EmotionSense: A Mobile Phones based Adaptive Platform for Experimental Social Psychology Research

Kiran K. Rachuri
Computer Laboratory
University of Cambridge
kkr27@cam.ac.uk

Mirco Musolesi
School of Computer Science
University of St. Andrews
mirco@cs.st-andrews.ac.uk

Cecilia Mascolo
Computer Laboratory
University of Cambridge
cecilia.mascolo@cl.cam.ac.uk

Peter J. Rentfrow
Faculty of Politics, Psychology, Sociology
and International Studies
University of Cambridge
pjr39@cam.ac.uk

Chris Longworth
Department of Engineering
University of Cambridge
cl336@cam.ac.uk

Andrius Aucinas
Computer Laboratory
University of Cambridge
aa535@cam.ac.uk

ABSTRACT

Today's mobile phones represent a rich and powerful computing platform, given their sensing, processing and communication capabilities. Phones are also part of the everyday life of billions of people, and therefore represent an exceptionally suitable tool for conducting social and psychological experiments in an unobtrusive way.

In this paper we illustrate EmotionSense, a mobile sensing platform for social psychology studies based on mobile phones. Key characteristics include the ability of sensing individual emotions as well as activities, verbal and proximity interactions among members of social groups. Moreover, the system is programmable by means of a declarative language that can be used to express adaptive rules to improve power saving. We evaluate a system prototype on Nokia Symbian phones by means of several small-scale experiments aimed at testing performance in terms of accuracy and power consumption. Finally, we present the results of real deployment where we study participants emotions and interactions. We cross-validate our measurements with the results obtained through questionnaires filled by the users, and the results presented in social psychological studies using traditional methods. In particular, we show how speakers and participants' emotions can be automatically detected by means of classifiers running locally on off-the-shelf mobile phones, and how speaking and interactions can be correlated with activity and location measures.

ACM Classification Keywords

H.1.2 User/Machine Systems, J.4 Social and Behavioral Sciences, I.5 Pattern Recognition.

General Terms

Algorithms, Design, Experimentation.

UbiComp '10, Sep 26–29, 2010, Copenhagen, Denmark.
Copyright 2010 ACM 978-1-60558-843-8/10/09...$10.00.

Author Keywords

Emotion Recognition, Speaker Recognition, Social Psychology, Mobile Phones, Energy Efficiency.

INTRODUCTION

Mobile phones represent an ideal computing platform to monitor behavior and movement, since they are part of the everyday life of billions of people [1]. Recently, systems such as Cenceme [22] and Betelgeuse [16] have shown the potential of mobile phone sensing in providing information such as user movement and activity for recreational and healthcare applications. One possible use of these technologies is arguably the support to sociology experiments [20] which involve studying people's daily life and interactions. In the past, this analysis has been performed with the help of cameras (in home/working environments or in laboratories), by using voice recorders attached to people, and self reports using daily diaries or PDAs [6]. However, these techniques may lead to biased results since people are aware of being constantly monitored. Instead, mobile phones offer an *unobtrusive* means of obtaining information about the behavior of individuals and their interactions.

In this paper, we present EmotionSense, a framework for collecting data in human interaction studies based on mobile phones. EmotionSense gathers participants' *emotions* as well as proximity and patterns of conversation by processing the outputs from the sensors of off-the-shelf smartphones. This can be used to understand the correlation and the impact of interactions and activities on the emotions and behavior of individuals. In terms of system design, the key characteristics of this framework are *programmability* (social scientists can describe the sensing tasks using a declarative language), and *run-time adaptation* (social scientists can write rules to activate and deactivate sensors according to the user context). Although energy efficient sensing has previously been investigated in works such as [30], this is the first paper to propose a fully context-aware programmable mobile sensing system for social psychology research. Social scientists can modify the behavior of the system both in terms of sensing operations based on the analysis of the available information about the user, and its environment. For example, they can write a rule to activate a voice sensor only if people are close by to the user.

More specifically, the key contributions of this work can be summarized as follows:

- We design, implement, deploy, and evaluate a complete system for experimental sociology and psychology that is able to provide information about social group dynamics, especially with respect to the influence of activity, group interactions, and time of day on the emotions of the individuals, in an unobtrusive way.
- We present the design of two novel subsystems for emotion detection and speaker recognition built on a mobile phone platform. These subsystems are based on Gaussian Mixture methods [27] for the detection of emotions and speaker identities. EmotionSense automatically recognizes speakers and emotions by means of classifiers running locally on off-the-shelf mobile phones.
- We propose a programmable adaptive system with declarative rules. The rules are expressed using first order logic predicates and are interpreted by means of a logic engine. The rules can trigger new sensing actions (such as starting the sampling of a sensor) or modify existing ones (such as the sampling interval of a sensor).
- We present the results of a real deployment designed in collaboration with social psychologists. We found that the distribution of the emotions detected through EmotionSense generally reflected the self-reports by the participants. Our findings confirm in a quantitative way the corresponding social psychological studies using traditional methods. We detected correlation of user activity and emotions as well as emotions and location, and we were able to characterize participants talking in groups by their amount of speaking.

