicted in Figure 6.2b

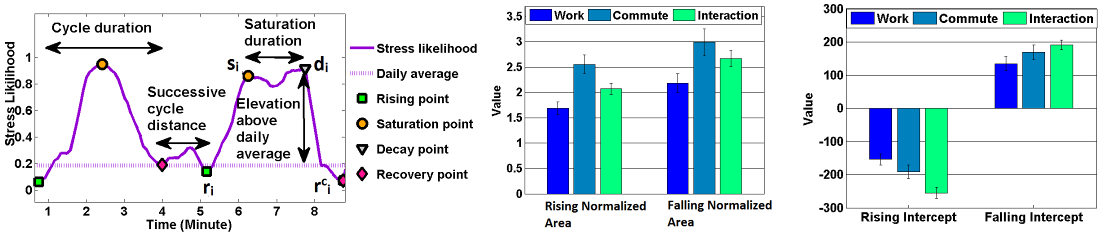


Fig. 6.3: (a) Several duration values for individual stress cycle. (b) Rising and falling normalized area. (c) Rising and falling intercepts of stress cycles.

(right portion). The cycle duration for interaction is significantly lower compared to work related cycle duration with p-value of 0.002 (using t-test).

6.4 Distinguishing Patterns in Wrist Motion Sensors

Researchers have studied the role of gestures during conversational interaction in assessing stress. The more stressful the situation, the higher the proportion of speech that is accompanied by hand gestures [96]. We observe distinct patterns in the wrist-worn motion sensor signals (accelerometer and gyroscope) during stressful interactions compared to work and commute related. We observe frequency of wrist movement is higher during stressful interpersonal interactions. While someone is working at a computer, motion will be more guided towards typing or mouse movement. Similarly, hand motion during driving is expected to be dominated by the steering wheel movement. On the other hand, wrist motion is more random during an interaction, possibly due to communicative gesturing. We hypothesize that wrist motion energy will be higher during a stressful interaction compared to a non-stressful interaction. Based on these insights, we extracted motion sensor features under each stress cycle to compare those differences in order to detect stressful interactions in daily living.

6.5 Feature Computation

To capture differences in stress cycle characteristics during stressful interactions compared to work and commute related stress, we identify new features from each stress cycle. From each cycle, we compute features from stress likelihood time-series and those from wrist-worn inertial sensors. In addition to computing features from individual cycles, we also compute features from two or three consecutive cycles, and all cycles in a stress event.

6.5.1 Features from Individual Stress Cycle

We compute the following features from each stress cycle of a stress event: fractional rising and fractional falling time, rising and falling normalized area, ratio of rising and falling normalized area, elevation above daily average, rising and falling slopes and intercepts, skewness, kurtosis, and entropy. We now describe these features and how they are computed.

To compute these features, we first calculate the following duration measurements from each stress cycle — cycle duration, saturation duration, and successive cycle distance as depicted in Figure 6.3a.

Let S_j be the stress likelihood at time t_j with new values produced every $\Delta t = t_j - t_{j-1} = 5$ seconds. A stress cycle is defined by four 2D points, i.e., $C_i = (r_i, s_i, d_i, r_i^c)$ (as shown in 6.3a). Here, $r_i = \langle t_{r_i}, S_{r_i} \rangle$, $s_i = \langle t_{s_i}, S_{s_i} \rangle$, $d_i = \langle t_{d_i}, S_{d_i} \rangle$, and $r_i^c = \langle t_{r_i^c}, S_{r_i^c} \rangle$.

Stress cycle duration: Stress cycle duration is defined as the temporal distance between stress rising and recovery point,i.e., $CD_i = t_{r_i^c} - t_{r_i}$.

Saturation duration: Saturation duration is the duration when the stress likelihood time-series stays in the upper region after reaching the saturation point before starting to decay, i.e., $SD_i = t_{d_i} - t_{s_i}$.

Successive cycle distance: Successive cycle distance is the distance between ending of one cycle and starting of next cycle, i.e., $SCD_i = t_{r_{i+1}} - t_{r_i^c}$.

With these duration measurements, we compute the following features from each stress cycle.

1. *Fractional rising and falling time:* Fractional rising time is defined as the ratio of rising duration to stress cycle duration where rising duration is defined as the temporal distance between stress cycle start and saturation point. Similarly,

fractional falling time is defined as the ratio of falling duration to stress cycle duration. Falling duration is the temporal distance between decay and recovery points. More specifically, $t_{rise_i} = (t_{s_i} - t_{r_i})/CD_i$ and $t_{fall_i} = (t_{r_i^c} - t_{d_i})/CD_i$.

2. *Rising and falling normalized area, Ratio of rising and falling normalized area:* Rising normalized area is computed as the area under rising region divided by the rising duration. Similarly, falling normalized area is computed as the area under falling region divided by the falling duration. We also use the ratio of these two values as a feature. The variation of this feature values are depicted in Figure 6.3b for different stressors, i.e., $A_{rise_i} = \frac{\sum_{k=r_i}^{s_i} S_k}{t_{rise_i}}$ and $A_{fall_i} = \frac{\sum_{k=d_i}^{r_i^c} S_k}{t_{fall_i}}$.

