StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones

Hong Lu
Intel Lab
hong.lu@intel.com

Mashfiqui Rabbi
Cornell University
ms2749@cornell.edu

Gokul T. Chittaranjan
EPFL
gokul.thattaguppa@idiap.ch

Denise Frauendorfer
University of Neuchâtel
denise.frauendorfer@unine.ch

Marianne Schmid Mast
University of Neuchâtel
marianne.schmid@unine.ch

Andrew T. Campbell
Dartmouth College
campbell@cs.dartmouth.edu

Daniel Gatica-Perez
Idiap and EPFL
gatica@idiap.ch

Tanzeem Choudhury
Cornell University
tanzeem.choudhury@cornell.edu

ABSTRACT

Stress can have long term adverse effects on individuals' physical and mental well-being. Changes in the speech production process is one of many physiological changes that happen during stress. Microphones, embedded in mobile phones and carried ubiquitously by people, provide the opportunity to continuously and non-invasively monitor stress in real-life situations. We propose *StressSense* for unobtrusively recognizing stress from human voice using smartphones. We investigate methods for adapting a one-size-fits-all stress model to individual speakers and scenarios. We demonstrate that the StressSense classifier can robustly identify stress across multiple individuals in diverse acoustic environments: using model adaptation StressSense achieves 81% and 76% accuracy for indoor and outdoor environments, respectively. We show that StressSense can be implemented on commodity Android phones and run in real-time. To the best of our knowledge, StressSense represents the first system to consider voice based stress detection and model adaptation in diverse real-life conversational situations using smartphones.

Author Keywords

mHealth, stress, sensing, user modeling, model adaptation

ACM Classification Keywords

H.1.2 User/Machine Systems; I.5 Pattern Recognition; J.3 Life and Medical Sciences: Health.

General Terms

Algorithms, Design, Human Factors, Performance

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.
Copyright 2012 ACM 978-1-4503-1224-0/12/09...$10.00.

INTRODUCTION

Stress is a universally experienced phenomenon in our modern lives. According to a 2007 study by the American Psychological Association, three quarters of Americans experience stress-related symptoms [1]. Studies have shown that stress can play a role in psychological or behavioral disorders, such as depression, and anxiety [2]. The amount of cumulative stress in daily life may have broad consequences on societal well-being, as stress-causing events have negative impact upon daily health and mood [2] and also contributes significantly to health care costs [3].

Because stress imparts negative public health consequences, it is advantageous to consider automatic and ubiquitous methods for stress detection. Ubiquitous stress detection can help individuals become aware of and manage their stress levels. Meanwhile, distributed stress monitoring may allow health professionals the ability to examine the extent and severity of stress across populations.

Many physiological symptoms of stress may be measured with sensors, e.g., by chemical analysis, skin conductance readings, electrocardiograms, etc. However, such methods are inherently intrusive upon daily life, as they require direct interaction between users and sensors. We therefore seek less intrusive methods to monitor stress. Researchers have widely acknowledged that human vocal production is influenced by stress [4, 5, 6, 7]. This fact poses the human voice as a potential source for nonintrusive stress detection. In this paper, we suggest that smartphones and their microphones are an optimal computer-sensor combination for the unobtrusive identification of daily stress.

To be operational in real life, a voice-based stress classifier needs to deal with both the diverse acoustic environments encountered everyday and the individual variabilities of portraying stress. Most existing research relating stress and speech has focused on a single acoustic environment using high-quality microphones. This paper presents a method for detecting the occurrence of stress using smartphone microphones and adapting universal models of stress to specific individuals or scenarios using Maximum A Posteriori

(MAP) adaptation. The contributions of this paper are as follows. We experimentally show that: 1) Stress from human voice can be detected using smartphones in real life acoustic environments that involve both indoor and outdoor conversational data. 2) A universal stress model can be robustly adapted to specific individual users, thereby increasing the accuracy across population of users. 3) A stress model can be adapted to unseen environments, thereby lowering the cost of training stress models for different scenarios. 4) The proposed stress classification pipeline can run in real-time on off-the-shelf Android smartphones.

BACKGROUND AND RELATED WORK

In this section, we describe how stress and stressors are modeled, as well as provide discussion on the differences between stress and emotion detection from speech. We discuss individual variability in portraying stress and why adaptation or personalization of models is necessary.

What is Stress?

In general terms, stress is the reaction of an organism to a change in its equilibrium. However a more quotidian definition is that stress is the tension one experiences in response to a threat. When a person is able to cope successfully with stress they experience eustress, the opposite of distress. Stress can therefore have positive or negative outcomes, dependent on a person's coping ability and the severity of the stressor. Stressors may be real or imagined; an event that produces stress for one individual may have no affect on another individual.

Stressors have been quantified in terms of "hassles" and "life events," the former referring to short-term intraday stress-causing events and the latter referring to larger events of sparser and more momentous occasion [8]. Two main categories of stressors are *physical* or *psychological*. Physical stressors are those which pose a threat to a person's physical equilibrium, such as a roller coaster ride, a physical altercation, or deprivation of sleep. A psychological stressor can be an event that threatens a person's self-esteem, e.g., the pressure of solving a difficult mental task within a time limit.

Stress is a subjective phenomenon, and stressors are the observable or imagined events and stimuli that cause stress. *In this paper, we focus on cognitive stress and estimating the frequency of stressors, rather than the severity of stress*. The reasons for this are severalfold. No universal metric exists for stress severity in the psychology literature. By definition, stress cannot exist without stressors. While it is difficult to objectively compare stress across individuals, research has shown that there exist sets of stressors that are shared by people. Finally, we simplify our objective by not determining the type of stress that an individual is experiencing, or their coping ability. We believe that mobile detection of the frequency of stressors in one's life is a more practical and realistic goal than determining the severity of stress.

People often react to stress emotionally. For instance, a person undergoing distress from an argument may experience anger or sadness. Stress detection in speech is therefore often wrapped up in speech emotion detection. Much of the literature treats stress and emotion as conjugate factors for speech production. While emotion detection in speech has been examined by a significant scientific population, speech under stress has received focussed attention by a smaller group of researchers. We specifically aim at modeling speech under stress, rather than emotional expression. In our view, a critical difference between stress and emotion detection is that stress can always be linked to a stressor, whereas it is not always possible to establish causal relationships between emotions and events.

Modeling and Detecting Stress From Speech

Pioneering work on voice analysis and cognitive stress was done in 1968 when amplitude, fundamental frequency, and spectrograms were analyzed for speech under task-induced stress [4]. It was found that neutral listeners could perceive differences in voice recordings of subjects undergoing stress-activating cognitive tasks.

