HHS Public Access

Author manuscript

Proc ACM Interact Mob Wearable Ubiquitous Technol. Author manuscript; available in PMC 2018 November 08.

Published in final edited form as:

Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018 March ; 2(1): . doi:10.1145/3191734.

rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor

RUMMANA BARI,
University of Memphis, Electrical and Computer Engineering, Memphis, TN, 38152, USA, rummanabari@gmail.com

ROY J. ADAMS,
University of Massachusetts Amherst, Computer Science, Amherst, MA, USA

MD. MAHBUBUR RAHMAN,
University of Memphis, Now works at Samsung Research America, Mountain View, CA, USA

MEGAN BATTLES PARSONS,
University of Memphis, Communication Science and Disorder, Memphis, TN, USA

EUGENE H. BUDER, and
University of Memphis, Communication Science and Disorder, Memphis, TN, USA

SANTOSH KUMAR
University of Memphis, Computer Science, Memphis, TN, USA.

Abstract

Monitoring of in-person conversations has largely been done using acoustic sensors. In this paper, we propose a new method to detect moment-by-moment conversation episodes by analyzing breathing patterns captured by a mobile respiration sensor. Since breathing is affected by physical and cognitive activities, we develop a comprehensive method for cleaning, screening, and analyzing noisy respiration data captured in the field environment at individual breath cycle level. Using training data collected from a speech dynamics lab study with 12 participants, we show that our algorithm can identify each respiration cycle with 96.34% accuracy even in presence of walking. We present a Conditional Random Field, Context-Free Grammar (CRF-CFG) based conversation model, called *rConverse*, 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 32 participants. rConverse identifies conversation episodes with 71.7%

ACM Reference Format:
Rummana Bari, Roy J. Adams, Md. Mahbubur Rahman, Megan Battles Parsons, Eugene H. Buder, and Santosh Kumar. 2018. rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor. *Proc. of ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 2 (March 2018), 27 pages. https://doi.org/10.1145/3191734

accuracy on 254 hours of field data. For comparison, the accuracy from a high-quality audio-recorder on the same data is 71.9%.

Keywords

Respiration Signal; Conversation Modeling; Wearable Sensing; Machine Learning

1 INTRODUCTION

Social interaction is a fundamental aspect of human life [11]. The most direct form of social interaction occurs through conversations. A deep understanding of social and psychological contexts during conversation helps improve interpersonal communication skills through taking and giving turns at the right moment in social and professional settings, and improves overall mental well-being, work performance, and productivity [43, 58].

Scientists for decades have proposed diverse methodologies to analyze audio data recorded during conversation episodes to characterize conversation through various attributes such as speech content [35], speaker identification [9, 33], group size [64], speaker's stress [34], and emotion [46]. However, audio based models depend on utterances and miss some interesting aspects of silent moments during conversation, especially unsuccessful attempts to take turns [50], cutting someone off, utterance planning [17], and the potential psycho-physiological stress of not being able to take a turn when a listener has an urge to speak [14, 27, 39].

In addition to generating sounds, conversation also causes specific changes in breathing patterns. Studies consistently show that speech production is achieved by a specific control of breathing, visible in the clear reduction of the inhalation duration relative to the exhalation duration, as compared with quiet breathing [38, 49, 60, 61, 63]. This gives speech breathing its well-known asymmetrical profile [21, 22, 50]. Thus, breathing kinematics can provide useful information about a person's speaking status.

Use of respiration measurements to assess conversations can open up numerous avenues for investigating the role of conversation in health outcomes. For example, respiration measurements are now routinely collected in field studies for smoking cessation [51] and stress regulation [26]. Conversation plays an important role in these and other health outcomes (e.g., depression) as well as in everyday work performance. With a respiration-based model for detecting conversations, conversation patterns can now be obtained from data already collected in such studies, and help investigate the role of conversation in determining health and performance outcomes.

Respiration-based conversation modeling can potentially enable an assessment of urge to speak, even in the absence of vocalized utterances. Listener's urge to speak and unsuccessful attempts to take turns influence respiratory patterns and interrupt physiological rhythms [27, 50], which may increase listener's stress [14, 39]. Another benefit of respiration sensing is that respiration along with other physiological data (e.g., heart rate variability) can be used to infer psycho-physiological stress even when there are no utterances (e.g., before and after

speech) [23]. The first step to enable the above mentioned analyses is to demonstrate the feasibility of detecting conversations from respiration infield settings.

Respiration based conversation assessment has traditionally been underexplored. Emerging connected wearables [6, 15] and contactless sensing technologies [8, 59] are making it increasingly feasible to reliably capture respiration data continuously infield setting. For example, a commercially available accelerometer-based small device called Spire (which can be clipped to clothing) helps people capture breathing and visualize the signals on a smartphone [13]. Moreover, Hao et al., [20] shows that real-time breathing waveforms can be monitored by analyzing data from a gyroscope sensor embedded in a smartwatch.

mConverse [47] was one of the first works to demonstrate that respiration measurements captured from a respiration sensor can be used to infer naturally-occurring conversation events. But, this early model could only work on 30-second windows, that usually contain a mixture of speech and non-speech events in a spontaneous conversation. Hence, a decision on a 30-second window cannot reveal moment-by-moment turn-taking and turn-holding behaviors, let alone urge to speak analysis. Finally, the mConverse model was not validated with audio ground-truth collected infield environment.

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. 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 respiration data.

A first challenge is the accurate identification of breathing cycles, i.e., 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.

A second 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 (as done in mConverse [47]) is not scalable to the field environment. Further, turn taking can occur swiftly, making it impossible to keep track of and synchronize the labels to the sensor data.

A third 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 fourth 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.

In this paper, we present a rigorous method for screening, cleaning respiration signals and improved algorithms for identifying respiration cycles captured infield setting. We propose a novel feature selection method to select cycle level features to improve lab to field generalizability. We apply a model from machine learning called the Conditional Random Field, Context-Free Grammar (CRF-CFG) model to infer conversation episodes at fine-granularity, achieving 95.9% accuracy in the lab and 71.7% accuracy in the field settings (using audio data for labels). These are comparable with conversation detection from high-quality audio recordings from the LENA device [4].

2 BACKGROUND AND RELATED WORKS

Conversation modeling, based on acoustic data captured with smartphone microphones [35] or with wearable microphones [24] 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 [35], but also to model various characteristics of a conversation, including turn-taking behavior [30], group size estimation [64], and speaker identification [9, 33]. Furthermore, acoustic researchers have also addressed speakers' emotions [46] and stress levels [34]; and developed socio-therapy applications [30] for children with autism.