3. *Elevation above daily average:* The amplitude difference between maximum value or the peak of a cycle and the daily average is defined as elevation above daily average. Peak amplitude of a cycle C_i is $peakAmp_i = max(S_{r_i}, S_{r_{i+1}}, ..., S_{r_i^c})$. Then, elevation above daily average is $E_i = (peakAmp_i - Avg(S_j, \forall j))$.

4. *Rising and Falling slope and Intercepts:* We fit a least square regression line in the rising phase. That is, we find slope m and intercept c of equation $y = mx + c$ using the sequence of points (t_k, S_k) between t_{r_i} and t_{s_i}. Similarly, falling slope and intercept are computed in the decay region. The variation of intercept values are depicted in Figure 6.3C for different stressors.

5. We also compute skewness, kurtosis, entropy for each stress cycle. Since, a stress cycle is defined by four points, i.e., $C_i = (r_i, s_i, d_i, r_i^c)$ therefore all the stress likelihood within the cycle C_i are $S_{r_i}, S_{r_{i+1}}, ..., S_{r_i^c}$. More specifically, skewness is $\frac{\sum_{i=r_i}^{r_i^c} (S_i - \bar{S})^3}{\|r_i^c - r_i\| * std^3}$ and kurtosis is $\frac{\sum_{i=r_i}^{r_i^c} (S_i - \bar{S})^4}{\|r_i^c - r_i\| * std^4}$.

6.5.2 Wrist Motion Features in Each Stress Cycle

From inertial sensor data coinciding with each stress cycle, we compute several time domain features from both accelerometer and gyroscope signals — mean, median, standard deviation, quartile deviation, skewness, and kurtosis of three axes of accelerometer and gyroscope. For wrist orientation features, we compute roll, pitch, and yaw that provide information about the orientation of the wrist with respect to gravity on the window of data. We also computed energy as the magnitude of the accelerometer and magnitude of the gyroscope ($a_{mag} = \sqrt{a_x^2 + a_y^2 + a_z^2}$).

6.5.3 Whole Stress Event Features

To compute features from the entire stress event, we compute number of stress cycles per event, duration of stress cycles per minute, and average stress likelihood across the entire event.

6.5.4 Features from Multiple Stress Cycles

We compute features from consecutive stress cycles (i.e., two cycles or three cycles) to determine the degree of performance improvement with more information. We note that using features from the entire event may delay the detection of stressor until after the stress event is over. The combination features include differential features from successive individual stress cycle features and statistical features such as mean and standard deviation across selected number of cycles. For wrist motion features, we compute only statistical features across selected number of cycles.

6.6 Model Selection and Training

We have grouped the stress events into two categories — interaction and non-interaction. Interaction group includes all the stressful social interactions. Non-interaction group includes all the stress events due to other common daily stressors i.e., work and commute. Our aim is to identify whether a stress cycle is induced due to interaction or non-interaction related stress activity. To do so, we identify each stress cycle automatically from the continuous stress time-series using previously mentioned stress cycle algorithm. Next, we compute the features from each stress cycle and train a machine learning model to identify whether the current stress event is due to interpersonal interactions.

In total, we obtain 13 features from stress cycles and 42 features from the motion sensor data. But, to avoid overfitting (as there are only 152 interaction stress cycles and 129 non-interaction stress cycles), we use selected features for modeling. The idea behind feature selection is to remove highly correlated and noninformative features. In this work, we used the Correlation-based Feature Selection (CFS) to select a subset of the features (15) as in other detection based work [22]. CFS selects features that are mutually uncorrelated but highly indicative of the interaction and non-interaction classes.

We use basic Logistic Regression model (LR) and Random Forest (RF) to train the model to discriminate stressful interaction from non-interaction related stress event. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Mathematically, a binary logistic model has a dependent variable with two possible values labeled as “0” and “1”. In the logistic model, the log-odds (the logarithm of the odds) for the value labeled “1” is a linear combination of one or more independent variables (stress cycle features). Random Forest is an ensemble learning method for classification. It constructs a collection of decision trees trained with random subsets of features and outputs the class which is the consensus of classes output by individual trees. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifier depends on the strength of the individual trees in the forest and the correlation between them.

We assess the performance of the model using standard classification metrics such as precision, recall, and F1 score. We use labelled stress cycles to train the model and run the model with several combinations of features.

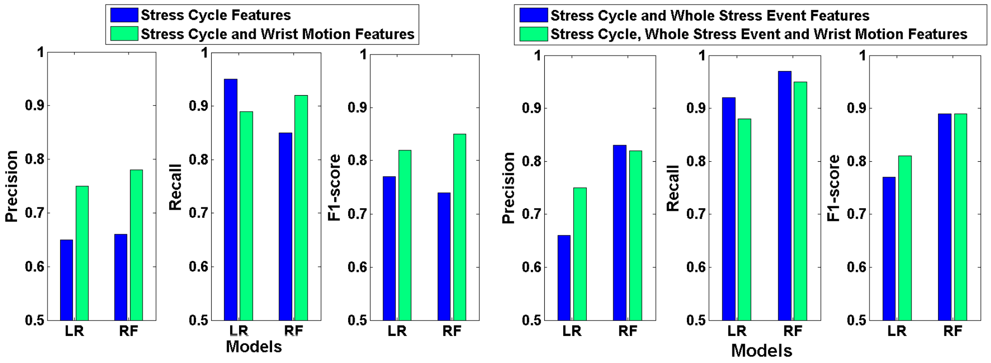


Fig. 6.4: (a) Performance of LR (Logistic Regression) and RF (Random Forest) models using individual stress cycle features and then fusing wrist motion features. (b) Performance of whole stress event features fusing with individual stress cycle features and wrist motion features.