EmotionSense has the potential to change how social scientists study human behavior. The psychologists involved in our project appreciate the flexibility of EmotionSense, and above all, the possibility of exploiting an unobtrusive technology for their experiments, and the ability of cross-validating the information obtained by means of self-report-based methods in a rigorous and quantitative way. Finally, we observe that privacy is not a major concern for this system, since all users voluntarily agree to carry the devices for constant monitoring, like in other psychological studies. In any case, no voice recording is stored since the samples are discarded immediately after processing. Bluetooth identifiers are also not stored by the system.

SENSING HUMAN EMOTIONS

What are the typical emotions exhibited by people? How does the frequency of emotions vary? Can we measure emotions in a quantitative way? What is the correlation of emotion with location and activity? Or with interaction? How does speech patterns among group of users vary over time? Generations of social psychologists have tried to answer these questions using a variety of techniques and methodologies involving experiments on people.

The research methods used in the behavioral sciences make little use of technology. In addition to traditional self-reports, researchers may also rely on one-time behavioral observations of participants in laboratory settings. Such methods can be useful, but the fact that they are based on behavior in a lab raises concerns about their generalizability to non-lab contexts. Recently, researchers have begun to use new methods in an effort to examine behavior in everyday life. For example, daily diary methods [6] and experience sampling methods [11] ask participants to report the social events and psychological states they experienced either at the end of the day or periodically throughout the day. Another method has used devices that take audio recordings (or snapshots) of participants' daily lives every few minutes, which are later transcribed and coded by teams of researchers. These methods have advantages over the traditional survey methods, but they nevertheless suffer from issues associated with forgetting events that took place during the day, and carrying an additional obtrusive electronic device.

We argue that mobile sensing technology has the potential to bring a new perspective to the design of social psychology experiments, both in terms of accuracy of the results of the study and from a practical point of view. Mobile phones are already part of the daily life of people, so their presence is likely to be "forgotten" by users, leading to accurate observation of spontaneous behavior. The overarching goal of EmotionSense is to exploit mobile sensing technology to study human social behavior.

The research challenges related to design of the EmotionSense system are three-fold. First, efficient inference algorithms needs to be exploited to extract high-level information from the available raw data of not always accurate sensors embedded in mobile phones. Second, an efficient system for this class of resource-constrained devices (especially in terms of power consumption) needs to be devised. Third, the system should be easily programmable and customizable for different types of experiments with changing requirements. Our goal is to use off-the-shelf devices so that inexpensive large-scale deployments can be possible. In the next section, we present a high level description of EmotionSense and then discuss in detail its key components, in particular those for speaker recognition, emotion detection, and rule-based dynamic adaptation.

SYSTEM OVERVIEW

In this section, we discuss the overall architecture of EmotionSense presenting the key design choices and features of each component of the system.

EmotionSense at a Glance

The EmotionSense system consists of several sensor monitors, a programmable adaptive framework based on a logic inference engine [25], and two declarative databases (*Knowledge Base* and *Action Base*). Each monitor is a thread that logs events to the *Knowledge Base*, a repository of all the information extracted from the on-board sensors of the phones. The system is based on a declarative specification (using first-order logic predicates) of:

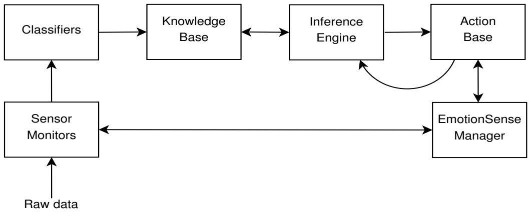

Figure 1. Information flow in EmotionSense.

- *facts*, i.e., the knowledge extracted by the sensors about user behavior (such as his/her emotions) and his/her environment (such as the identity of the people involved in a conversation with him/her).
- *actions*, i.e., the set of sensing activities that the sensors have to perform with different duty cycles, such as recording voices (if any) for 10 seconds each minute or extracting the current activity every 2 minutes.

By means of the inference engine and a user-defined set of rules (a default set is provided), the sensing actions are periodically generated. The actions that have to be executed by the system are stored in the *Action Base*. The Action Base is periodically accessed by the *EmotionSense Manager* that invokes the corresponding monitors according to the actions scheduled in the Action Base. Users can define sensing tasks and rules that are interpreted by the inference engine in order to adapt dynamically the sensing actions performed by the system. One example of such a rule is to start the GPS receiver only if the user is currently moving. The flow of information in EmotionSense is illustrated in Figure 1.