In this paper, 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 [47].

Respiration-based conversation modeling is, however, underexplored, perhaps due to the lack of reliable respiration signals collected infield 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 [5]. 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 [42, 45] 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 conned to a bed or any furniture [8, 18]. For example, Adib et al., developed a radar based, contactless Vital-Radio [8] 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 [44] and WiFi [59]. 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 [6, 15] and smart garments [1] 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 [3] help users capture breathing information and visualize on a smartphone to aid in breathing regulation. The Philips Health Watch [2], an FDA[1] 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[2]. 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 [47] 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 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

[1]US Food and Drug Administration. https://www.fda.gov/
[2]See https://md2k.org/studies for a list of these deployments.

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 [7, 41, 53, 54]. For example, Tang et al., [54] and Sung et al., [53] use conditional segmentation models for labeling and segmenting activities in video streams. Adams et al., [7] use a hierarchical segmentation model to label and segment smoking activities in respiration data. Most closely related to our approach, [41] 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.

3 DATA COLLECTION AND LABELING

For development, training, and testing of our model, we collected data in both lab and field settings as described below. All studies were approved by the Institutional Review Board (IRB) at University of Memphis, and all participants provided written informed consent.

3.1 Data Collection

3.1.1 Lab Data.—High quality lab data is collected in two settings - (1) in a true laboratory setting and (2) in a natural environment. Data collected in the laboratory setting is designed to: validate the performance of the chest band sensor with a hospital grade system and to collect conversation data in sitting position from a heterogeneous group of 12 participants (6 couples) recruited from a diverse population i.e., students, full-time professionals and part-time employees. Within the sample recruited in this study, there are 7 women (mean age: 29.9 ± 7.4 years) and 5 men (mean age: 27.2 ± 2.9 years). The ‘Field’ training data is collected to enhance the generalizability of the model to detect conversation in presence of free-living activity since activity also affects respiration measurements.

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 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 sensor [15] that collects respiration and 3-axis accelerometer signals (Figure 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. Electrocardiogram (ECG) data were also collected for stress monitoring: These data will be analyzed and reported separately to investigate relationships between stress and conversation.

A headset microphone as shown in Figure 1b was placed in front of the participant's mouth and processed through an analog amplifier. Participants also wore a throat microphone (see Figure 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 1d shows the whole lab setup where conversation partners were seated face-to-face, as captured using the side view video camera.

Participants engaged in several tasks. For the 'Quiet Breathing' task, 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 [10] which elicits goal-oriented conversation. During this phase, a (blocking) screen was placed between the participants. 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. They then switched roles and were given another set of maps to generate another conversation to complete the task. A screen was placed between them for visual separatation. 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. We did not use this data for modeling due to difficulty in labeling in the presence of rapid turn taking. Finally, to obtain spontaneous natural dialogue, participants were encouraged to engage in continuous speech on their chosen topic for fifteen minutes.

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

For this study, participants wore an AutoSense chestband sensor underneath their clothes (Figure 2a). They also carried an android smartphone (Xperia X10) and an audio recorder [4] shown in Figures 2b and 2c. The LENA recorder is housed in a fixed position relative to the speaker's mouth in a secure manner that minimized noise and maintained orientation of the microphone. We designed a phone interface with labels: Walk-Talk and Walk-NoTalk. Participants 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.

3.1.2 Field Study—A 'Field' study was designed to evaluate the performance of the conversation model in the natural environment with 38 participants (19 couples). Participants wore the sensors for a day during their awake hours. As raw recording of audio data in the natural environment poses significant privacy concerns, the data collection was limited to one day, while the number of participants was increased to broaden the diversity across individuals as well as situations.

Both partners wore the AutoSense respiration band underneath their clothes, carried a smart phone and a wearable audio recorder LENA (see Figure 2a, 2b and 2c). They were told to carry the recorder in a waist pouch which was placed around abdomen to reduce occlusion of microphone. The recorder was able to record 16 hours of continuous audio. This setup maximized the chance to capture high quality audio from field.

3.2 Data Labeling

3.2.1 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 3a. 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 3b. 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.

3.2.2 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 le 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.

4 DATA SCREENING AND PROCESSING TO LOCATE EACH BREATH CYCLE

The first stage in detecting conversation from respiratory waveforms is the automated detection of individual breath cycles. In this section, we describe a method to identify respiration cycles automatically in both the lab and field settings.

4.1 Data Pre-processing

Since respiration data from wireless on-body sensors exhibit significant baseline drift, the first step is to account for baseline drift. We normalize the signal within each five minute window by subtracting off the mean within the segment. All of our subsequent processing steps are applied on baseline corrected respiration data. Respiration signals are impacted by physical movement and positioning of the chestband. We mark the signal acceptable as long as the signal retains the characteristic morphology of a respiration signal. After removing poor quality signals, we apply the following cycle identification method on baseline removed data to locate each breath cycle.

4.2 Cycle Identification

The simplest procedure for detecting breaths is a threshold level detector [32, 47, 48, 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 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) [12, 31, 37]. 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 [36]. 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 [56], 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 [36]. 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. We present the details of this method in the following section.

4.2.1 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 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 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 1.

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

Step 2: Moving Average Centerline (MAC).: The next step is to compute a moving average centerline (MAC) curve using Equation 2, where y is the smoothed respiratory signal, L its duration, t is time, and $\overline{y(t)}\Big|_{t-T}^{t+T}$ the average value of y during [$t1$, $t2$]. The MAC appears as a center line (shown as red dotted line in Figure 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.

$$MAC(t) = \overline{y(t)}\Big|_{t-T}^{t+T},\ if\ T < t \leq L - T \quad (2)$$

After visual inspection we found that, in cases of large baseline drift infield 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 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: $I_{dn}(1) < I_{up}(1) < I_{dn}(2) < I_{up}(2) \ldots < I_{dn}(m) < I_{up}(m)$ 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 $i = 1, 2, \ldots, p$ and p = number of peaks. In cases of a regular breathing signal (as Figure 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 4d and 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 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 [19, 52], and total cycle duration lies within the range of 0.8 seconds to 12.5 seconds [23, 52], 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 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 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, 23, 52], 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 [52] 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 4h. Second, if an inspiration or expiration amplitude is too small, 10% of the mean cycle amplitude, the associated cycle is not considered to be of good quality and is screened out.

4.3 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 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.