A large body of research on stress detection from speech centered around the Speech Under Simulated and Actual Stress (SUSAS) dataset [9]. Evidence from experiments with SUSAS suggests that pitch plays a prominent role in stress. Other important features are those using energy, spectral characteristics of the glottal pulse, and phonetic variations (speaking rate) and spectral slope. Nonlinear features based on the Teager Energy Operator [10, 11] have also shown promising discriminative capabilities, especially for talking styles such as "anger" and "loud" [12].

Fernandez and colleagues performed voice-based stress classification in a scenario involving both physically and physiologically induced stress [13]. Four subjects were asked to answer mathematical questions while driving a car simulator at two different speeds and with two different response intervals. Several models were tested on two levels of speech analysis: intra-utterance and whole utterance. The best results were found by using a mixture of hidden Markov models at the intra-utterance level. Although this work is a comprehensive study of several models for speech under stress the data collection is done in a quiet environment using a high-quality professional microphone.

Paltial et al. propose a Gaussian mixture model based framework for physical stress detection [14]. Stress is induced by execrising on a stair-stepper at 9-11 miles per hour. They investigate the effect of number of speakers in the training set that consists of all female subjects. The data is recorded in a single acoustic environment using professional grade microphone. Adaboost is applied to combine the mel-scale cepstral coefficients and Teager energy operator to achieve a 73% classification accuracy with a generic stress model.

Mobile Stress Detection

There has been growing interest in inferring stress and emotion from mobile phone sensor data. Chang, Fisher, and Canny describe a speech analysis library for mobile phone voice classification in [15]. They implement an efficient voice feature extraction library for mobile phones and show

the processing is able to run on off-the-shelf mobile phones in realtime. However, their work does not consider stress classification in mobile scenarios, mixed acoustic contexts, or any form of adaptation. No data is collected from mobile phones. The evaluation is done solely on the SUSAS data set. EmotionSense [16] presents a multi-sensor mobile system for emotion detection. The system is specifically designed for social psychology researchers, and includes several parameterizations toward that end. A universal background model (UBM) is trained on all of the emotion categories of an artificial dataset of portrayed emotions. Specific emotion classifiers are subsequently generated by maximum a posteriori (MAP) adapting from the UBM to specific emotions in the dataset. Unlike our work, EmotionSense does not train its emotion classifier on actual user data. Rather, their classifiers are trained from the artificial dataset and kept static. Again, there is no verification of the classifiers' accuracies on audio collected in the wild, nor is there any personalization of the universal models to individual users.

As Scherer [17] states, there is "little evidence for a general acoustic stress profile" because correlation between voice and human emotion is subject to large differences between individuals. Humans respond to the same stressors in different ways, and they also respond differently from each other depending on their coping ability, coping style, and personality. It is therefore imperative to develop a method of stress detection that can model individual speakers in different scenarios. This evidence leads us to believe that no universal model of stress will be effective for all individuals or all scenarios. However, it is cumbersome to learn a new model of stress for every individual or scenario. It is therefore advantageous to have a means for adapting models. Finally, most stress detection research has not considered diverse acoustic environments. We believe it is important to evaluate classifiers trained in more realistic acoustic environments, and through microphones in existing mobile devices.

EXPERIMENTAL SETUP

Data Collection

We study stress associated with the cognitive load experienced by a participant during a job interview as an interviewee and conducting a marketing task as an employee. We also consider a neutral task where participants are not under stress. These three tasks are designed with the help of behavioral psychologists. As SUSAS and many previous studies [9, 4, 13, 14], we assume that the subject's voice are stressed once the stressor is present. And reading without stressor is neutral. Participants were recruited through flyers advertised at multiple locations within a university campus and through messages shared using social media sites. Data is collected from a total of 14 participants (10 females, 4 males). The mean age was 22.86 years and participants had in average few experience in job interviews (mean experience of 2.86 within a scale from 1 (no experience at all) to 5 (a lot of experience)). Thirteen participants were undergraduate students in different domains such as geology, psychology, biology, and law. One participant was a PhD student. The data collection is done in three phases:

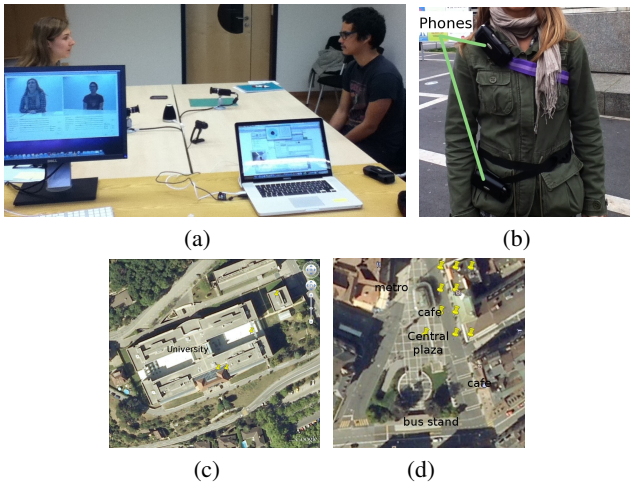

Figure 1: (a) the setup of the interview room. (b) the setup of data collection (c) and (d) map of the university and city centre respectively with yellow markers indicating locations where participants had conversation with other people

1. **Job Interviews:** Job interviews of the participants are conducted indoors, as shown in Fig. 1(a). The subjects are informed that they first have to go through the job interview, and whether they would be hired for the marketing task depends on their interview performance. The interview scenario comprises a structured interview with 8 questions in French; the translation is as follows: i) Can you describe yourself in a few words? ii) What is the motivation for you to apply for this job? iii) What is the meaning and importance of scientific research to you? iv) Can you give an example that shows your competence to communicate? v) Can you give an example of persuading people? vi) Can you give an example of working conscientiously? vii) Can you give an example of mastering a stressful situation? viii) Can you tell me your strengths and weaknesses? This set of questions is designed to test the subjects' competence to do the tasks, especially the last four questions.

   Audio is continuously collected using a Google Nexus One Android smartphone and a microcone[1] microphone array. In addition to audio data, video cameras record the interviewer and interviewee.