EmotionSense Manager

The *EmotionSense Manager* starts all the sensor monitors, the inference engine, and instantiates the *Knowledge Base*. Each monitor is a thread which collects data from the sensor with a given duty cycle. The EmotionSense manager periodically invokes the inference engine to process the latest facts from the *Knowledge Base* and generates actions that are stored in the *Action Base*. The manager is then responsible for scheduling all the sensing actions. The sensing actions are scheduled by updating the state and parameters of each monitor according to the actions generated by the inference engine. Example output actions are setting *sampling interval* of GPS sensor, accelerometer sensor, and so on.

Speaker and Emotion Recognition Component

This monitor is responsible for speaker and emotion recognition. It records audio samples with a variable sampling interval. Each sample is processed to extract speaker and emotion information by comparing it against a set of preloaded emotion and speaker-dependent models, collected offline during the setup phase of the system. The EmotionSense manager updates this monitor with the latest co-location data (obtained via Bluetooth). Thus, only the models associated with co-located users need to be examined, improving the efficiency and accuracy of the speaker recognition component. The speaker recognition component also includes a silence model. When no speech is detected, the computationally intensive emotion classification algorithm is not executed. Details of these subsystems are presented in the next section.

Sensor Monitors

The system is based on several sensor *monitors*. The *Accelerometer Monitor* infers the current activity by evaluating the mean and the average of the amplitudes of the accelerometer signal. The movement is extracted by means of a discriminant function classifier [5]. The classifier is able to discriminate between two types of activities: movement and non-movement. The detection of complex actions was not one of the main goals of the design of this prototype; in fact, many classifiers can be plugged in such as those presented in [22]. The *Bluetooth Monitor* is responsible for detecting other Bluetooth devices that are in proximity, and we used the *lightblue* [18] module for Python For Symbian S60 (PyS60) to get this information. When the system is set up, the Bluetooth identifier of the phone is associated with each user. Finally, the *Location Monitor* is responsible for tracking the location of the user by analyzing the output of the GPS receiver. We used the *positioning* module of PyS60 to get location information. The monitor tries to extract valuable information even if the GPS data are not complete. An example of an incomplete position is one that contains data about the satellites used to obtain a GPS fix but no latitude or longitude data. This can at the least be used to infer whether the user is indoors/outdoors.

Action and Knowledge Base

Another key feature of the system is the *Knowledge Base*, which stores the current facts that are inferred from the raw data generated by the various sensors. All the monitors log facts to the Knowledge Base, which are in turn used by the inference engine to generate actions. The *Knowledge Base* loads in memory only a snapshot of the facts (i.e., not all facts that are generated so far but only the unprocessed facts) to reduce application footprint. The older facts are logged to a file. The format of facts is as follows:

```
fact(<fact_name>, <value>)
```

The corresponding timestamps of these facts are also stored. Actions are also treated as facts, but with an extra identifier which is of the form:

```
fact(‘action’, <action_name>, <value>)
```

Some examples are:

```
fact(Activity, 1)
fact(‘action’, ‘ActivitySamplingInterval’, 10)
```

The former indicates that the user is currently moving, and the latter means that the sampling interval of accelerometer should be set to 10 seconds.

Inference Engine

The adaptation framework is based on a set of *adaptation rules* that allow for changing the behavior of the system at run-time by monitoring the current activity, co-location with other people, and location of the person carrying the mobile phone. The adaptation rules are used to modify the sampling behavior of the system according to the observed status of the user, (e.g., if a person is moving or not) and his/her surroundings (e.g., if there are other people around, if they are currently talking, and so on). The adaptation framework helps in saving energy by reducing the amount of data sampling and information processing without compromising considerably on the accuracy of the inference. We will present more details about the inference engine in the implementation section.

IMPLEMENTATION

In this section we provide more details about the implementation of the fundamental components of our system, describing the key design choices and original solutions. We implemented the EmotionSense system on a Nokia 6210 Navigator phone using PyS60 for most of the components. The speaker recognition component is implemented in C++ since it is based on tools of the Hidden markov model ToolKit (HTK) suite for speech processing originally written in that language [13].

Speaker Recognition Subsystem

The speaker recognition subsystem is based on a Gaussian Mixture Model classifier [5, 26], which is implemented using HTK [13][1]. HTK is a portable toolkit for building and manipulating Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) and provides sophisticated facilities for speech analysis, model training, testing and results analysis. At present HTK is available for Windows and Linux systems only. It was therefore necessary to adapt the main components of the toolkit to work on the Nokia Symbian S60 platform.