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

2. ***Missed cycle rate*.** Missing of one or more cycles results in elongated cycle duration as shown in Figure 5b. *Missed cycles Rate:* Percentage of cycles that are missed with respect to total number of actual cycles (N).

   $$Error(\%) = \text{Number of missed cycles}/N * 100$$

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 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 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 dened as *Error in Cycle duration* (Δ_C).

4.4 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 [47] where the threshold is set by taking the average of the signal for every 30 second window. The second one is a change point detection method described in [31]. We also present the performance evaluation of the semi-automatic method [36], which we call the 'base method'.

4.4.1 Evaluation on Lab Data.—We compare the performance of the current method on lab data (1,938 marked respiration cycles) with the base method [36] as well as two other methods i.e., the threshold based and Maxima-Minima based methods. The results are presented in Table 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 [47] 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.

4.4.2 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 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 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 4b.

Table 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.

5 SPEECH DETECTION USING CONDITIONAL RANDOM FIELD-CONTEXT FREE GRAMMAR (CRF-CFG)

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 [16] 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 [41]. 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 [28] 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 8, we present experiments validating this model on the lab and field data described in previous sections.

5.1 Conditional Random Fields

Conditional randomfields (CRFs) are a sub-class of probabilistic graphical models [25] that encode correlations between label variables. A CRF denes a conditional distribution over a set of L label variables $\mathrm{Y} = \{Y_1, \ldots, Y_L\}$ given a corresponding set of M feature variables $\mathrm{X} = \{\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{M}\}$. We assume each feature variable $\mathrm{X}_i \in \mathbb{R}^D$ 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 Y. Throughout this work, we will use upper-case to refer to random variables (e.g., Y) and lower case to refer to particular assignments to those variables (e.g., 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 Y and X:

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

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 $P_\theta(y|x)$ of a setting of the label variables $y = [y_1,\ldots, y_L]$ conditioned on the observed feature variables $x = [x_1,\ldots, x_L]$ is given below. $Z_W(x)$ is referred to as the *partition function* and is the normalization term of the probability distribution.

$$p_\theta(y|x) = \frac{\exp(-E_\theta(y, x))}{\sum_{y \in \mathcal{Y}^L} \exp(-E_\theta(y, x))} \quad (3)$$

The parameters of a CRF can be estimated using either maximum likelihood estimation (MLE) or max-margin learning [57]. 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 [25]. 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 randomfield model which remedies this problem.

5.2 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[3]. 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, consider a simple CFG with $\mathcal{I} = \{\gamma, A, B\}, \mathcal{V} = \{a, b\}$ and the production rules $\gamma \rightarrow AB, A \rightarrow aA, A \rightarrow a, B \rightarrow bB, B \rightarrow b$.[4] The recursive application of 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 offinding 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

[3]We assume a slightly relaxed form equivalent to Chomsky normal form.
[4]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 $B \rightarrow bB|B$.

probable parse tree under this distribution. The conditional randomfield context free grammar (CRF-CFG) model presented in the next section further conditions weighted CFG on features of the input sequence.

5.3 The CRF-CFG Model

The conditional randomfield context free grammar (CRF-CFG) model is a CRF model that denes a distribution over parse trees given a grammar $G = (\mathcal{I}, \mathcal{V}, \mathcal{R}, \gamma)$ and a length L feature sequence $x = [x_1,\ldots, x_L]$.[16]. The set of all parse trees is represented by a set of binary random variables $Y=\{y_{A,BC,i,j,l} | A \rightarrow BC \in \mathcal{R}, 1 \leq i \leq j < l \leq L\}$ $y_{A,BC,i,j,l}$ 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 $r \in \mathcal{R}: f_k^r(y_{r,i,j,l}, i, j, l, x)$ for $k = 1, ..., K^r$. $f_k^r(y_{r,i,j,l}, i, j, l, x)$ takes the value 0 if $y_{r,i,j,l} = 0$ otherwise it may be any function of the input sequence x and the indices of the) production rule i, j, and l which leads to tremendousflexibility.

Finally, the probability of a parse tree y given an input sequence x is given by

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

where $\mathbb{1}$ 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. Specically, 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 [29].

5.4 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 6.

$$\begin{aligned} \gamma &\rightarrow \alpha|\beta \\ \alpha &\rightarrow C\beta|C \\ \beta &\rightarrow O\alpha|O \\ O &\rightarrow sO|qO|s|q \\ C &\rightarrow sC|qC|s|q \end{aligned} \tag{4}$$

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 [55, 57]. For the augmentation loss, we use the Hamming loss between the true and predicted sequence labels.

6 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 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 andfield domains.

6.1 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 7

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 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 7.

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 inspiratoryflow 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 instantaneousflow 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 andflow 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 byfindingfive 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.

6.2 Feature Selection - Reducing Covariate Shift for Lab to Field Generalization

Covariate shift refers to a significant difference between the lab andfield feature distributions. This difference can result in decreased generalization performance of models trained on lab data to afield setting. While several methods exist to address covariate shift in the independent classification setting (e.g. [40]), 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, [40] used the following importance weighted logistic regression model:

$$\underset{x}{\operatorname{argmin}} \sum_{i=1}^{N} \delta(y_i, x_i) \log(1 + \exp(-y_i(w^T x_i + w_0))) + \lambda \left\| w \right\|^2 \quad (5)$$

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

$$\delta_i(y_i, x_i) = 1/(1 - Q(x_i)) \quad (6)$$

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 andfield 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 andfield data and plotted the distribution of weighted sums of feature vectors. Figure 8a shows this distribution for all features and Figure 8b shows this distribution for selected features.

6.3 Resampled Lab Data - Handling Prior Probability Shi

The way participants spent time within conversations in lab environment may not be representative of their behavior in thefield. Figure 9 shows the amount of time participants

spend in conversation activities in the lab andfield. A smaller fraction of time is spent in conversation in thefield (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 infield. 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 infield: 1) cohabiting couples were recruited to maximize conversational interaction; 2) most of the couples conducted theirfield 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).

6.4 Conversation in Presence of Activity

Data collected in lab typically exercises a very limited number of contexts relative tofield 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 [40], 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 8c).

7 EMPIRICAL PROTOCOLS

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

7.1 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.

7.2 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.

7.3 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 randomfield model (CRF-LC). All models are trained using max-margin learning and all models include ℓ_2 regularization on the parameters [57].

For all models, the regularization strength parameter,λ was tuned over a logarithmic grid, $\{10^{-1}, 10^{0},\ldots, 10^{5}\}$, 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.

7.4 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 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).

8 RESULTS

8.1 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 10.

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 4 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.

8.2 Experiment 2: Conversation Detection in the Field

In order to test the various feature selection and data augmentation methods proposed in Section 6 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).