6.7 Model Evaluation and Results

In this section, we evaluate the impact of design choices on the model performance and also compare it with a baseline model.

6.7.1 Performance Improvement by Using Wrist Sensor Features and Whole Stress Event Features

We compare the model performance using both logistic regression and random forest, when using stress cycles features in each cycle, performance improvement when adding features from wrist-worn inertial sensors, and further performance gain when features from the entire stress event are used.

1. *Performance using individual stress cycle features:* The logistic regression (LR) classifier can distinguish whether a stress cycle belongs to interaction class with an F1 score of 0.77 using stratified 10-fold cross validation method using stress cycle features from one cycle (see Figure 6.4a). Precision and recall values are 0.65 and 0.95 respectively. On the other hand, Random Forest (RF) classifier achieves precision, recall and F1-score of 0.66, 0.85, and 0.74, respectively. We consider a stress cycle as the smallest unit for detecting an on-going stressful event. Therefore, to detect a stressor, we need to observe at least one stress cycle in a stress time-series data.

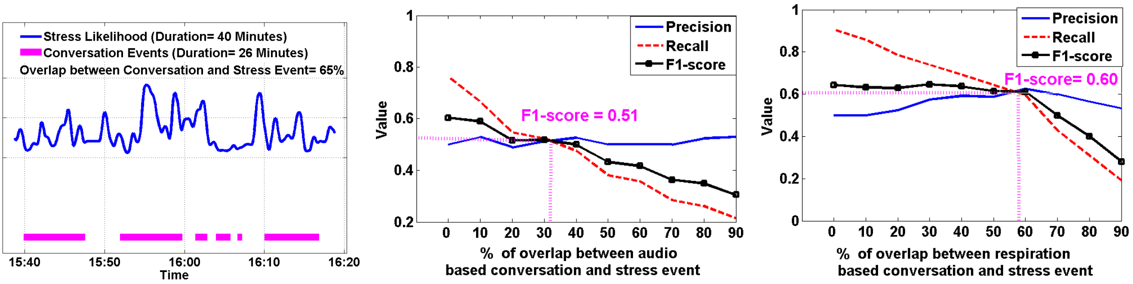


Fig. 6.5: (a) Whether a stress event is due to interpersonal interaction depending on the overlap ratio between detected conversation and stress events. Here conversation is measured from the LENA audio device. (b) If conversation is measured using respiration based model, then the accuracy of finding whether a stress event is due to interaction.

2. *Performance using individual stress cycle and wrist motion features:* After fusing wrist motion features with individual stress cycle features, Logistic regression (LR) can classify with 0.75 precision, 0.89 recall, and 0.82 F1-score. The Random Forest (RF) model can classify the two classes with precision, recall and F1 score of 0.78, 0.92, 0.85, respectively. This shows that adding computationally inexpensive and less power-hungry motion sensors, which are already part of wrist devices, can significantly improve the accuracy of detecting stressful interpersonal interaction.

3. *Performance using whole stress event features:* When we augment whole stress event features with the individual stress cycle features, the precision improves to 0.83, recall improves to 0.97, and F1-score becomes 0.89 using Random Forest Model as shown in Figure 6.4b. After fusing wrist motion features with whole stress event and individual stress cycle features, the precision, recall and F1-score becomes 0.82, 0.95, and 0.89, respectively using Random Forest Model. Adding whole event features gives the best performance among all the evaluation setup. However, to achieve this performance, we need to observe the whole duration of the stressful event. Therefore, this will produce fewer false alarms but at a cost of longer waiting times for interventions.

6.8 Comparison with Baseline Models

To compare the performance of our model, we construct a baseline model. We consider a natural model that compare percentage of stress event duration that is detected to be spent in conversation by user from another data source motivated by the work presented in [35]. To detect the timing of interpersonal interaction, we use an audio based model available from LENA [27] and a respiration based model from [18]. We search for the right value for overlap for each of these two models that results in optimal performance (see Figures 6.5b and 6.5c).

We observe an F1-score of 0.51 for the audio based model with around 32% overlap with the stress event and an F1 score of 0.60 for the respiration based model with 58% overlap with the stress event. Lower F1-scores for these two baseline models can be explained by the fact that models detecting conversations are not perfect (F1-score of 0.7). Secondly, people usually multitask and therefore, even when a user may be in conversation during a stress event, (s)he might be stressed for other reasons. For example, a driver may be in conversation with a co-passenger during driving but stressor can be traffic events.

In addition to 15-30% performance improvement over baseline models, our model also has the advantage of detecting stressor from the stress time series itself, without needing concurrent detection of potential stressors (e.g., conversation from audio, work status from computer logs).