The speaker recognition process is performed as follows:

- Speech data are collected from all users enrolled in the current experimental study. The data are then parameterized using a frame rate of 10ms and a window size of 30ms and a vector of 32 Perceptual Linear Predictive (PLP) coefficients [12] (16 static and 16 delta) are extracted from each frame.
- A 128-component universal background GMM (representative of all speakers) is then trained using all available enrollment speech to optimize a maximum likelihood criterion. This training procedure is currently executed offline. However, the training procedure could also be executed at run-time by sending samples to the back-end servers by means of a WiFi or 3G connection.
- Next, a set of user-dependent GMMs are obtained by performing Maximum A Posteriori (MAP) adaptation of the background model using the enrollment data associated with each user. The adaptation constant used for the MAP process was set to 15.
- Finally, at run-time the likelihood of each audio sequence is calculated given each user model. Each sequence is then associated with the model that assigns it the highest likelihood. This is the Bayes decision rule [5] in the case that the prior probability associated with each user is equal.

In order to improve accuracy and efficiency of the system, two key mechanisms were implemented:

- *Silence detection.* Successfully detecting silence can improve the efficiency of the system by eliminating the need to compare each sequence with each user-dependent model. Silence detection was implemented by training an additional GMM using silence audio data recorded under similar background conditions to the enrollment data. Each audio sequence is initially classified as either silence/non-silence by comparing the likelihood of the sequence given the silence and the background GMMs. The silence detector can also be used to infer information about user environment, sleep patterns, and so on.
- *Comparisons driven by co-location information.* To reduce the total number of comparisons required, the speaker recognition component compares a recorded audio sequence only with the models associated with co-located users. Co-location is derived from the facts extracted by the Bluetooth monitor. This avoids unnecessary comparisons against models of people who are not in proximity of the user, both considerably speeding up the detection process and potentially avoiding misclassifying the sequence as belonging to users that are not present.

Emotion Recognition Subsystem

The emotion recognition subsystem is also based on a GMM classifier. The classifier was trained using emotional speech taken from the Emotional Prosody Speech and Transcripts library [17], the standard benchmark library in emotion and speech processing research. This corpus contains recordings of professional actors reading a series of semantically neutral utterances (dates and numbers) spanning fourteen distinct emotional categories. The selection is based on Banse and Scherer's study [3] of vocal emotional expression. Actor participants were provided with descriptions of each emotional context, including situational examples adapted from those used in the original study. Flashcards were used to display series of four-syllable dates and numbers to be uttered in the appropriate emotional category.

The emotion recognition process is performed as follows:

- A 128-component background GMM representative of all emotional speech is initially trained using all the emotion data.
- MAP adaptation of the background model is performed offline using emotion specific utterances from the emotion database to obtain a set of emotion-dependent models. These models are then loaded onto the phones.
- At run-time, the component periodically calculates the likelihood of the recorded audio sequence given each emotion-dependent model and assigns the sequence to the emotion characterized by the highest likelihood.

We initially tested a total of 14 "narrow" emotions based on the classes defined in the emotion library. These were then clustered into 5 standard broader emotion groups generally used by social psychologists [10]. It is difficult to distinguish with high accuracy between utterances related to emotions in the same class given their similarity. In any case, we also

[1] Alternative SVM-based schemes, including the popular GMM-supervector [7] and MLLR [28] kernel classifiers, were not considered as they are generally suitable for binary classification tasks.

Table 1. Emotion clustering

Broad emotion	Narrow emotions
Happy	Elation, Interest, Happy
Sad	Sadness
Fear	Panic
Anger	Disgust, Dominant, Hot anger
Neutral	Neutral normal, Neutral conversation, Neutral distant, Neutral tete, Boredom, Passive

```
set_location_sampling_interval
  foreach
    facts.fact($factName, $value)
    check $factName == ’Activity’
    facts.fact($actionName, $currentInterval)
    check $actionName == ’LocationInterval’
    $interval = update($value, $currentInterval)
  assert
     facts.fact(’action’, ’LocationInterval’, $interval)
```

Figure 2. An example rule to set sampling rate of GPS sensor.

note that it is also hard for a person involved in an experiment to distinguish exactly among the emotions belonging to the same class in a questionnaire and for this reason broad classes are commonly used. The details of each grouping is given in Table 1.

Adaptation Framework

The adaptation framework is based on Pyke [25], a knowledge-based inference engine written in Python. It takes a set of facts as inputs and derives additional facts through forward chaining rules. It can also be used to prove goals using backward chaining rules. However, these are not necessary in our system and were removed when we adapted Pyke to the Nokia Symbian 60 platform in order to reduce the memory footprint. We have defined adaptation rules which drive the behavior of the entire EmotionSense system. Each of these rules sets the sampling interval of a sensor based on the data extracted from it. *EmotionSense Manager* instantiates the Pyke inference engine, and periodically invokes it to process facts and generate actions, and it in turn updates the tasks for the sensor monitors.