8.2.1 Performance using lab data trained on all features.—The lab data model trained with all features can identify the conversation episodes infield with an accuracy of 52.03% (Figure 11). The precision and recall is 43.02% and 97.02%, respectively.

8.2.2 Performance using lab data trained on selected features.—Deploying the lab model trained with selected features that reduce covariate shift from lab tofield data, the conversation episode detection accuracy infield is 60.8%, precision is 58.6% and recall is 98.01% (Figure 11) while the false positive rate is 87.5%. Thus, feature selection method has improved the accuracy by 8.8% infield. The F1 score is 0.72 for this model.

However, in comparison with the performance with lab data, conversation episode detection accuracy drops from 95.9% (see Figure 10) to 58.6% on thefield data using this model. Still there is a large gap of performance between lab andfield.

8.2.3 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% infield. 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% infield.

8.2.4 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 thefield. 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%.

8.2.5 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 onfield 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). 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.

9 DISCUSSION, LIMITATIONS, AND FUTURE WORKS

In this work, we used a dedicated audio device to capture long-duration, high-quality audio throughout the day and compared our model output with the ground-truth derived from the audio data. Future work may additionally collect smartphone microphone data as well to have a three-way comparison to understand the extent of loss in accuracy due to energy-efficient subsampling and audio occlusion due to the recorder being in a pocket or purse.

To the best of our knowledge, this is thefirst model to show feasibility of respiration cycle based conversation modeling withfield validation against audio ground truth with promising cross-subject test accuracy (71.7%). Accuracy can be further improved by reducing false positive rate infield deployment in several ways.

Previous research on respiration cycle based smoking detection [51] reduced false positive rates infield data by incorporating an additional sensor modality, i.e., tracking hand-to-mouth gesture via a wrist-worn sensor. We could similarly combine respiration data with hand gesture data to capture gesturing during conversations.

Future studies can also incorporate personality information (e.g., extrovert vs introvert) and optimize the model parameters to further reduce false positive rate infield. Although in the lab training data each participant generally contributed equally, their speech rates could be be quite different in real life as a function of personality factors.

Since respiration based stress-relaxation devices are emerging in the market for daily use (e.g., Spire, Prana, Bellabeat Leaf Urban) [3], respiratory cycle based social interaction modeling should significantly improve and expand the capabilities of such devices. For example, our model enables analysis of stress due to speech planning and unsuccessful attempts to take turns within a conversation, which can provide richer contexts for interpretation. Such models can help assess whether an interaction is stressful or soothing, and help indicate how a user could improve their interaction behaviors (e.g., turn-taking, turn-yielding) to ease the conversation for herself and the conversation partner, to make interactions more enjoyable and productive for both. Combining stress and conversation patterns detected from the respiratory signals may also improve assessments and and treatments for depression in users via real-time intervention through mobile devices.

10 CONCLUSION

This paper presented a conversation episode identification model from respiration signals by classifying each breathing cycle into speech and non-speech. Audio captured in thefield 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 thefield setting hadn’t been realized. This work can contribute a comprehensive approach to processing of respiration data in thefield setting and lead to momentary detection of various daily behaviors from respiration data and enhance the growing utility of respiration sensing.

ACKNOWLEDGMENTS

The authors would like to thank volunteers who participated in the study and Soujanya Chatterjee from University of Memphis who helped to setup data analysis environment. The authors would also like to acknowledge advice from Dr. Benjamin Marlin from UMass Amherst on model development. The authors acknowledge support by the National Science Foundation under award numbers IIS-1722646, ACI-1640813, CNS-1212901 and IIS-1231754 and by the National Institutes of Health under grants R01CA190329, R01MD010362, R01DE025244, UG1DA040309, UH2DA041713, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative.

REFERENCES

[1]. 2017 Mobile Health News. http://www.mobihealthnews.com/content/31-new-digital-health-tools-showcased-ces-2017. (Accessed: May 2017).

[2]. 2017 Philips Watch. http://www.usa.philips.com/c-m-hs/health-programs/health-watch. (Accessed: May 2017).

[3]. 2017 Stress Beating Tech. https://www.wareable.com/wearable-tech/stress-beating-tech-to-keep-you-sane. (Accessed: May 2017).

[4]. 2017 LENA Research Foundation. http://www.lenafoundation.org//. (Accessed: May 2017).

[5]. 2018 GigaOm. https://gigaom.com/2013/09/20/could-a-breath-monitoring-headset-improve-your-health/. (Accessed: January 2018).

[6]. 2018 BioHarness Zephyr. https://www.zephyranywhere.com/. (Accessed: January 2018).

[7]. Adams Roy J, Parate Abhinav, and Marlin Benjamin M. 2016 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 334–343.

[8]. Adib Fadel, Mao Hongzi, Kabelac Zachary, Katabi Dina, and Miller Robert C. 2015 Smart homes that monitor breathing and heart rate. In ACM CHI.

[9]. Ahmed Mohsin Y, Kenkeremath Sean, and Stankovic John. 2015 Socialsense: A collaborative mobile platform for speaker and mood identification. In European Conference on Wireless Sensor Networks.

[10]. Anderson Anne H, Bader Miles, Bard Ellen Gurman, Boyle Elizabeth, Doherty Gwyneth, Garrod Simon, Isard Stephen, Kowtko Jacqueline, McAllister Jan, Miller Jim, et al. 1991 The HCRC map task corpus. Journal of Language and speech (1991).

[11]. Aran Oya and Gatica-Perez Daniel. 2011 Analysis of group conversations: Modeling social verticality In Computer Analysis of Human Behavior.

[12]. Daluwatte Chathuri, Scully Christopher G, Kramer George C, and Strauss David G. 2015 A robust detection algorithm to identify breathing peaks in respiration signals from spontaneously breathing subjects. In Computing in Cardiology.

[13]. Schüll Natasha Dow. 2016 Sensor technology and the time-series self. continent. (2016).

[14]. Duncan Starkey. 1972 Some signals and rules for taking speaking turns in conversations. APA personality and social psychology (1972).

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

[16]. Finkel Jenny Rose, Kleeman Alex and Manning Christopher D. 2008 Efficient, Feature-based, Conditional Random Field Parsing.. In ACL, Vol. 46 959–967.

[17]. Fuchs Susanne, Petrone Caterina, Krivokapić Jelena, and Hoole Philip. 2013. Acoustic and respiratory evidence for utterance planning in German. Journal of Phonetics (2013).