6.9 Implications for Delivery of Just-In-Time Stress Intervention

A just-in-time intervention needs information on most opportune moments for delivering the intervention. In this section, we investigate the trade-off between the accuracy of detecting the stressor and how quickly since the start of the stress event the stressor can be detected.

The model performance is expected to improve when features are computed over longer intervals, but it also comes at the cost of delayed detection. We also observe that when we use features spanning multiple cycles (i.e., two cycles, three cycles), the number

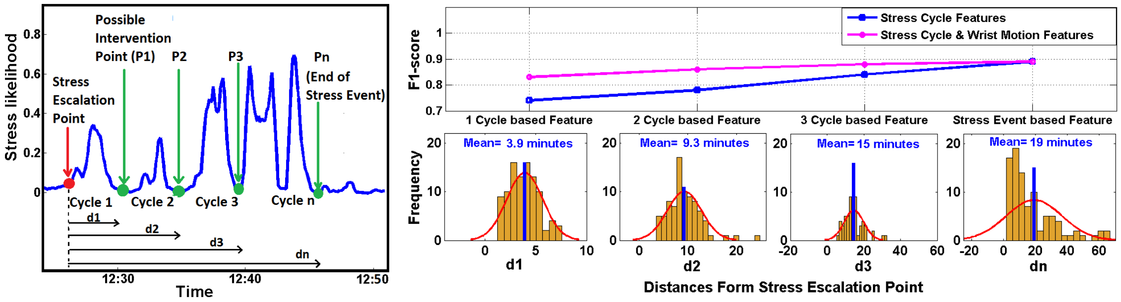


Fig. 6.6: (a) Possible timing/points of delivering intervention depending on features computed from one stress cycle or span of multiple stress cycles. d1, d2, d3 and dn are representing the temporal distances from the beginning of a stress event. (b) Trade-off between sources of stress detection accuracy and timing of stress intervention delivery based on features computed from one stress cycle or features from multiple cycles.

of instances for classification reduces and the model may tend to overfit. The total number of instances while taking features from one cycle, two cycles, and three cycles are 346, 258, and 183, respectively. After computing features from three cycles, the dataset becomes too small to test any further combination of featureset. This issue does not arise when using whole event features as the unit of analysis is still each cycle within the stress event. If d_i^m is the time difference between beginning of the stress event and end of the i^{th} cycle, C_i^m. We can estimate the expected value of d_n^m as

$d_n^m = n * \overline{CD} + (n-1) * \overline{SCD}$; where $\overline{CD}$ and $\overline{SCD}$ are the expected value of cycle duration and successive cycle duration, respectively.

For example if $CD \sim N(\mu_{CD}, \sigma_{CD})$ and $SCD \sim N(\mu_{SCD}, \sigma_{SCD})$ then, $d_n^m = n * \mu_{CD} + (n-1) * \mu_{SD}$

The F1-score using only stress cycle features for one cycle, two cycle, three cycle and whole event based featureset is 0.74, 0.78, 0.84, and 0.89, respectively. Fusing the wrist motion features with stress cycle features increases the F1-score to 0.83, 0.86, 0.88, and 0.89 for one cycle, two cycle, three cycle, and whole event based feature set, respectively. Figure 6.6 shows these results.

Stress intervention designers can consider the trade-offs between the timing of intervention and accuracy of detecting stressful conversations on one or multiple stress cycles. For example, if a quicker intervention is called for, then they can consider intervening after one cycle which will allow them to intervene within 3.9 minutes on average duration from the stress rising point with an F1-score of 0.83 when the stress cycle features are used with wrist motion features (Figure 6.6). To achieve higher accuracy, one can use the model that fuses two cycles together. In that case, F1 score improves to 0.86, but the timing of the delivery will be further delayed (on average, 9.3 minutes from the first stress rising point shown in Figure 6.6a). We note that the best accuracy is achieve when using whole event features, but this will further delay the detection of stressor, as the average duration of a stress event is 19 minutes. These analysis can help find the sweet spot between timing and the stressor detection accuracy for intervention designers.

6.10 Conclusions

Just-in-time (JIT) interventions delivered on personal mobile devices to manage stress have the potential to improve individuals' mental well-being. Novel stress interventions are being developed and recent work on detecting stress and availability using passively collected data can reveal the most opportune moments for delivering such interventions. However, given that stress can be due to very different sources and the intervention suitable for one may be ineffective or even counterproductive for another necessitates knowing the likely source of stress prior to delivering a JIT intervention. This work opens the door for automatically identifying the stressful social interactions by showing that it is feasible to detect from the stress time series data.

Chapter 7

Conclusion

This dissertation introduced the concept of stress cycles within stress event and presented an algorithm to identify them in a stress likelihood timeseries and characterize points of interest in them. It further showed that features derived from stress cycles have sufficient patterns to distinguish stressful conversations from other stressors (with improved accuracy when combined by features derived from hand gestures). This work opens the doors to future research that can collect larger datasets consisting of a large number of other daily stressors and develop models to identify each of them. Such models can be used to determine various sources of stress for each stress event detected by wearable sensors. This information can not only inform the timing of intervention delivery, but also the right content, the adaptation mechanisms for personalizing it to the individual and the user's context, and selecting the right modality for delivery (e.g., smartphone or smartwatch). But, there are several limitations in this work that can inspire future research.