An example of a rule used in the EmotionSense system is given in Figure 2. The rule updates the value of location sampling interval based on the data from the accelerometer sensor. It gets the fact `Activity` and the current location sampling interval `LocationInterval` from *Knowledge Base*, and then updates it based on a function (`update()`). The idea is to provide a simple interface to add rules in order to change the behavior of the system. In the EmotionSense system, the function `update()` is based on a back-off mechanism as shown in Figure 3. If the user is moving then the sampling interval is set to a minimum value otherwise it is increased by doubling each time until it reaches a maximum value. The sampling interval stays at this maximum as long as user is idle, but, as soon as movement is detected, it is set to a minimum value. In addition to the GPS sensor, we have similar rules for microphone sensor, Bluetooth sensor, and accelerometer sensor. This way, users can write very simple functions to adapt the system to external changes. The parameters `MIN_INTERVAL` and `MAX_INTERVAL` play a crucial role in the back-off function. In order to find the optimum values of these parameters, we conducted several benchmark tests for each of the sensors. We will present the results of some of these tests as methodological example in the next section.

```
def update(value, currentInterval):
  if value == 1:
    samplingInterval = MIN_INTERVAL
  elif value == 0:
    samplingInterval = min(2*currentInterval, MAX_INTERVAL)
  return samplingInterval
```

Figure 3. A back-off function for updating sampling interval.

EVALUATION

We first present an evaluation of EmotionSense by means of several micro-benchmark tests to study the system performance and to tune the parameters of the adaptation mechanisms. In particular, we discuss the choice of the optimal sample length value for speaker and emotion recognition and a generic methodology for the selection of the optimal values for the parameters of the rules. These initial experiments involved 12 users. We then describe the results of a larger scale deployment involving 18 users for 10 days to evaluate the prototype in a realistic setting and demonstrate its usefulness for social science.

Performance Benchmarks

We ran a series of micro-benchmarks to test the performance of the different components and mechanisms of our system. The data used for benchmarking the adaptation rules were collected from 12 users over 24 hours. Each user carried a Nokia 6210 mobile phone, which continuously monitored and recorded the outputs of the accelerometer, the microphone, and the Bluetooth sensors. We then used these data as a trace to benchmark the different components and tune the various parameters of our system. We used a different data set for benchmarking the speaker and emotion recognition subsystems as discussed later in this section. We also explored the trade-offs in performing local computation on the phones and remote computation on a back-end server.

Speaker Recognition

In this subsection, we present the results of speaker recognition subsystem benchmarks. Voice samples from 10 users were used for this test using approximately 10 minutes of data for training the speaker-dependent models. A separate, held-out dataset was used to test the accuracy of the speaker recognition component. We varied the sample length from 1 to 15 seconds and each sample was classified against 14 possible models. We used 15 samples per user per sample length, resulting in a total of 150 test samples per sample length. Figure 4 shows the speaker recognition accuracy with respect to the sample length. As the sample length was increased the accuracy improved, converging at around 90% for sample lengths greater than 4. From Figure 5, it can be seen that this corresponds to a latency of 55 seconds in the case of local computation on the phone.

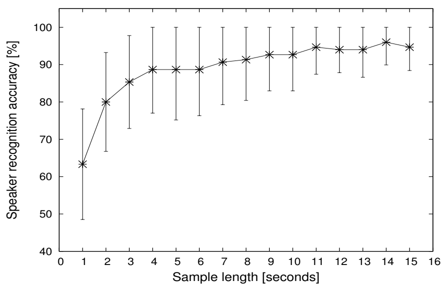

Figure 4. Speaker recognition accuracy vs audio sample length.

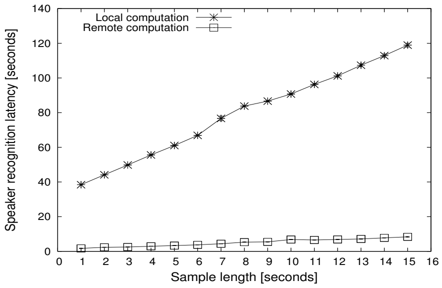

Figure 5. Speaker recognition latency vs audio sample length.

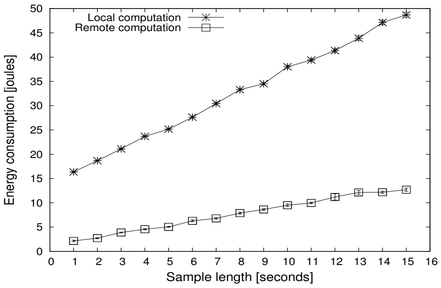

Figure 6. Speaker recognition energy consumption vs audio sample length.