[18]. Gao Ju, Ertin Emre, Kumar Santosh, and al'Absi Mustafa. 2013 Contactless sensing of physiological signals using wideband RF probes. In Asilomar Conference on Signals, Systems and Computers.

[19]. George R, Vedam SS, Chung TD, Ramakrishnan V, and Keall PJ. 2005 The application of the sinusoidal model to lung cancer patient respiratory motion. Medical physics (2005).

[20]. Hao Tian, Bi Chongguang, Xing Guoliang, Chan Roxane, and Tu Linlin. 2017 MindfulWatch: A Smartwatch-Based System For Real-Time Respiration Monitoring During Meditation. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 3 (2017), 57.

[21]. Hixon Thomas J, Goldman Michael D, and Mead Jere. 1973 Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung. Journal of Speech, Language, and Hearing Research (1973).

[22]. Hoit Jeannette D and Lohmeier Heather L. 2000 Influence of continuous speaking on ventilation. Journal of Speech, Language, and Hearing Research (2000).

[23]. Hovsepian Karen, al'Absi Mustafa, Ertin Emre, Kamarck Thomas, Nakajima Motohiro, and Kumar Santosh. 2015 cStress: towards a gold standard for continuous stress assessment in the mobile environment. In ACM UbiComp.

[24]. Kim Taemie, Chang Agnes, Holland Lindsey, and Pentland Alex Sandy. 2008 Meeting mediator: enhancing group collaborationusing sociometric feedback. In Proceedings of the ACM conference on Computer supported cooperative work.

[25]. Koller Daphne and Friedman Nir. 2009 Probabilistic graphical models: principles and techniques. MIT press.

[26]. Kumar S, Abowd GD, Abraham WT, al'Absi M, Beck JG, Chau DH, Condie T, Conroy DE, Ertin E, Estrin D, Ganesan D, Lam C, Marlin B, Marsh CB, Murphy SA, Nahum-Shani I, Patrick K, Rehg JM, Sharmin M, Shetty V, Sim I, Spring B, Srivastava M, and Wetter DW. 2015 Center of excellence for mobile sensor Data-to-Knowledge (MD2K). JAMIA (2015).

[27]. Kumar Santosh, al'Absi Mustafa, Beck J, Ertin Emre, and Scott M. 2014 Behavioral monitoring and assessment via mobile sensing technologies Behavioral Healthcare Technol.: Using Science-Based Innovations to Transform Practice (2014).

[28]. Lafferty John, McCallum Andrew, and Pereira Fernando C N. 2001 Conditional randomfields: Probabilistic models for segmenting and labeling sequence data. (2001).

[29]. Lari Karim and Young Steve J. 1990 The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer speech & language 4, 1 (1990), 35–56.

[30]. Lee Youngki, Min Chulhong, Hwang Chanyou, Lee Jaeung, Hwang Inseok, Ju Younghyun, Yoo Chungkuk, Moon Miri, Lee Uichin, and Song Junehwa. 2013 Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In ACM MobiSys.

[31]. Long Xi, Yang Jie, Weysen Tim, Haakma Reinder, Foussier Jérôme, Fonseca Pedro, and Aarts Ronald M. 2014 Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. Physiological measurement (2014).

[32]. Lopez-Meyer Paulo and Sazonov Edward. 2011 Automatic breathing segmentation from wearable respiration sensors. In IEEE ICST.

[33]. Lu Hong, Brush AJ Bernheim, Priyantha Bodhi, Karlson Amy K, and Liu Jie. 2011 Speakersense: Energy efficient unobtrusive speaker identification on mobile phones. In Pervasive Computing.

[34]. Lu Hong, Frauendorfer Denise, Rabbi Mashfiqui, Schmid Mast Marianne, Chittaranjan Gokul T, Campbell Andrew T, Gatica-Perez Daniel, and Choudhury Tanzeem. 2012 StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In ACM UbiComp.

[35]. Lu Hong, Pan Wei, Lane Nicholas D., Choudhury Tanzeem, and Campbell Andrew T.. 2009 SoundSense: scalable sound sensing for people-centric applications on mobile phones. In ACM MobiSys.

[36]. Lu Wei, Nystrom Michelle M, Parikh Parag J, Fooshee David R, Hubenschmidt James P, Bradley Jeffrey D, and Low Daniel A. 2006 A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms. Medical physics (2006).

[37]. Lukočius Robertas, Virbalis Juozapas Arvydas, Daunoras Jonas, and Vegys Algis. 2015 The respiration rate estimation method based on the signal maximums and minimums detection and the signal amplitude evaluation. Elektronika ir Elektrotechnika (2015).

[38]. McFarland David H. 2001 Respiratory markers of conversational interaction. Journal of Speech, Language, and Hearing Research (2001).

[39]. Mengis Jeanne and Eppler Martin J. 2005 Understanding and enabling knowledge sharing in conversations: a literature review and management framework. In Citeseer KMAP.

[40]. Natarajan Annamalai, Angarita Gustavo, Gaiser Edward, Malison Robert, Ganesan Deepak, and Marlin Benjamin M. 2016 Domain adaptation methods for improving lab-to-field generalization of cocaine detection using wearable ECG. In ACM UbiComp.

[41]. Nguyen Thai, Adams Roy J., Natarajan Annamalai, and Marlin Benjamin M.. 2016 Parsing Wireless Electrocardiogram Signals with Context Free Grammar Conditional Random Fields. In IEEE Wireless Health.

[42]. Nukaya Shoko, Shino Toshihiro, Kurihara Yosuke, Watanabe Kajiro, and Tanaka Hiroshi. 2012. Noninvasive bed sensing of human biosignals via piezoceramic devices sandwiched between thefloor and bed. IEEE Sensors journal 12, 3 (2012), 431–438.

[43]. Olguín Daniel Olguín and Pentland Alex. 2010 Assessing group performance from collective behavior. In CSCW.

[44]. Penne Jochen, Schaller Christian, Hornegger Joachim, and Kuwert Torsten. 2008. Robust real-time 3D respiratory motion detection using time-of-flight cameras. International Journal of Computer Assisted Radiology and Surgery 3, 5 (2008), 427–431.

[45]. Perez-Macias Jose M, Jimison Holly, Korhonen Ilkka, and Pavel Misha. 2014 Comparative assessment of sleep quality estimates using home monitoring technology. In IEEE EMBC.

[46]. Rachuri Kiran K, Musolesi Mirco, Mascolo Cecilia, Rentfrow Peter J, Longworth Chris, and Aucinas Andrius. 2010 EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In ACM UbiComp.