First, this work used stress event detection from chest-worn ECG and respiration sensors. These sensors provide a firmer attachment than pulse plethysmography or electrodermal sensing from conveniently worn wrist devices. Wearing electrodes or a chest belt in field for long term to monitor stress is burdensome and sometimes interferes with daily activities. Therefore, it is a challenge to deploy such systems in the field. However, smart-watches are becoming increasingly popular and recent research work shows feasibility of detecting stress from wrist worn sensors. Future work can assess how well the presented model can be adapted to work with potentially noisier stress time series obtained from wrist-worn sensor data.

Second, this study used data from 38 participants, but the data was collected only for one full day (due to privacy concerns with audio capture in the natural environment).

Future work can investigate the generalizability of the presented model on data collected in longer-term studies and those involving a more general population.

Third, the limited size of dataset (in terms of number and diversity of detected stress events) in this work was insufficient to develop and test a three class classifier to distinguish interaction, work and commute. In future, a larger dataset can enable identification of other stressors as well as support construction of data-driven features in a deep learning model.

Fourth, in addition to stressful conversations, work, and commute, there are numerous other sources of stress such as financial difficulties, health issues, news about friends, family, colleagues, region, country, and the world, among others. Future work can investigate the possibilities of detecting these and other stressors, by potentially exploring novel methods to combine the data collected from other sources with the stress dynamics data.

Fifth, the labeling of stress events was done based on participants interview. As we asked the participants to recall what was causing the stress after showing the detected stress events using the visualization, it may introduce some bias. To better assess recall and detect false positives in stress event detection, a future study can present the surrounding contexts and time segments (both when stress is detected and not detected) without disclosing whether stress event was detected at those times. Another way to reduce bias is to first ask the participants recall major periods of stress and then show them the visualizations to verify the stress events.

Sixth, in this study, participants were asked to recall the main source of stress for the detected stress events. Several situations in real-life involve multitasking where a stress event may be due to confluence of multiple concurrent factors. Future work can investigate methods for detection of multiple concurrent sources and their prevalence in inducing the current stress event.

References

[1] K. H. Teigen, “Yerkes-dodson: A law for all seasons,” *Theory & Psychology*, vol. 4, no. 4, pp. 525–547, 1994.

[2] S. C. Segerstrom and G. E. Miller, “Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry.” *Psychological bulletin*, vol. 130, no. 4, p. 601, 2004.

[3] D. M. Almeida, E. Wethington, and R. C. Kessler, “The daily inventory of stressful events: An interview-based approach for measuring daily stressors,” *Assessment*, vol. 9, no. 1, pp. 41–55, 2002.

[4] C. Jungbluth, I. M. MacFarlane, P. M. Veach, and B. S. LeRoy, “Why is everyone so anxious?: an exploration of stress and anxiety in genetic counseling graduate students,” *Journal of Genetic Counseling*, vol. 20, no. 3, pp. 270–286, 2011.

[5] G. Stadler, K. A. Snyder, A. B. Horn, P. E. Shrout, and N. P. Bolger, “Close relationships and health in daily life: A review and empirical data on intimacy and somatic symptoms,” *Psychosomatic Medicine*, 2012.

[6] J. K. Kiecolt-Glaser, J.-P. Gouin, and L. Hantsoo, “Close relationships, inflammation, and health,” *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 1, pp. 33–38, 2010.

[7] A. J. Fuligni, E. H. Telzer, J. Bower, S. W. Cole, L. Kiang, and M. R. Irwin, “A preliminary study of daily interpersonal stress and c-reactive protein levels among adolescents from latin american and european backgrounds,” *Psychosomatic Medicine*, 2009.

[8] L. Coleman, J. Mitcheson, and G. Lloyd, “Couple relationships: Why are they important for health and wellbeing?” *Journal of Health Visiting*, vol. 1, no. 3, pp. 168–172, 2013.

[9] D. K. Snyder and W. K. Halford, “Evidence-based couple therapy: Current status and future directions,” *Journal of Family Therapy*, vol. 34, no. 3, pp. 229–249, 2012.

[10] R.-P. Juster, B. S. McEwen, and S. J. Lupien, “Allostatic load biomarkers of chronic stress and impact on health and cognition,” *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 1, pp. 2–16, 2010.

[11] J. Taelman, S. Vandeput, E. Vlemincx, A. Spaepen, and S. Van Huffel, “Instantaneous changes in heart rate regulation due to mental load in simulated office work,” *European journal of applied physiology*, vol. 111, no. 7, pp. 1497–1505, 2011.

[12] D. MacLean, A. Roseway, and M. Czerwinski, “Moodwings: a wearable biofeedback device for real-time stress intervention,” in *Proceedings of the 6th*

international conference on PErvasive Technologies Related to Assistive Environments, 2013, pp. 1–8.