2. **Marketing Jobs:** When the interview is complete, participants are then briefed about the marketing job that they have been hired to conduct. The marketing task involves recruiting new participants from the general public for other studies that are frequently conducted at a local university. Each participant is rewarded 200 CHF for about 4 hours work. The marketing task provides us with the opportunity to study participants executing this real-world job at two different locations: (i) at a university campus and, (ii) in the center of a city, as shown in Fig. 1(c) and Fig. 1(d), respectively.The participants are given flyers with contact information for the study and a paper to collect the contact information of interested people. Participants conducting recruitment campaigns wear two Nexus One smartphones positioned at two different places on

[1]http://www.dev-audio.com

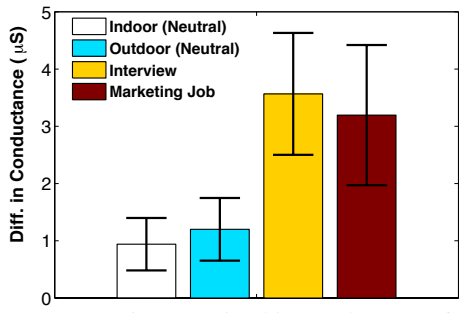

Figure 2: Average increase in skin conductance for tasks

their body, as shown in Fig. 1(b). Participants were informed that their remuneration for the job would have a fixed and performance-based component. The performance-based component would depend on the number of participants they would recruit. This was done to motivate the participants to perform the outdoor task. For each participants, the marketing job was designed to be carried in four independent sessions in different days. However, data for all 4 sessions are not available for all subjects due to technical issues with data collection or subjects failing to attend all the sessions.

3. **Neutral Task:** In addition to capturing stressed audio, we also collect audio data from neutral scenarios where participants are not stressed. In neutral scenarios, participants had to read both indoor and outdoor. The reading materials are simple stories that often used to study different accents in languages. Compared to the job interview situation or to the recruitment there was no performance expected of the participants. Thus, there was no stress induction during this situation and participants were therefore unlikely to be stressed.

The study is done over multiple days - one for the interview, one for the neutral task, two or more for the marketing task. All audio data is collected using Nexus One phones at 8kHz 16 bit PCM using a custom android program that runs in the background. The same program also continuously records accelerometer data and collects GPS data every 3 minutes. During all the tasks, every participant wore a wrist band called Affectiva[2], which includes a galvanic skin resistance (GSR) sensor used for ground-truth, as discussed next.

Ground-truth of Stress from GSR Sensor:

During stress, the human body goes into an alert mode resulting in increased skin conductance, which can be measured by a GSR sensor [2]. We used the Affectiva wrist band to collect skin conductance data. As an external sensor, the GSR sensor is only used for ground truth purpose. For each session, there was a 5 minute calibration phase for the GSR sensor where participants wear the sensor and relax. This calibration phase provides the baseline GSR readings (i.e., skin conductance) for each session. To measure changes in stress level for each session, we compute the difference between the average baseline readings of the calibration phase and the average GSR readings during the actual tasks. Multiple sessions of outdoor marketing job recordings are averaged. Figure 2 shows the average increase of GSR readings in different types of tasks. Clearly, the increase of GSR reading is higher for marketing and job interview sessions than both neutral scenarios (indoor or outdoor). This result suggests that participants are reacting to stressors induced during both the job interview and marketing tasks.

[2]http://www.affectiva.com/

STRESSSENSE CLASSIFICATION MODEL

In what follows, we detail the initial data preprocessing steps and then discuss our StressSense features, classification framework and model adaptation.

Data preprocessing

Prior to stress classification, the audio from the recording sessions are preprocessed using a two step process: (i) audio is classified into voice and non-voice regions and (ii) the individual speakers are segmented based on the output of the voice classification. This process is carried out differently for the indoor and outdoor scenario since we could leverage the additional microphone array that was part of the instrumented room, as shown in Figure 1(a).

In the indoor setting, information from the microcone microphone array is used to segment the speakers. The microcone comprises of 7 microphones in an array and records audio at 16kHz in 16 bit PCM. The microcone manufacturer provides software to do automatic speaker turn segmentation from the recorded audio. We time align the microcone and Android smartphone audio by identifying the first peak in the cross correlation between the center channel of the microcone device and the audio recorded on the phone.

For the outdoor audio, we do not have the advantage of using the microphone array. We use a different classifier to segment voice and non-voice region that has been demonstrated to be robust in outdoor noisy environments [18, 19, 20]. In our test on 4 minutes of labeled audio data including human speech, the classifier yields an accuracy of 83.7%, with precision 90% and recall 84%. This 4 minutes of audio is acquired from our outdoor dataset with high environmental noise, and the above mentioned performance of the classifier is compatible with results found in [18, 19, 20]. For speaker segmentation, the participants wear two smartphones during data collection – one attached to their shoulder and another to their waist, as shown in Fig. 1(b). In the future, we envision utilizing multiple microphones embedded within a single smartphone. To align the two audio streams, mutual information between the voiced regions is used. Direct cross-correlation between audio streams from the microphones is avoided due to outdoor noisy environments (even different rubbing pattern of phones at different positions on the body can cause confusing differences). Mutual information has been successfully used for alignment and conversation detection in outdoor scenarios [19]. Upon alignment, energy comparison among waist and shoulder audio is used for speaker segmentation. We exploit the fact that audio energy is inversely proportional to the distance of the microphone. If person A and person B are in a conversation with person A being instrumented with two phones, A's mouth is at a much shorter distance from the shoulder mi-

Feature	Description
Pitch std	standard deviation of pitch
Pitch range	difference of max and min pitch
Pitch jitter	perturbation in pitch
Spectral centroid	centroid frequency of the spectrum
High frequency ratio	ratio of energy above 500Hz
Speaking rate	rate of speech
MFCCs	cepstral representation of the voice
TEO-CB-AutoEnv	Teager Energy Operator based nonlinear transformation

Table 1: StressSense Acoustic Features

crophone compared to the waist microphone. On the other hand, A's microphones are almost equidistance from B's mouth. Thus when B is speaking the energy ratio between A's two microphones will be close to one. On the other hand, while person A is speaking, the energy ratio between shoulder and waist microphone will be greater than one.

Therefore, audio energy ratio can be used as a discriminatory feature for segmenting the subject from his conversational partners. We use the receiver operating characteristic (ROC) analysis to find an appropriate threshold for energy ratio to differentiate between speakers. We test this threshold based classifier on a total of 20 minutes of manually labeled audio data with labels indicating which speaker is speaking and when. This 20 minutes of audio data is gathered from two separate outdoor conversations in our outdoor dataset that involved distinct participants. On this test data, the classifier yields an accuracy of 81% with precision 81% and recall 95%. After running the speaker segmentation algorithm on our outdoor data set, we find that in most cases participants are talking for 10-20 minutes in each session with the exception of one participant who only talks for 4 minute.

StressSense Features

The most widely investigated acoustic feature for stress is pitch (F0). Pitch reflects the fundamental frequency of vocal cord vibration during speech production. Normally, the mean, standard deviation and range of pitch increase when somebody is stressed [7, 21, 22, 23, 24], while the pitch jitter in voice usually decreases [25, 26]. The degree of change varies from person to person depending on the person's experience and arousal level.