[47]. Rahman Md. Mahbubur, Ali Amin, Plarre Kurt, al'Absi Mustafa, Ertin Emre, and Kumar Santosh. 2011 mConverse: Inferring Conversation Episodes from Respiratory Measurements Collected in the Field. In ACM Wireless Health.

[48]. Rahman Tauhidur, Adams Alexander T, Ravichandran Ruth Vinisha, Zhang Mi, Patel Shwetak N, Kientz Julie A, and Choudhury Tanzeem. 2015 Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In ACM UbiComp.

[49]. Rochet-Capellan Amélie and Fuchs Susanne. 2013 The interplay of linguistic structure and breathing in German spontaneous speech. In Interspeech.

[50]. Rochet-Capellan Amélie and Fuchs Susanne. 2014. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. Philosophical Transactions of the Royal Society of London (2014).

[51]. Saleheen Nazir, Ali Amin Ahsan, Hossain Syed Monowar, Sarker Hillol, Chatterjee Soujanya, Marlin Benjamin, Ertin Emre, al'Absi Mustafa, and Kumar Santosh. 2015 puffMarker: a multi-sensor approach for pinpointing the timing offirst lapse in smoking cessation. In ACM UbiComp.

[52]. Suh Yelin, Dieterich Sonja, Cho Byungchul, and Keall Paul J. 2008 An analysis of thoracic and abdominal tumour motion for stereotactic body radiotherapy patients. Physics in medicine and biology (2008).

[53]. Sung Jaeyong, Ponce Colin, Selman Bart, and Saxena Ashutosh. 2012 Unstructured Human Activity Detection from RGBD Images. In IEEE Robotics and Automation.

[54]. Tang Kevin, Fei-Fei Li, and Koller Daphne. 2012 Learning Latent Temporal Structure for Complex Event Detection. In IEEE CVPR.

[55]. Taskar Ben, Klein Dan, Collins Michael, Koller Daphne, and Manning Christopher D. 2004 Max-Margin Parsing.. In EMNLP, Vol. 1 Citeseer, 3.

[56]. Tehrany R. 2015 Speech breathing patterns in health and chronic respiratory disease. Ph.D. Dissertation.

[57]. Tsochantaridis Ioannis, Joachims Thorsten, Hofmann Thomas, and Altun Yasemin. 2005 Large margin methods for structured and interdependent output variables. In Journal of Machine Learning Research.

[58]. Waber Benjamin N, Olguin Daniel Olguin, Kim Taemie, and Pentland Alex. 2010 Productivity through coffee breaks: Changing social networks by changing break structure. (2010).

[59]. Wang Hao, Zhang Daqing, Ma Junyi, Wang Yasha, Wang Yuxiang, Wu Dan, Gu Tao, and Xie Bing. 2016 Human respiration detection with commodity wifi devices: do user location and body orientation matter?. In ACM UbiComp.

[60]. Wang Y-T, Green Jordan R, Nip Ignatius SB, Kent Ray D, and Kent Jane Finley. 2010 Breath group analysis for reading and spontaneous speech in healthy adults. Folia Phoniatrica et Logopaedica (2010).

[61]. Whalen Doug H and Kinsella-Shaw Jeffrey M. 1997 Exploring the relationship of inspiration duration to utterance duration. Phonetica (1997).

[62]. AJ Wilson CI Franks, and IL Freeston. 1982. Algorithms for the detection of breaths from respiratory waveform recordings of infants. Medical and Biological Engineering and Computing (1982).

[63]. Winkworth Alison L, Davis Pamela J, Adams Roger D, and Ellis Elizabeth. 1995 Breathing patterns during spontaneous speech. Journal of Speech, Language, and Hearing Research (1995).

[64]. Xu Chenren, Li Sugang, Liu Gang, Zhang Yanyong, Miluzzo Emiliano, Chen Yih-Farn, Li Jun, and Firner Bernhard. 2013 Crowd++: unsupervised speaker count with smartphones. In ACM UbiComp.

CCS Concepts: • **Human-centered Computing** → **Ubiquitous and Mobile Computing; • Information Systems** → *Data Mining*;

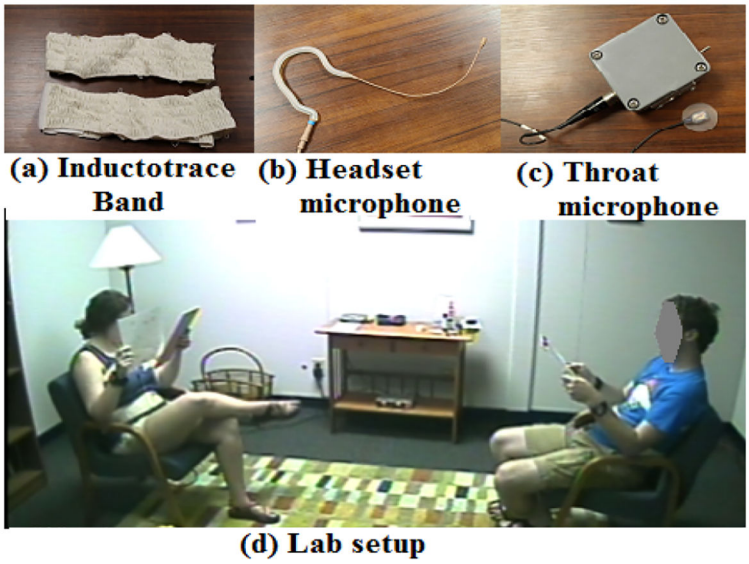
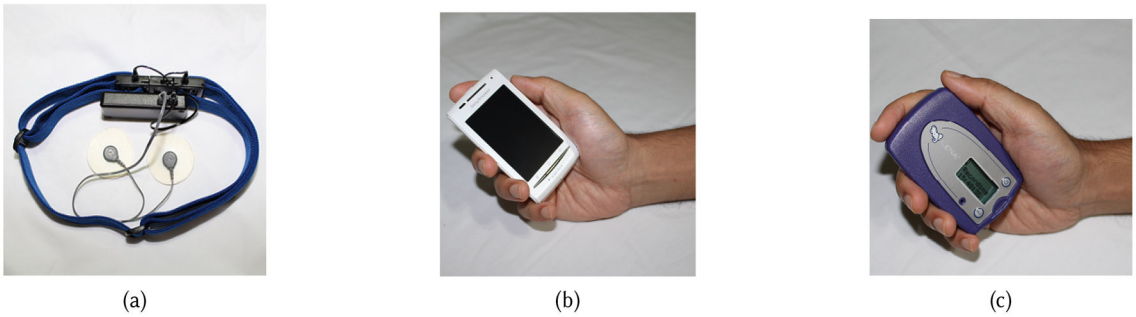