We then compare the efficiency of classifying each audio sample either locally on the phone or remotely on a powerful server reached via the 3G network. We can observe from Figures 5 and 6 that remote computation is more efficient in terms of both latency and energy consumption. The use of the 3G network is acceptable and telephone contracts can be purchased easily for experiments, although sometimes costs might be an issue, especially if the scale of the planned experiment is very large. However, by avoiding to use 3G data transmission, the software can be distributed to participants who want to use their own SIM cards (with their own phone number and contacts) on the experiment phone, without impacting on their bills (with transmission of our potentially large quantity of data): this can be very important in the context of the experiment. Since no personal data (such as voice recordings, which are very sensitive) are sent to a back-end server, privacy issues related to the transmission to a remote server can be avoided by adopting the local computation approach. In the case of local computation, the process of comparing n voice samples with all the pre-loaded models is performed on a Nokia 6210 mobile phone which is equipped with ARM 11 369MHz processor. Instead, with respect to the remote computation case, an audio sample to be classified is sent over the 3G network using the *HTTP Connection* module of PyS60 to a powerful back-end server (Intel Xeon Octa-core E5506 2.13GHz processor, and 12 GB RAM). An audio sample of 5 seconds length has a size of about 78KB. The energy consumption shown in the results is end-to-end consumption including all computation, and radio transmission costs. We measured the energy consumption using the Nokia Energy Profiler.

We also conducted a test to evaluate the effect of noise on speaker recognition accuracy. We used Audacity [2], an open source cross-platform sound editor, to inject noise into voice samples. Audacity provides an easy way to add a particular type of noise into voice samples. We injected Brownian noise into all the test samples for their entire length with amplitudes ranging from 0 to 0.1, in increments of 0.02. Figure 7 shows the effect of Brownian noise on speaker recognition accuracy. As expected, the accuracy decreases as the amplitude of noise increases.

Emotion Recognition

In order to benchmark the emotion recognition subsystem, we used both test and training data from the Emotional Prosody Speech and Transcripts library [17]. The advantage of using this library is that it is difficult for non professionals to deliver emotional utterances. An alternative is to use "natural" speech recordings (i.e., taken from everyday life situations without acting). However, it is difficult to determine appropriate reference labels, required to evaluate performance on this speech, since many natural utterances are emotionally ambiguous. The use of a pre-existing library also allowed us to avoid explicitly annotating collected data with emotional labels. We used a total of 14 narrow emotions, which were then grouped into 5 broad emotion categories. For each narrow emotion, we used a total of 25 test samples per narrow emotion per sample length, resulting in a total of 350 test samples per sample length.

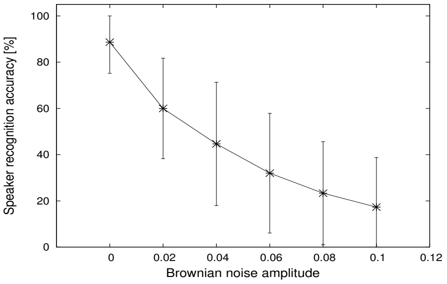

Figure 7. Effect of Brownian noise on speaker recognition accuracy for a sample length of 5 seconds.

Figures 8, 9, and 10 show the emotion recognition accuracy, latency, and energy consumption with respect to the sample length, respectively. As the sample length increases the accuracy improves, converging to about 71% for the broad emotions for sample lengths greater than 5 seconds. Based on the speaker and emotion recognition accuracy results (Figures 4 and 8), we used a sample length of 5 seconds in the EmotionSense system which is the point where the convergence becomes evident. The confusion matrix for broad emotions for a sample length of 5 seconds is shown in Table 2. Among non-neutral emotions, anger has the highest accuracy out of all. This is confirmed in [4], where the authors show that intense emotions (like anger) are easier to detect than emotional valence. They also mention that the emotions that are similar in intensity, like anger and fear (panic), are hard to distinguish: the same can be observed in our confusion matrix.

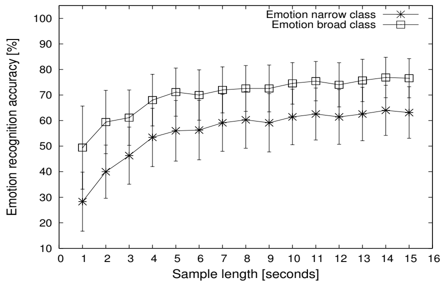

Figure 8. Emotion recognition accuracy vs audio sample length.