It is known that the locations of formants across the spectrum are different between stressed and neutral speech. Generally, the distribution of spectral energy shifts toward higher frequency when somebody is stressed. This change reflects in spectral centroid which goes up during stressed speech and more energy concentrates at frequencies above 500Hz [25, 26]. In addition to changes in the spectrum energy distribution, previous studies also show an increase in speaking rate under stressed conditions [24, 26]. We adopt the speaking rate estimator proposed in [27], which derives the speaking rate directly from the speech waveform. The processing window is set to one second for better dynamics. More recently, studies on stress analysis show that feature based on a nonlinear speech production model is very useful in stress detection. The TEO-CB-AutoEnv feature [12] is based on a multi-resolution analysis of the Teager Energy profile. It is designed to characterize the level of regularity in the segmented TEO response and can capture variations in excitation characteristics including pitch and its harmonics [12, 28, 13]. It is robust to the intensity of recorded audio and background noise. We adopt the 17 dimension TEO-CB-AutoEnv proposed in [29] which is an extension to the version proposed in [12]. MFCC is a set of acoustic features modeling the human auditory system's nonlinear response to different spectral bands. It is a compact representation of the short-term power spectrum of a sound. It is widely used in speech analysis, such as speech recognition and speaker recognition. As a generic speech feature, MFCCs are used in detecting stressed speech in a number of prior studies [12, 14]. We use 20-dimension MFCCs with the DC component removed since it corresponds to intensity of the sound.

Another widely investigated category of features for stress classification are intensity based, such as, the mean, range and variability of intensity. Intensity based features require consistent control over ambient noise and both distance and orientation of the microphone with respect to the speaker. In a mobile ubiquitous setting, it is impractical to make such strong assumptions about the orientation, body placement, and environmental contexts. Therefore, we do not adopt intensity based features. Also, the intensity of audio clips are normalized to ensure the stress classification is not affected by different intensity levels found in different sessions.

Table 1 summarized the acoustic features used by StressSense. In the feature extraction stage, the audio samples captured by the phone are divided into frames. Each frame is 256 samples (32 ms). It is known that human speech consists of voiced speech and unvoiced speech [30]. Voiced speech is voice generated from periodic vibrations of the vocal chords and includes mostly vowels. In contrast, unvoiced speech does not involve the vocal chords and generally includes consonants. We use only voiced frames for analysis, i.e., features are only computed from voiced frames of speech. The reason for this is twofold. First, the pitch and TEO-CB-AutoEnv features can only be reliably extracted from voiced frames. Second, voiced speech contains more energy than unvoiced speech, thus, it is more resilient to ambient noise. We use the method introduced in [31] where zero crossing rate and spectral entropy are used to select voiced frames. The final dimensionality of feature we extracted is 42.

StressSense Classification

Classification Framework.

Our classification framework uses Gaussian Mixture Models (GMMs) with diagonal covariance matrix. We use one GMM for each of the two classes, i.e., stressed speech and neutral speech. The framework makes decisions according to the likelihood function $p(X|\lambda)$ of each class with equal prior, where X is a feature vectors and $\lambda(w, \mu, \Sigma)$ is a GMM model with weight, mean, and covariance matrix parameters. We choose 16 as the number of components after evaluating the Akaike Information Criterion (AIC) for several models on the subjects. To avoid over fitting the training data, the variance limiting technique [32] is applied with a standard expectation maximization (EM) algorithm to train the GMM speaker models [32]. To initialize the EM algorithm, $k-$Means is used to set the initial means and vari-

ances of the GMM components. We investigate three training schemes to model stress: universal model, personalized model and speaker adapted universal model. In what follows, we discuss these three models in more detail:

- **Universal model** uses a one-size-fits-all approach, where one universal stress classifier is trained for all users. It is the most widely adopted scheme in many mobile inferencing systems due to its simplicity. The universal classifier is static after deployment.
- **Personalized model** uses a completely speaker dependent approach. It requires model training on each individual user's own data and generate a speaker dependent model for each user. Clearly, this scheme is superior to the universal model in terms of performance, but its usability and scalability greatly limits its application. Each user has to label their own data and train their own stress model before they can start using the system. The cumbersome bootstrapping phase and significant computational resource required for this scheme render it infeasible in most practical applications. We study this scheme mainly to use it as a reference for best achievable performance.
- **Model adaptation** represents a middle ground between the previous two schemes. All users start with a universal stress classification model, which in turn gets adapted to each individual user for better performance when more personal data is available as users carry the phone. We design two adaptation methods: i) *supervised adaptation*, where a user explicitly contributes labelled data for adaptation; and ii) *unsupervised adaptation*, which leverages self-train [33] technique by utilizing unlabeled data – note, we refer to this type of adaptation as *self-train* in the remaining paper. Both methods use the Maximum A Posteriori algorithm to adapt the universal model, but are different in the way they assign samples for adaptation.

Model Adaptation

In order to adapt the parameters of a GMM, we use the Maximum A Posteriori (MAP) method developed in [34] for speaker verification. The paradigm of MAP adaptation for speaker recognition is similar to ours for stress detection. Given the variation in stress parameters for individuals and different scenarios (e.g., noisy outdoor), we expect that universal model of stress may be adapted to individual speakers and scenarios. The MAP adaptation is a non-iterative process and is therefore performed only once for each new set of observations. It is a modification of the Maximization step in the EM algorithm. It attempts to maximize the posterior probabilities of the unimodal gaussian components given new observations. We begin with a universal GMM model λ and new training observations (from individuals or scenarios), $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}\}$, where T is the number of observations. For gaussian component i, the component's posterior probability given observation $\boldsymbol{x_t}$ is

$$p(i|\boldsymbol{x_t}) = \frac{w_i p_i(\boldsymbol{x_t})}{\sum_{j=1}^{M} w_j p_j(\boldsymbol{x_t})} \tag{1}$$

Using the posterior probability of component i, we can calculate the sufficient statistics for the weight, mean and variance:

$$n_i = \sum_{t=1}^{T} p(i|\boldsymbol{x_t}),\ E_i(\boldsymbol{x}) = \frac{1}{n_i}\sum_{t=1}^{T} p(i|\boldsymbol{x_t})\boldsymbol{x_t},$$

$$E_i(\boldsymbol{x}\boldsymbol{x'}) = \frac{1}{n_i}\sum_{t=1}^{T} p(i|\boldsymbol{x_t})\boldsymbol{x_t}\boldsymbol{x'_t} \tag{2}$$

The updated parameters of the MAP adapted GMM are calculated using α as follows:

$$\alpha_i = n_i/(n_i + r) \tag{3}$$

$$\hat{w}_i = [\alpha_i n_i/T + (1-\alpha_i)w_i]\gamma \tag{4}$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \alpha_i E_i(\boldsymbol{x}) + (1-\alpha_i)\boldsymbol{\mu_i} \tag{5}$$

$$\hat{\boldsymbol{\sigma}}_{\boldsymbol{i}}^{\boldsymbol{2}} = \alpha_i E_i(\boldsymbol{x}\boldsymbol{x'}) + (1-\alpha_i)(\boldsymbol{\sigma_i^2} + \boldsymbol{\mu_i^2}) - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\boldsymbol{2}}, \tag{6}$$

where r is a relevance factor that determines how much relevancy the original model should hold. We set r to 16, as suggested by [34] . γ is a scaling coefficient that ensures $\sum_i^M \hat{w}_i = 1$.