[13] J. Costa, A. T. Adams, M. F. Jung, F. Guimbretière, and T. Choudhury, "Emotioncheck: leveraging bodily signals and false feedback to regulate our emotions," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 758–769.

[14] J. Costa, M. F. Jung, M. Czerwinski, F. Guimbretière, T. Le, and T. Choudhury, "Regulating feelings during interpersonal conflicts by changing voice self-perception," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[15] J. Costa, F. Guimbretière, M. F. Jung, and T. Choudhury, "Boostmeup: Improving cognitive performance in the moment by unobtrusively regulating emotions with a smartwatch," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–23, 2019.

[16] P. E. Paredes, Y. Zhou, N. A.-H. Hamdan, S. Balters, E. Murnane, W. Ju, and J. A. Landay, "Just breathe: In-car interventions for guided slow breathing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–23, 2018.

[17] M. M. Rahman, A. A. Ali, K. Plarre, M. al'Absi, E. Ertin, and S. Kumar, "mConverse: Inferring Conversation Episodes from Respiratory Measurements Collected in the Field," in *ACM Wireless Health*, 2011.

[18] R. Bari, R. J. Adams, M. M. Rahman, M. B. Parsons, E. H. Buder, and S. Kumar, "rconverse: Moment by moment conversation detection using a mobile respiration sensor," *ACM IMWUT*, vol. 2, no. 1, p. 2, 2018.

[19] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cstress: towards a gold standard for continuous stress assessment in the mobile environment," in *ACM UbiComp*, 2015.

[20] D. M. Almeida and R. C. Kessler, "Everyday stressors and gender differences in daily distress." *Journal of personality and social psychology*, vol. 75, no. 3, p. 670, 1998.

[21] A. Natarajan, D. Ganesan, and B. M. Marlin, "Hierarchical active learning for model personalization in the presence of label scarcity," in *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2019, pp. 1–4.

[22] H. Sarker, M. Tyburski, M. M. Rahman, K. Hovsepian, M. Sharmin, D. H. Epstein, K. L. Preston, C. D. Furr-Holden, A. Milam, I. Nahum-Shani *et al.*, "Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data," in *ACM SIGCHI*, 2016, pp. 4489–4501.

[23] D. Kahneman, A. B. Krueger, D. A. Schkade, N. Schwarz, and A. A. Stone, “A survey method for characterizing daily life experience: The day reconstruction method,” *Science*, pp. 1776–1780, 2004.

[24] A. C. Timmons, T. Chaspari, S. C. Han, L. Perrone, S. S. Narayanan, and G. Margolin, “Using multimodal wearable technology to detect conflict among couples,” *Computer*, no. 3, pp. 50–59, 2017.

[25] A. Gujral, T. Chaspari, A. C. Timmons, Y. Kim, S. Barrett, and G. Margolin, “Population-specific detection of couples’ interpersonal conflict using multi-task learning,” in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018.

[26] H. Sarker, K. Hovsepian, S. Chatterjee, I. Nahum-Shani, S. A. Murphy, B. Spring, E. Ertin, M. AlAbsi, M. Nakajima, and S. Kumar, “From markers to interventions: The case of just-in-time stress intervention,” in *Mobile health*, 2017, pp. 411–433.

[27] “LENA research foundation,” http://www.lenafoundation.org//, Accessed: February 2019.

[28] S. J. Wilson, B. E. Bailey, L. M. Jaremka, C. P. Fagundes, R. Andridge, W. B. Malarkey, K. M. Gates, and J. K. Kiecolt-Glaser, “When couples hearts beat together: Synchrony in heart rate variability during conflict predicts heightened inflammation throughout the day,” *Psychoneuroendocrinology*, vol. 93, pp. 107–116, 2018.

[29] B. Lamichhane, U. Großekathöfer, G. Schiavone, and P. Casale, “Towards stress detection in real-life scenarios using wearable sensors: normalization factor to reduce variability in stress physiology,” in *eHealth 360*, 2017, pp. 259–270.

[30] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, “Wearable affect and stress recognition: A review,” *arXiv preprint arXiv:1811.08854*, 2018.

[31] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, “Development and evaluation of an ambulatory stress monitor based on wearable sensors,” *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2011.

[32] E. J. de Geus, L. J. Van Doornen, and J. F. Orlebeke, “Regular exercise and aerobic fitness in relation to psychological make-up and physiological stress reactivity.” *Psychosomatic medicine*, vol. 55, no. 4, pp. 347–363, 1993.

[33] S. I. Powers, P. R. Pietromonaco, M. Gunlicks, and A. Sayer, “Dating couples’ attachment styles and patterns of cortisol reactivity and recovery in response to a relationship conflict.” *Journal of personality and social psychology*, vol. 90, no. 4, p. 613, 2006.

[34] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, “A non-eeg biosignals dataset for assessment and visualization of neurological status,” in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2016, pp. 110–114.

[35] M. Lee, J. Moon, D. Cheon, J. Lee, and K. Lee, “Respiration signal based two layer stress recognition across non-verbal and verbal situations,” in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 638–645.

[36] M. al’Absi, S. Bongard, T. Buchanan, G. A. Pincomb, J. Licinio, and W. R. Lovallo, “Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors,” *Psychophysiology*, vol. 34, no. 3, pp. 266–275, 1997.