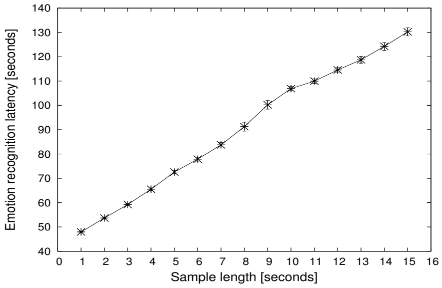

Figure 9. Emotion recognition latency vs audio sample length.

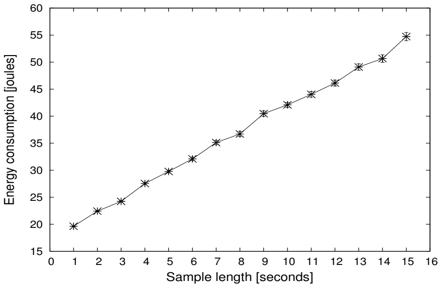

Figure 10. Emotion recognition energy consumption vs audio sample length.

We also note that distinguishing between some narrow emotions in a group is difficult given the similarity of the utterances corresponding to them: for a sample length of 5 seconds, the narrow emotion "happy" matches "interest" with a probability of 0.28. Grouping these increases the accuracy of classification which can be observed from Figure 8. The usage of a limited number of broader emotional classes is also advocated by social psychologists [10]: in general, classifying own emotions using narrow categories when filling self-report questionnaires is also difficult.

Dataset for Sensor Benchmarks

Trace files with ground-truth information for all sensors were generated based on the data collected from 12 users for 24 hours. In order to extract the microphone sensor trace, audio samples of 5 seconds length were recorded continuously with a sleep period of 1 second between consecutive recordings. Co-location data for the Bluetooth sensor trace is queried continuously with a sleep duration of 3 seconds between successive queries. The accelerometer sensor is sampled continuously for movement information with a gap of 1 second. The data from these sensors were processed and trace files with event information were generated. In the traces, the events generated from the data of each sensor can be of two types, viz., "unmissable" and "missable" events. An unmissable event is an event of interest observed in the environment that should not be missed by the sensor. A missable event indicates that no interesting external phenomenon has happened and the corresponding sensor can sleep during this time. Finally, the accuracy of a rule is measured in terms of the percentage of events missed by the corresponding sensor. An event is said to be missed when there is an unmissable event recorded in the trace file while the corresponding sensor monitor is sleeping. Also, the percentage of missed events is a relative value with respect to the percentage of missed events for the lowest possible sampling rate; for this reason, the results of all the plots for accuracy start with zero percentage of missed events.

Given the space constraints we present, as an example, the tuning of the microphone sensor: in this case, an unmissable event corresponds to some audible voice data being heard in the environment, and a missable event corresponds to silence. These events are generated by comparing a recorded audio sample with two pre-loaded models, the background GMM (same as that used for speaker recognition benchmarks) and silence model. The trace file is a list of events (tagged as missable and unmissable) with timestamps. The main goal of benchmarks is to find optimal values for the `MIN_INTERVAL` and `MAX_INTERVAL` parameters discussed in the previous section.

Table 2. Confusion matrix for broad emotions.

Emotion [%]	Happy	Sad	Fear	Anger	Neutral
Happy	58.67	4	0	8	29.33
Sad	4	60	0	8	28
Fear	8	4	60	8	20
Anger	6.66	2.66	9.34	64	17.33
Neutral	6	5.33	0	4	84.66

Figures 11 and 12 show the effect of increasing the value of `MIN_INTERVAL` on percentage of missed events and energy consumption for the microphone sensor rule. Figures 13 and 14 show the effect of increasing the value of `MAX_INTERVAL` on percentage of missed events and energy consumption for the microphone sensor rule. We can observe that all these plots exhibit asymptotic behavior. Based on these plots and energy saving as the main motivation, we set `MIN_INTERVAL` to 45 seconds and `MAX_INTERVAL` to 100 seconds. We performed similar experiments for the other sensors. As a result, for the Bluetooth sensor, we set `MIN_INTERVAL` to 30 seconds and `MAX_INTERVAL` to 100 seconds; for the accelerometer, we set `MIN_INTERVAL` to 10 seconds and `MAX_INTERVAL` to 30 seconds; for the GPS sensor, we set `MIN_INTERVAL` to 180 seconds and `MAX_INTERVAL` to 800 seconds.

Social Psychology Experiment

After evaluating the accuracy of the system by means of the micro-benchmark tests, we conducted a social psychology experiment to evaluate the usefulness of the EmotionSense system for social scientists. The data extracted by means of the EmotionSense system running on the mobile phones were compared to information provided by participants by means of traditional questionnaires.