The two adaptation schemes, supervised adaptation and self-train adaptation (which is unsupervised adaptation) are different in how they acquire new training samples. In case of supervised scheme, new training observations are labeled explicitly by the user. Once the user labels new data, MAP is applied to adapt the universal model. However, the data annotation process is tedious and error-prone, thus, the amount of labeled data might be scarce. The self-train scheme leverages only unlabeled data without any user input. It reuses the predicted label and confidence statistics generated by the universal stress model during the inference process to select new training samples. Given the imperfection of the universal stress model, obviously it is not wise to completely trust its predictions. Therefore, the self-train method determines whether a data sample is suitable for adaptation according to the confidence level of the inference and uses high confidence samples. Furthermore, in most conditions, stress exhibits temporal dependence. Successive voice samples are very likely captured from the same stress status, therefore, the likelihoods of the data samples from the GMM models should be relatively stable. An abrupt change in likelihood sometimes indicates a change of stress conditions, but it can be an outlier. Therefore, the system apply a low pass filter on the output of each GMM to smooth the likelihood.

$$l_{t_i} = \alpha * l_{t_{i-1}} + (1-\alpha) * l'_{t_i}$$

where l' is the original likelihood estimated by the GMM and l is the smoothed likelihood. The weight α determines the balance between the history and incoming data.

Once the likelihood function is estimated, the system uses an entropy based confidence score to estimate the quality of classification. For each data sample, the entropy of normalized likelihoods indicates the confidence level of the inference result.The entropy is computed on the smoothed likelihood $[l_1, l_2]$ for stress and neutral class, respectively. By

$$Entropy = \sum_{i=1}^{2}(l_i/Z) \times \log(l_i/Z),\ \ Z = \sum_{i=1}^{2} l_i$$

	Feature set A				*Feature set A B*				*Feature set A B C*			
	precision	recall	f-score	accuracy	precision	recall	f-score	accuracy	precision	recall	f-score	accuracy
universal	73.1%	58.9%	64.4%	68.6%	69.6%	70.0%	69.5%	69.6%	70.5%	74.8%	72.2%	71.3%
unsupervised	74.9%	63.4%	68.1%	70.8%	75.8%	72.1%	73.7%	74.3%	78.8%	76.8%	77.5%	77.8%
supervised	72.7%	70.3%	71.1%	71.9%	78.5%	80.1%	79.2%	79.0%	79.1%	85.0%	81.8%	81.1%
personalized	72.5%	71.9%	72.1%	72.2%	77.1%	83.1%	79.9%	79.1%	80.5%	87.1%	83.6%	82.8%

Table 2: Stress classification on indoor data

	Feature set A				*Feature set A B*				*Feature set A B C*			
	precision	recall	f-score	accuracy	precision	recall	f-score	accuracy	precision	recall	f-score	accuracy
universal	66.1%	56.0%	60.1%	63.6%	65.7%	63.6%	64.2%	65.5%	67.0%	64.2%	65.2%	66.5%
unsupervised	67.7%	53.4%	58.8%	63.8%	68.3%	62.4%	64.7%	67.1%	76.7%	59.1%	65.7%	70.5%
supervised	66.7%	62.6%	64.2%	65.5%	72.0%	72.0%	71.7%	71.8%	75.6%	75.9%	75.5%	75.5%
personalized	66.6%	65.1%	65.7%	66.1%	72.6%	78.3%	75.2%	74.0%	76.0%	82.4%	78.9%	77.9%

Table 3: Stress classification on outdoor data

where Z is a normalization term. If the universal classifier has a high confidence on the prediction, the entropy will be low. Otherwise, if the normalized likelihoods of the two classes are quite close to each other, the entropy will be high. An entropy threshold is used to control whether a data sample is selected for adaptation. Because unlabeled data is cheap and abundant, the system can use a tight threshold to ensure data quality. Once adaptation data set is selected, the MAP algorithm is applied to adapt the universal model as the supervised scheme.

We test the effectiveness of two adaptation schemes in two use cases: adapting universal model to individual speakers, and adapting universal model trained from one scenario to another. A detail evaluation is presented in the next section.

STRESSSENSE EVALUATION

We conduct three experiments using the dataset and stress models described earlier. First, we evaluate the importance and effectiveness of different vocal features for different acoustic environments. Next, we study the success of stress models trained and tested in specific scenarios, i.e., either fully indoors or fully outdoors. Finally, we investigate the stress classification performance under mixed acoustic environments. The amount of neutral data for each subject is about 3 minutes for both indoor and outdoor scenarios. The amount of stress data for indoor scenario ranges between 4-8 minutes depending on how talkative the subject is during the 10-min interview session. The average amount of indoor stress data is 4.11 mins. For each subject, we also use 4 minutes of speech segments from different outdoor recruiting sessions. For all experiments, the universal models are tested using the leave-one-user-out cross validation method. In cases where adaptation is performed, the target user's data is equally partitioned into an adaptation set and a test set. The adaptation set for a target user is used to customize the universal model learned using training examples that exclude data from that user. Adaptation is done in two ways: (i) supervised adaptation, where the adaptation set includes labels and (ii) self-train, where the adaptation set is unlabeled. The personalized model is evaluated using five-fold cross validation.

Stress Classification in Individual Scenarios

In the first experiment, we measure the relevance of the different features using information gain based feature ranking [35]. We treat the indoor and outdoor scenarios as two separate data sets. Table 4 shows the top 10 features in each scenario. It is clear that pitch features and speaking rate are ranked as the most predictive in both scenarios. MFCC is more relevant in the indoor environment, while TEO-CB-AutoEnv is more useful in outdoor environments. This finding is in line with what Fernandez and Picard observed in [13]; that is, the Teager energy operator is robust to ambient noise. Even though MFCC is able to capture valuable stress related information, it is sensitive to ambient noise by its nature as a generic acoustic feature. Therefore it is less discriminative outdoors. Spectral features (i.e., high frequency energy ratio and frequency centroid) are more discriminative in outdoor scenarios, but not indoor. Since MFCC is able to capture information similar to spectral features, the contribution of spectral features is lower in the indoor scenario.