[37] M. Myrtek and G. Brügner, “Perception of emotions in everyday life: studies with patients and normals,” *Biological psychology*, vol. 42, no. 1-2, pp. 147–164, 1996.

[38] J. A. Healey and R. W. Picard, “Detecting stress during real-world driving tasks using physiological sensors,” *IEEE TITS*, 2005.

[39] K. Plarre and et. al., “Continuous inference of psychological stress from sensory measurements collected in the natural environment,” in *ACM IPSN*, 2011.

[40] R. K. Nath, H. Thapliyal, and A. Caban-Holt, “Validating physiological stress detection model using cortisol as stress bio marker,” in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, 2020, pp. 1–5.

[41] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, “Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study,” *Sensors*, vol. 19, no. 8, p. 1849, 2019.

[42] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al’Absi, T. Kwon, S. Mitra, S. Shah, and J. Jeong, “AutoSense: Unobtrusively Wearable Sensor Suite for Inferencing of Onset, Causality, and Consequences of Stress in the Field,” in *ACM SenSys*, 2011.

[43] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, “The hcrc map task corpus,” *Journal of Language and speech*, 1991.

[44] C. Sas, S. Challioner, C. Clarke, R. Wilson, A. Coman, S. Clinch, M. Harding, and N. Davies, “Self-defining memory cues: Creative expression and emotional meaning,” in *ACM CHI Extended Abstract*, 2015.

[45] R. Montoliu, J. Blom, and D. Gatica-Perez, “Discovering places of interest in everyday life from smartphone data,” *Springer Multimedia Tools and Applications*, pp. 179–207, 2013.

[46] K. P. Tang, J. I. Hong, and D. P. Siewiorek, “Understanding how visual representations of location feeds affect end-user privacy concerns,” in *ACM UbiComp*, 2011.

[47] R. W. Bohannon, “Comfortable and maximum walking speed of adults aged 20–79 years: reference values and determinants,” *Age and Aging*, pp. 15–19, 1997.

[48] S. Vhaduri, A. Ali, M. Sharmin, K. Hovsepian, and S. Kumar, "Estimating drivers' stress from gps traces," in *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2014, pp. 1–8.

[49] M. M. Rahman, R. Bari, A. A. Ali, M. Sharmin, A. Raij, K. Hovsepian, S. M. Hossain, E. Ertin, A. Kennedy, D. H. Epstein, and Others, "Are we there yet?: feasibility of continuous stress assessment via wireless physiological sensors," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 479–488.

[50] M. Sharmin, A. Raij, D. Epstien, I. Nahum-Shani, J. G. Beck, S. Vhaduri, K. Preston, and S. Kumar, "Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 505–516.

[51] A. Ståhl, K. Höök, and E. Kosmack-Vaara, "Reflecting on the design process of affective health," in *IASDR, 31 October-4 november, Delft, the Netherlands*, 2011, pp. 1–12.

[52] T. Munzner, *Visualization analysis and design*. AK Peters/CRC Press, 2014.

[53] A. Muaremi, B. Arnrich, and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," *BioNanoScience*, 2013.

[54] G. B. Spanier, "Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads," *Journal of Marriage and the Family*, 1976.

[55] D. H. McFarland, "Respiratory markers of conversational interaction," *Journal of Speech, Language, and Hearing Research*, 2001.

[56] R. George, S. Vedam, T. Chung, V. Ramakrishnan, and P. Keall, "The application of the sinusoidal model to lung cancer patient respiratory motion," *Medical physics*, 2005.

[57] Y. Suh, S. Dieterich, B. Cho, and P. J. Keall, "An analysis of thoracic and abdominal tumour motion for stereotactic body radiotherapy patients," *Physics in medicine and biology*, 2008.

[58] X. Long, J. Yang, T. Weysen, R. Haakma, J. Foussier, P. Fonseca, and R. M. Aarts, "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging," *Physiological measurement*, 2014.

[59] W. Lu, M. M. Nystrom, P. J. Parikh, D. R. Fooshee, J. P. Hubenschmidt, J. D. Bradley, and D. A. Low, "A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms," *Medical physics*, 2006.

[60] P. Lopez-Meyer and E. Sazonov, "Automatic breathing segmentation from wearable respiration sensors," in *IEEE ICST*, 2011.

[61] A. Wilson, C. Franks, and I. Freeston, "Algorithms for the detection of breaths from respiratory waveform recordings of infants," *Medical and Biological Engineering and Computing*, 1982.

[62] T. Rahman, A. T. Adams, R. V. Ravichandran, M. Zhang, S. N. Patel, J. A. Kientz, and T. Choudhury, "Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar," in *ACM UbiComp*, 2015.

[63] R. Lukočius, J. A. Virbalis, J. Daunoras, and A. Vegys, "The respiration rate estimation method based on the signal maximums and minimums detection and the signal amplitude evaluation," *Elektronika ir Elektrotechnika*, 2015.

[64] C. Daluwatte, C. G. Scully, G. C. Kramer, and D. G. Strauss, "A robust detection algorithm to identify breathing peaks in respiration signals from spontaneously breathing subjects," in *Computing in Cardiology*, 2015.