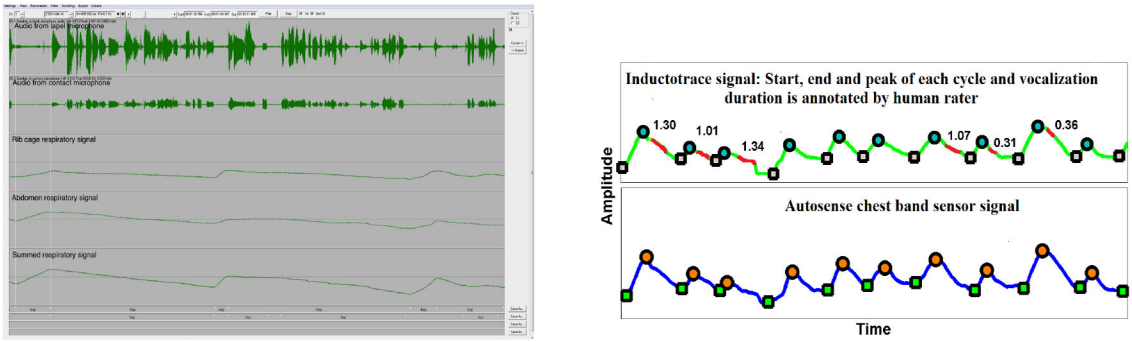
Fig. 1.
Lab equipment and lab setup.



(a) (b) (c)

Fig. 2.

(a) Chest band sensor. (b) Study phone (Sony Ericsson Xperia X10, Android Smart phone). (c) LENA audio recorder.



(a) (b)

Fig. 3.

(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.

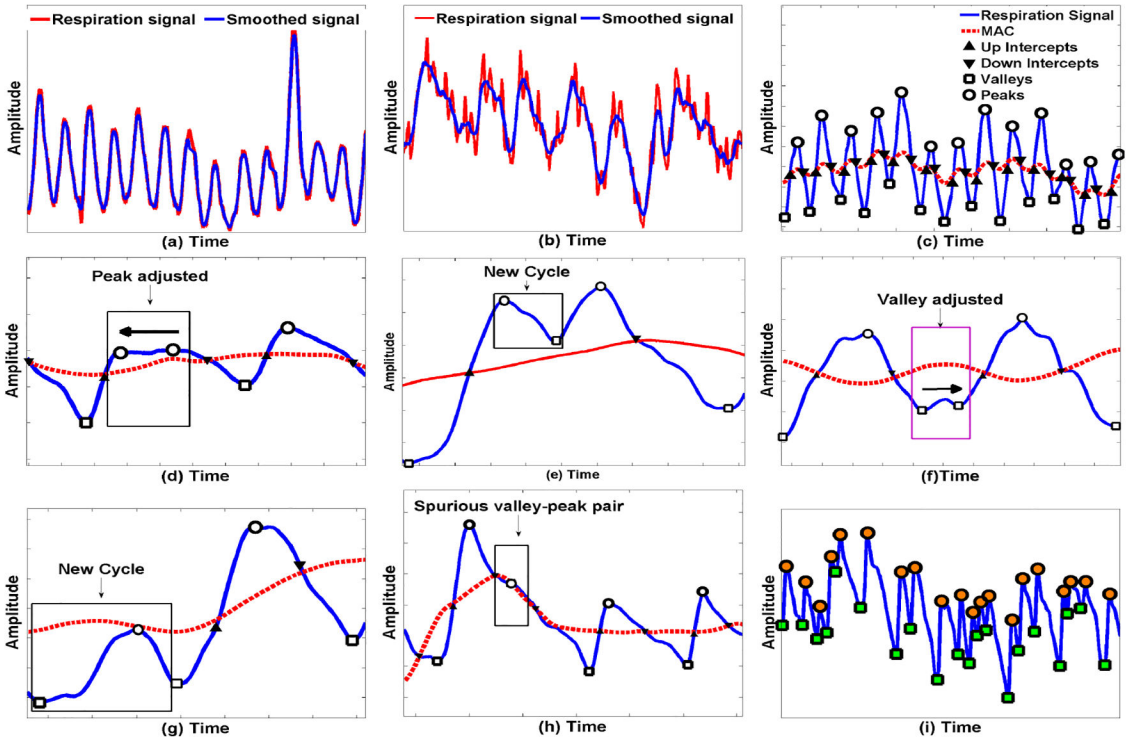


Fig. 4.

(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.

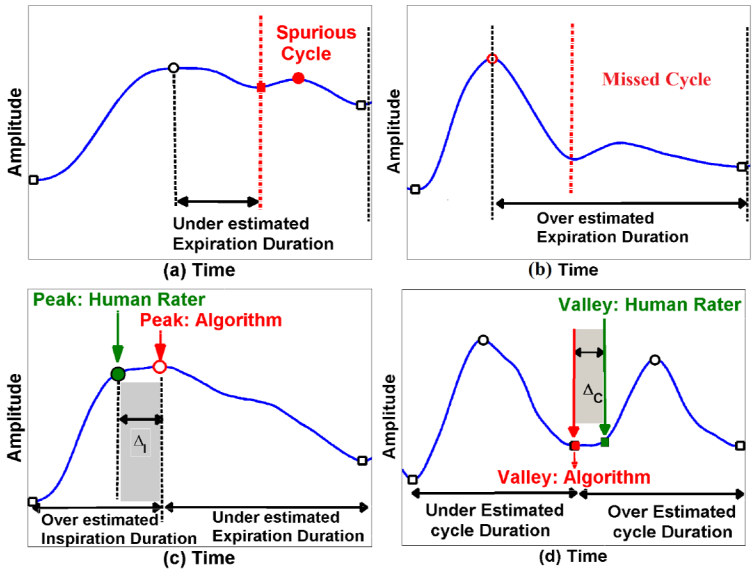


Fig. 5.

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.

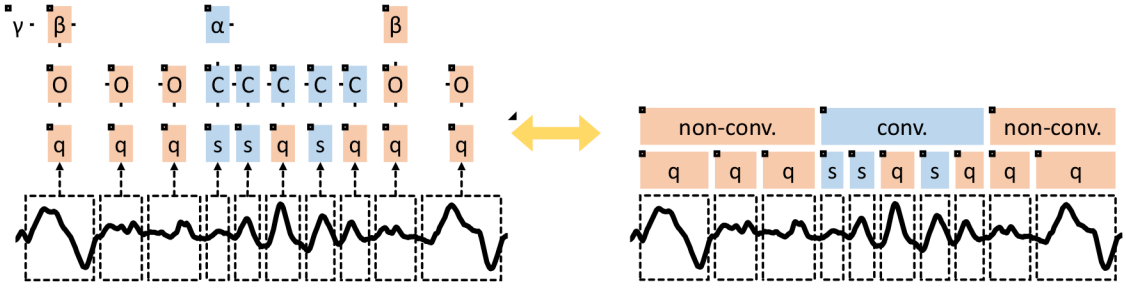


Fig. 6.