Rank	*Indoor*	*Outdoor*
1	pitchStd	pitchStd
2	speakingRate	speakingRate
3	pitchRange	pitchRange
4	MFCC4	TEO-CB-AutoEnv17
5	MFCC3	HighFrequencyRatio
6	MFCC14	TEO-CB-AutoEnv7
7	TEO-CB-AutoEnv2	MFCC2
8	MFCC15	TEO-CB-AutoEnv10
9	MFCC19	MFCC1
10	MFCC1	Centroid

Table 4: Feature ranking in different environments

To study the impact of the different features on classification, we divide the 42 features into three groups : A) pitch based features and the two spectral features, B) TEO-CB-AutoEnv feature set, and C) the MFCC feature set. We add them one by one in the indoor and outdoor stress classification tasks. Table 2 and Table 3 show how performance changes as the feature set grows. Adding TEO-CB-AutoEnv to feature set A improves the classification performance significantly particularly for the outdoor scenario. MFCC provides a small additional improvement of 3% on top of feature set A and B. To reproduce speech intelligibly at least the information of three formants are required [36]. Therefore, our features set A and B can't reconstruct the speech. But with C, it is possible. In our future work, we will consider better privacy sensitive features set without using MFCC.

Figure 3 and Figure 4 show the accuracy for each of the subjects for each model when all the features are used. Note that subject 1's indoor data was corrupted due to cellular connectivity issue during the experiment, so we leave the data out for the indoor experiment. The same problem was experienced by subject 5 in the outdoor case and was excluded in the outdoor experiment. As shown in the plots the universal model is penalized for its one size fits all philosophy. The universal stress model provides the lowest accuracy of

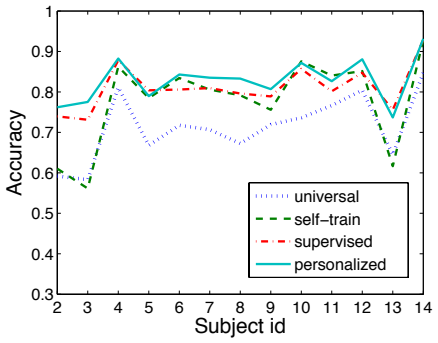

Figure 3: Accuracy of indoor scenario

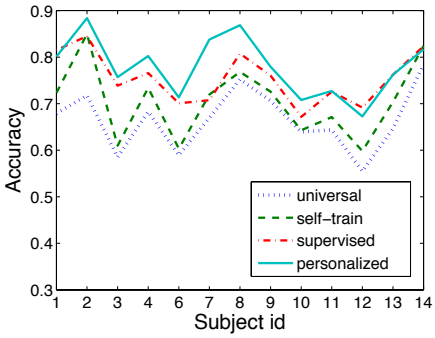

Figure 4: Accuracy of outdoor scenario

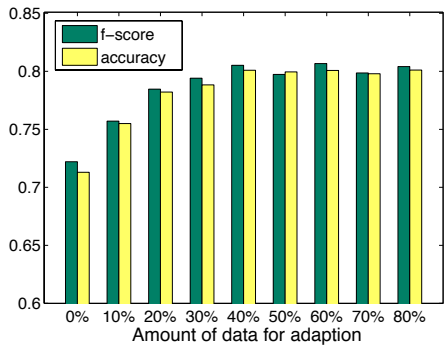

Figure 5: Supervised adaptation indoor

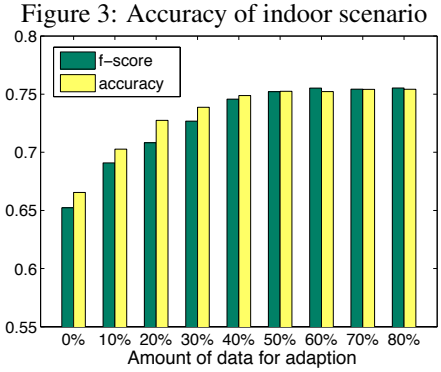

Figure 6: Supervised adaptation outdoor

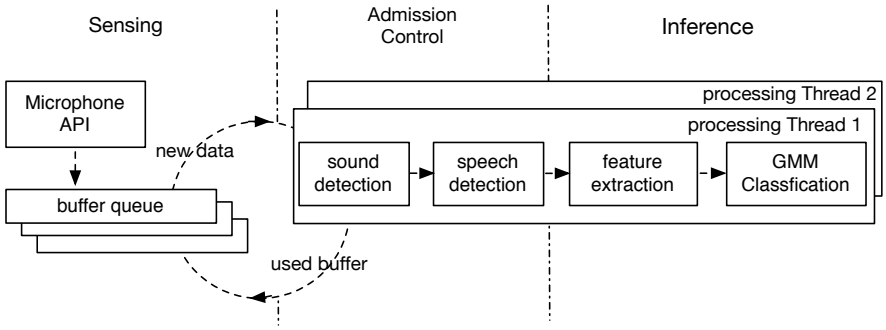

Figure 7: StressSense implementation component diagram.

71.3% for the indoor scenario and 66.6% for the outdoor scenario. For comparison [14] achieved 72.63% accuracy with only female speakers on data from controlled indoor environment using professional microphones. The personalized model provides the highest accuracy of 82.9% for the indoor scenario and 77.9% for the outdoor scenario.

The two adaptation schemes are intended to adapt a universal model to a new user when limited training data is available. None of the new user's data is part of the universal model's training data set. From Table 2 and Table 3, it is clear that each of the adaptation methods perform well above the universal stress model. Not surprisingly, the supervised adaptation scheme, which take advantages of user input, does a better job than the unsupervised self-train scheme. In both cases, the supervised adaptation scheme provides a 10% boost in accuracy, and is only about 2% lower than the personalized model. Therefore supervised adaptation is a practical alternative to a personalized scheme, when it is impractical to collect enough labeled training data for user-dependent models. We conduct another experiment to study the optimal amount of labeled data required for supervised adaptation. In this experiment, 20% of the data is held as test set, and we increase the adoption set from 10% to 80%. Figure 5 and Figure 6 show the performance of the supervised adapted model improves as more data is used for adaptation. For both indoor and outdoor scenario, the increase in stress classification performance levels off when approximately 40% of the data is used for adaptation. Generally speaking, using about 2 minutes of labeled data will increase the accuracy by about 8.3%, with minimal increase in accuracy there after.