[65] R. Tehrany, "Speech breathing patterns in health and chronic respiratory disease," Ph.D. dissertation, 2015.

[66] J. R. Finkel, A. Kleeman, and C. D. Manning, "Efficient, Feature-based, Conditional Random Field Parsing." in *ACL*, vol. 46, 2008, pp. 959–967.

[67] T. Nguyen, R. J. Adams, A. Natarajan, and B. M. Marlin, "Parsing wireless electrocardiogram signals with context free grammar conditional random fields," in *IEEE Wireless Health*, 2016.

[68] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[69] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[70] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, 2005.

[71] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Computer speech & language*, vol. 4, no. 1, pp. 35–56, 1990.

[72] B. Taskar, D. Klein, M. Collins, D. Koller, and C. D. Manning, "Max-Margin Parsing." in *EMNLP*, vol. 1, no. 1.1. Citeseer, 2004, p. 3.

[73] A. Natarajan, G. Angarita, E. Gaiser, R. Malison, D. Ganesan, and B. M. Marlin, "Domain adaptation methods for improving lab-to-field generalization of cocaine detection using wearable ecg," in *ACM UbiComp*, 2016.

[74] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, “Soundsense: scalable sound sensing for people-centric applications on mobile phones,” in *ACM MobiSys*, 2009.

[75] T. Kim, A. Chang, L. Holland, and A. S. Pentland, “Meeting mediator: enhancing group collaborationusing sociometric feedback,” in *Proceedings of the ACM conference on Computer supported cooperative work*, 2008.

[76] Y. Lee, C. Min, C. Hwang, J. Lee, I. Hwang, Y. Ju, C. Yoo, M. Moon, U. Lee, and J. Song, “Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion,” in *ACM MobiSys*, 2013.

[77] C. Xu, S. Li, G. Liu, Y. Zhang, E. Miluzzo, Y.-F. Chen, J. Li, and B. Firner, “Crowd++: unsupervised speaker count with smartphones,” in *ACM UbiComp*, 2013.

[78] M. Y. Ahmed, S. Kenkeremath, and J. Stankovic, “Socialsense: A collaborative mobile platform for speaker and mood identification,” in *European Conference on Wireless Sensor Networks*, 2015.

[79] H. Lu, A. B. Brush, B. Priyantha, A. K. Karlson, and J. Liu, “Speakersense: Energy efficient unobtrusive speaker identification on mobile phones,” in *Pervasive Computing*, 2011.

[80] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, “Emotionsense: a mobile phones based adaptive platform for experimental social psychology research,” in *ACM UbiComp*, 2010.

[81] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, “Stresssense: Detecting stress in unconstrained acoustic environments using smartphones,” in *ACM UbiComp*, 2012.

[82] “GigaOm,” https://gigaom.com/2013/09/20/could-a-breath-monitoring-headset-improve-your-health/, Accessed: January 2018.

[83] S. Nukaya, T. Shino, Y. Kurihara, K. Watanabe, and H. Tanaka, “Noninvasive bed sensing of human biosignals via piezoceramic devices sandwiched between the floor and bed,” *IEEE Sensors journal*, vol. 12, no. 3, pp. 431–438, 2012.

[84] J. M. Perez-Macias, H. Jimison, I. Korhonen, and M. Pavel, “Comparative assessment of sleep quality estimates using home monitoring technology,” in *IEEE EMBC*, 2014.

[85] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, “Smart homes that monitor breathing and heart rate,” in *ACM CHI*, 2015.

[86] J. Gao, E. Ertin, S. Kumar, and M. al’Absi, “Contactless sensing of physiological signals using wideband rf probes,” in *Asilomar Conference on Signals, Systems and Computers*, 2013.

[87] J. Penne, C. Schaller, J. Hornegger, and T. Kuwert, "Robust real-time 3d respiratory motion detection using time-of-flight cameras," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, no. 5, pp. 427–431, 2008.

[88] H. Wang, D. Zhang, J. Ma, Y. Wang, Y. Wang, D. Wu, T. Gu, and B. Xie, "Human respiration detection with commodity wifi devices: do user location and body orientation matter?" in *ACM UbiComp*, 2016.

[89] "Zephyr Bioharness," https://www.zephyranywhere.com/, Accessed: February 2019.

[90] "Mobile Health News," http://www.mobihealthnews.com/content/31-new-digital-health-tools-showcased-ces-2017, Accessed: May 2017.

[91] "Stress Beating Tech," https://www.wareable.com/wearable-tech/stress-beating-tech-to-keep-you-sane, Accessed: May 2017.

[92] "Philips Watch," http://www.usa.philips.com/c-m-hs/health-programs/health-watch, Accessed: May 2017.

[93] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *IEEE CVPR*, 2012.

[94] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *IEEE Robotics and Automation*, 2012.

[95] R. J. Adams, A. Parate, and B. M. Marlin, "Hierarchical span-based conditional random fields for labeling and segmenting events in wearable sensor data streams," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 334–343.

[96] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "Recognizing stress using semantics and modulation of speech and gestures," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 162–175, 2015.