An example parse (left) using the grammar described in equation 4. Also shown is the mapping from the parse to a labeled segmentation (right) where q and s stand for quiet and speaking respectively.

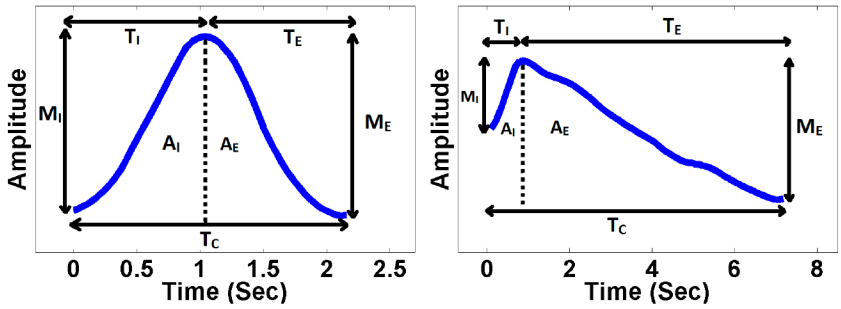


Fig. 7.

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.

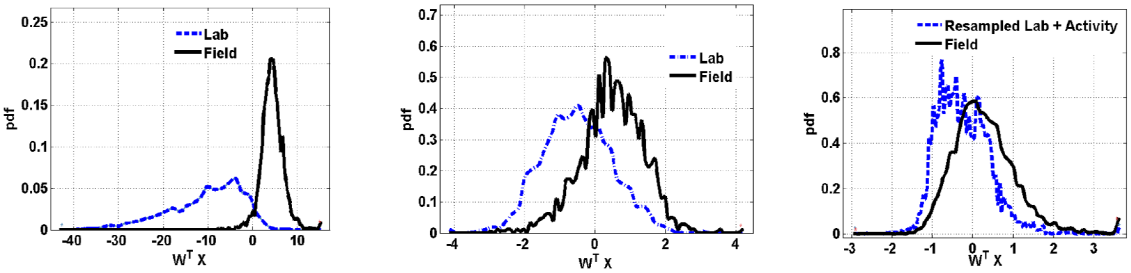


Fig. 8.

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

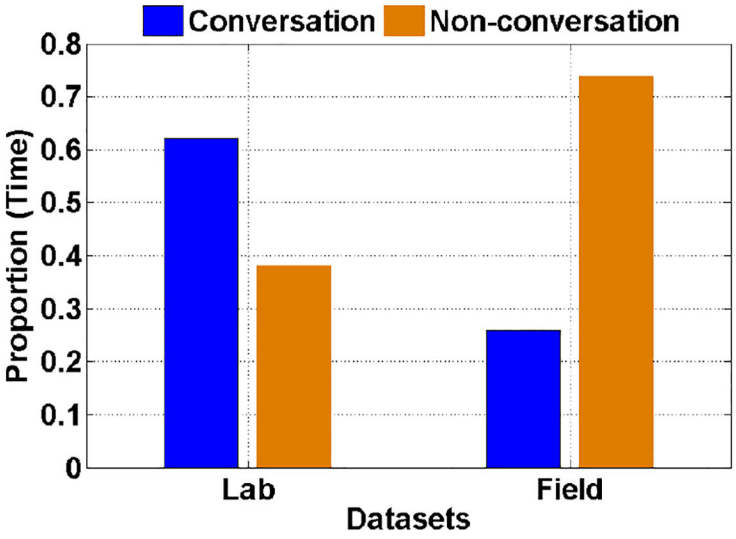


Fig. 9.

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

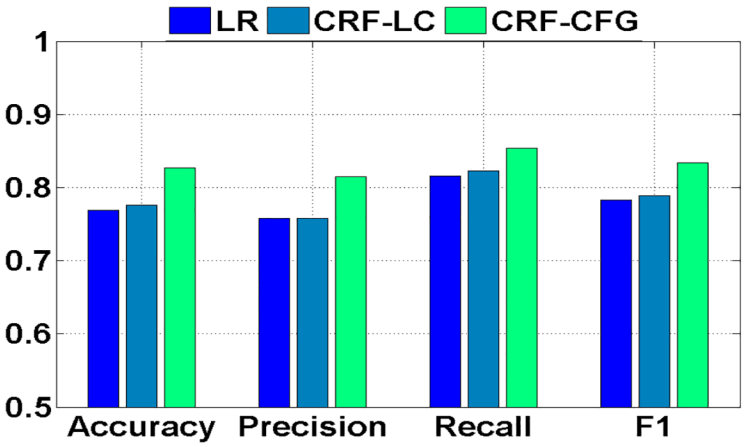


Fig. 10.

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

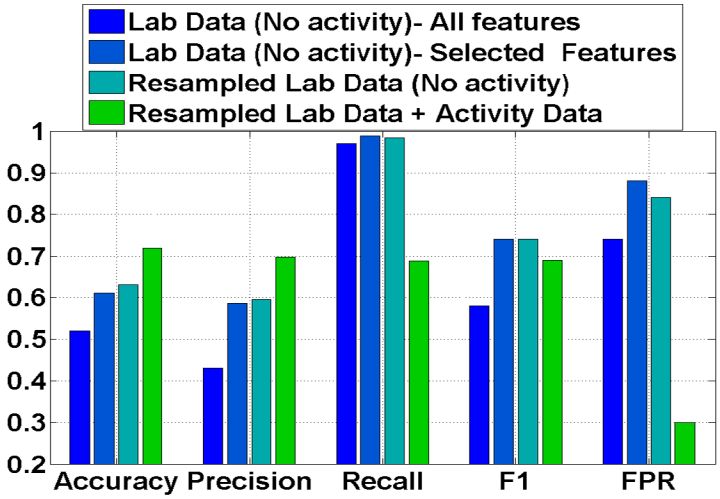


Fig. 11.

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 beer performance (higher accuracy, lower false positive rate) over other models to detect conversation episodes on field data.

Table 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

Table 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 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

Table 4.

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

Table 5.

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