The self-train scheme is a fully unsupervised. In comparison to the universal scheme, the self-train scheme provides an increase of performance of 6.5% and 4.0% for indoor and outdoor scenario, respectively. The self-train scheme is more effective in the indoor case because the indoor universal classifier works better than the outdoor one. Unlike the supervised adaptation scheme which receive new label information from the user, the self-train scheme relies on the class information encoded in the universal models. Therefore, it is sensitive to the performance of the original stress model that adaptation starts with.

As expected, stress detection in an uncontrolled outdoor environment is more challenging than the indoor environment. Even though human speech is the dominant sound in our data sets, the impact of different context is not negligible. In the following section, we investigate the effectiveness of cross scenario model adaptation, i.e., whether a stress model trained in a given acoustic environment can be adapted to perform well in another very different acoustic environment.

Cross Scenario Model adaptation

To test how well stress model trained in one scenario performs in a different scenario, we apply the indoor model to outdoor data and outdoor model to indoor data. The universal and personalized models are tested directly, while the adaptation models are adapted from the universal model of the original scenario using adaptation data from the new scenario. We use a 50% split between adaptation and test sets.

Table 5 shows the performance of stress classifier trained using indoor data is applied to unconstrained outdoor data. Compared to Table 2, both the universal and personalized classifiers perform poorly and are incapable of properly han-

	precision	recall	f-score	accuracy
universal	67.6%	28.2%	38.9%	57.8%
self-train	78.1%	26.4%	38.4%	60.1%
supervised	78.1%	70.1%	72.9%	74.8%
personal	65.3%	39.4%	47.7%	59.7%

Table 5: Indoor model tested outdoor

	precision	recall	f-score	accuracy
universal	64.8%	60.8%	62.2%	63.6%
self-train	66.5%	59.0%	61.7%	64.6%
supervised	74.1%	77.3%	75.5%	74.9%
personal	77.4%	77.7%	77.5%	77.4%

Table 6: Outdoor model tested indoor

Component	Avg. Runtime(sec)	
	Nexus S	Galaxy Nexus
admission control	0.005	0.003
feature extraction	1.95	1.20
classification	0.27	0.22
full pipeline	2.23	1.43

Figure 8: Runtime Benchmark

dling noisier outdoor data. Consequently, the self-train adaptation also provides limited performance gain but supervised adaptation is able to increases the accuracy by 17%. On the other hand, models trained on unconstrained outdoor data works better in the controlled indoor environment. Table 6 shows the performance of stress classification when outdoor stress classifiers are applied to indoor data. Due to the change of environment, the performance of outdoor classifier is lower when compared to the Table 3, where the native indoor classifiers are tested on indoor data. However, the performance drop is moderate compared to when the indoor model is applied to outdoor data. It is likely that classifiers trained on the real world data model speech under stress more precisely than classifiers trained in controlled environments and is more resilient to context changes. Supervised adaptation can further improve the accuracy by 10%. This result suggests that real world data is more important than data collected in a controlled environment and leads to more robust classifiers. When it is too difficult to collect a large real world dataset, a small amount of labeled data (e.g., 2 mins) from the new environment can make a big difference.

STRESSSENSE PROTOTYPE IMPLEMENTATION

Figure 7 shows our current proof-of-concept implementation of the StressSense classification pipeline. Our results show that it is feasible to implement a computationally demanding stress classification system on off-the-shelf smartphones. Insights and results from this initial implementation will serve as a basis for the further development and release of the StressSense App. The StressSense prototype is implemented on the Android 4.0 platform. The StressSense software comprises approximately 5,000 lines of code and is a mixture of C, C++ and Java. Most computationally demanding signal processing and feature (as listed in Table 1) extraction algorithms are written in C and C++ and interfaced with Java using JNI wrappers. The speaking rate is computed by cross-complied Enrate library[37]. Java is used to build an Android application which allows us to access the microphone data and construct a simple GUI to drive the app. Training is done offline (future work will consider online training) - the offline server side training code is implemented primarily in Matlab.

The StessSense software components and their interactions are shown in Figure 7. In current pipeline implementation, the processing comprises several computational stages with increasing computational cost. Each stage triggers the next more computationally demanding stage on an on-demand basis – presenting an efficient pipeline. Using the standard Android API, we collect 8 kHZ, 16-bit, mono audio samples from the microphone. The PCM formatted data is placed in a circular queue of buffers, with each buffer in the queue holding one processing window. Each processing window includes 40 non-overlapping frames. Each frame contains 256 16-bit audio samples. Once a buffer is full, it is provided to the sound detector component. If the sound detector detects the data as non-silence then data proceeds to voice detector to determine whether the incoming sound is human speech. We use sound and voice detection algorithm based on [31]. If human speech is present in the data then stress detection is applied. In this case, a stress/non-stress inference result is made each processing window (every 1.28 sec). The current implementation does not consider speaker segmentation due to the technical difficulty of recording from a second external microphone on current Android phone (we intend to study the use of multiple mics on phones as well as bluetooth headsets in future work).

StressSense is a CPU bound application due to the need to continuously run a classification pipeline. Note, however, that the full classification pipeline is only fully engaged when necessary – in the presence of human voice – else it runs at a low duty cycle. The current prototype is optimized for lower CPU usage at the cost of a larger memory footage. To best understand the cost of running StressSense on a phone we conducted a detailed component benchmark test, as shown in Table 8. We implement and benchmark StressSense on two Android phones: Samsung Nexus S and Samsung Galaxy Nexus. The Nexus S comes with a single core 1 GHz Cortex-A8 CPU while the newer Galaxy Nexus has a dual-core 1.2 GHz Cortex-A9 CPU. For a fair comparison, the runtime shown in Table 8 is measured with one single processing thread for both phones. When the whole pipeline is fully engaged, neither phone is able to process the data in real time with a single core. The Nexus S and Galaxy Nexus take 2.23s and 1.43s, respectively, to process a window (1.28s) of audio data. However, the Galaxy Nexus is able to achieve real-time processing when using two processing threads in parallel with two CPU cores. During full operation, the CPU usage is 93% - 97% and 46% - 55% for Nexus S (one CPU core) and Galaxy Nexus (two CPU cores), respectively. The computational power in newer generation phone is critical for real-time operation of StressSense. The memory usage is 7.8MB on Nexus S, whereas the memory usage on Galaxy Nexus is 15MB due to the memory used to process the extra processing thread. In terms of power consumption, the average current draw on a Galaxy Nexus is 53.64mA when the audio recorded is not voice (the pipeline stops at admission control stage in this case); when the audio recorded is human speech (full StressSense pipeline engages), the average current draw increases to 182.90mA. With the standard 1750 mAh battery, a Galaxy Nexus will last for 32.6 hours and 9.6 hours in above two cases respectively. Note, that the increased availability of multi-core phones opens the opportunity to implement more sophisticated pipelines in the future. Our implementation is one of the first that exploits dual-cores for continuous sensing; we show that this complex pipeline is

capable of running continuously in real-time without significant degradation of the phone's user experience. We believe future quad-core phones will lead to increased performance and make continuous sensing applications more commonplace in our daily lives.

CONCLUSION AND FUTURE WORK

We presented an adaptive method for detecting stress in diverse scenarios using the microphone on smartphones. Unlike physiological sensors, microphones do not require contact with the human body and are ubiquitous to all mobile phones. Using a non-iterative MAP adaptation scheme for Gaussian mixture models, we demonstrated that it is feasible to customize a universal stress model to different users and different scenarios using only few new data observations and at a low computational overhead. Our proof-of-concept software demonstrates that StressSense can run on off-the-shelf smart phone in real time. In our current work, we conducted an initial study of cognitive load related stress in job interviews and outdoor job execution tasks. For neutral voice, we used reading data. But in real world the stress and neutral scenarios are much more diverse. As part of future work – and building on our initial prototype implementation – we plan to design, deploy, and evaluate a StressSense Android App that harvests a diverse range of stress and neutral speech data from phone calls and conversations occurring in the wild from heterogeneous population of user. Each of the StressSense classification models evaluated in this paper present tradeoffs in terms of training and the burden on users.We plan to base our future StressSense application release on an adaptive pipeline that uses the self-train model for speaker adaption and the supervised adaption model for environment adaption.

ACKNOWLEDGMENTS

This work is supported in part by the SNSF SONVB Sinergia project, NSF IIS award #1202141, and Intel Science and Technology Center for Pervasive Computing. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of any funding body. A special thank goes to Andy Sarroff for his early contributions to this work.

REFERENCES

1. http://www.apa.org/pubs/info/reports/2007-stress.doc.
2. S. Cohen, R.C. Kessler, and L.U. Gordon. *Measuring stress: A guide for health and social scientists*. Oxford University Press, USA, 1997.
3. A. Perkins. Saving money by reducing stress. *Harvard Business Review*, 72(6):12, 1994.
4. Michael H. L. Hecker, Kenneth N. Stevens, Gottfried von Bismarck, and Carl E. Williams. Manifestations of Task-Induced Stress in the Acoustic Speech Signal. *The Journal of the Acoustical Society of America*, 44(4):993–1001, 1968.
5. K.R. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143–165, 1986.
6. K.R. Scherer. *Dynamics of stress: Physiological, psychological and social perspectives*, chapter 9: Voice, Stress, and Emotion, pages 157–179. Plenum Press, New York, 1986.
7. F.J. Tolkmitt and K.R. Scherer. Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12(3):302–313, 1986.
8. Allen D. Kanner, James C. Coyne, Catherine Schaefer, and Richard S. Lazarus. Comparison of two modes of stress measurement: Daily hassles and uplifts versus major life events. *Journal of Behavioral Medicine*, 4(1):1–39, 1981.
9. John H. L. Hansen. SUSAS. Linguistic Data Consortium, Philadelphia, 1999.
10. H. Teager. Some Observations on Oral Air Flow During Phonation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(5):599–601, 1980.
11. J.F. Kaiser. Some Useful Properties of Teager's Energy Operators. In *Proc. ICASSP '93*, volume 3, pages 149–152, 1993.
12. G. Zhou, J.H.L. Hansen, and J.F. Kaiser. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 9(3):201–216, 2001.
13. Raul Fernandez and Rosalind W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1–2):145–159, 2003.
14. SA Patil. Detection of speech under physical stress: Model development, sensor selection, and feature fusion. 2008.
15. Drew Fisher Keng-hao Chang and John Canny. Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. In *Proceedings of PhoneSense 2011*, 2011.
16. Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research. In *Proc. UbiComp '10*, pages 281–290, 2010.
17. K.R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger. Acoustic correlates of task load and stress. In *Proc. ICSLP2002*, pages 2017–2020, 2002.
18. S. Basu. A linked-hmm model for robust voicing and speech detection. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–816. IEEE, 2003.
19. D. Wyatt, T. Choudhury, and J. Bilmes. Conversation detection and speaker segmentation in privacy sensitive situated speech data. In *Proceedings of Interspeech*, pages 586–589. Citeseer, 2007.
20. T. Choudhury and S. Basu. Modeling conversational dynamics as a mixed memory markov process. In *Proc. of Intl. Conference on Neural Information and Processing Systems (NIPS)*. Citeseer, 2004.
21. J.P. Henry. Stress, neuroendocrine patterns, and emotional response. 1990.
22. K.R. Scherer. Voice, stress, and emotion. *Dynamics of stress: Physiological, psychological and social perspectives*, pages 157–179, 1986.
23. Robert Ruiz, Emmanuelle Absil, Bernard Harmegnies, Claude Legros, and Dolors Poch. Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions. *Speech Communication*, 20(1-2):111 – 129, 1996.
24. J.H.L. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication*, 20(1-2):151–173, 1996.
25. William A Jones Jr and Air Force Inst of Tech Wright-pattersonafb School of Engineer An Evaluation of Voice Stress Analysis Techniques in a Simulated AWACS Environment, 1990.
26. Lambert M Surhone, Miriam T Timpledon, and Susan F Marseken. Voice Stress Analysis, jul 2010.
27. N Morgan and E Fosler. Speech recognition using on-line estimation of speaking rate. *Conference on Speech*, 1997.
28. John Hansen and Sanjay Patil. Speech under stress: Analysis, modeling and recognition. In Christian Müller, editor, *Speaker Classification I*, volume 4343 of *Lecture Notes in Computer Science*, pages 108–137. Springer Berlin / Heidelberg, 2007.
29. John H L Hansen, Wooil Kim, Mandar Rahurkar, Evan Ruzanski, and James Meyerhoff. Robust Emotional Stressed Speech Detection Using Weighted Frequency Subbands. *EURASIP Journal on Advances in Signal Processing*, 2011:1–10, 2011.
30. J. Saunders. Real-time discrimination of broadcast speech/music. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 993–996. IEEE, 1996.
31. H. Lu, A. Bernheim Brush, B. Priyantha, A. Karlson, and J. Liu. Speakersense: Energy efficient unobtrusive speaker identification on mobile phones. *Pervasive 2011*, pages 188–205, 2011.
32. D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995.
33. D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.
34. Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1–3):19–41, 2000.
35. M.A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering*, pages 1437–1447, 2003.
36. R.E. Donovan. Trainable speech synthesis. *Univ. Eng. Dept*, page 164, 1996.
37. N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Fifth European Conference on Speech Communication and Technology*, 1997.