Overview of the Experiment

The experiment was conducted for a duration of 10 days involving 18 users. Users were members of the local Department of Computer Science. Since the system did not require any user-phone interaction, the fact that the partici-

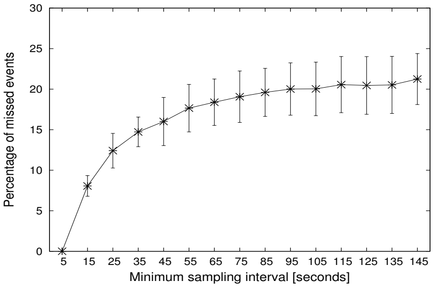

Figure 11. Percentage of missed events vs minimum sampling interval for microphone rule.

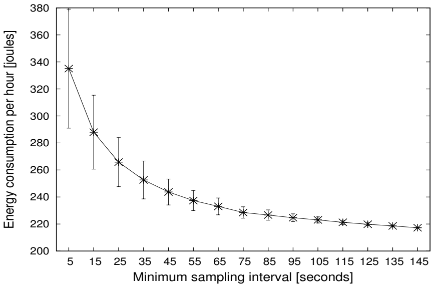

Figure 12. Energy consumption vs minimum sampling interval for microphone rule.

pants were technology-savvy is not a determinant factor for the outcomes of the experiment. Each user carried a Nokia 6210 mobile phone for the total duration of the experiment. Users filled in a daily dairy questionnaire for each day of the experiment which was designed by a social psychologist of our team following the methodology described in [6]. We divided a day into 30-minute slots, and asked the users to fill a questionnaire about the activity/event they were involved in at a particular time of day. We also asked them if the event happened indoors or outdoors, their location, if there were other people present (in particular, participants involved in our study). Furthermore, we asked them to specify their mood at that time.

Results and Discussion

We analyzed the distribution of emotions detected, and also the effect of time of day, activity, and co-location, on the distribution of emotions. Figure 15 shows the distribution of "broad" emotions detected by the system during the experiment. We can infer that people tend to exhibit neutral emotions far more than other emotions. Fear is the least shown emotion of all. This is very much in accordance with existing social psychological studies [9], where it is shown that most of social activity is effectively neutral, with the exception of rare arguments. Figure 16 shows the distribution of emotions from the questionnaires filled by users, which shows a distribution of emotions similar to that extracted by means of our system, except for the emotion "happy". Based on our discussions with the participants, we found that the users have indicated the "happy" emotion to represent their mental state but this does not necessarily mean that they were expressing happiness in a verbal way.

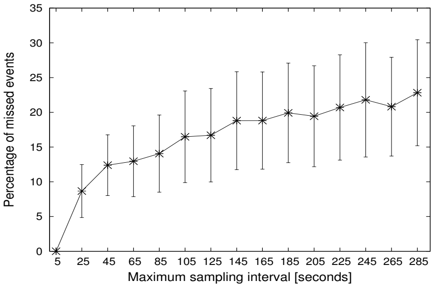

Figure 13. Percentage of missed events vs maximum sampling interval for microphone rule.

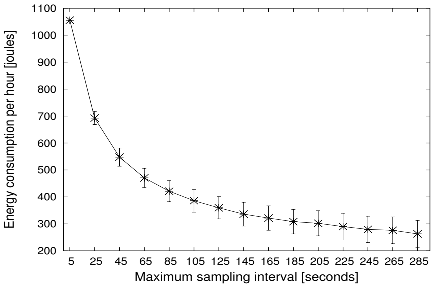

Figure 14. Energy consumption vs maximum sampling interval for microphone rule.

Figure 17 shows the distribution of emotions with respect to the time of day. We can infer that users tend to exhibit non-neutral emotions more frequently during evenings than mornings. This is particularly true with respect to most of the users of this experiment who were more relaxed in the evenings than mornings. Instead of studying a global percentage of each of the emotions with respect to activity, we plotted the distribution of relative percentage of broad emotions when users are stationary and mobile. From Figure 18, we can observe that these distributions are very close, and the relative ordering is the same in both the cases. Figure 19 shows the effect of the number of co-located participants on the emotions of users. We can observe that the total number of emotions detected in smaller groups is higher than that in larger ones. However, this can also be due to the fact that our users spent more time in smaller groups than larger. In order to fathom this phenomenon better, we compared the distributions of relative percentage of emotions detected, which is shown in Figure 20. We can infer that the number of co-located participants has little effect on emotions like neutral and happy. Furthermore, we can also observe that people tend to exhibit sad and anger emotions lesser in larger groups than smaller groups. These results which we obtained are in-line with that of results found in social psychology studies [21]. We were able to associate predominant non-neutral emotion to location categories: we found that the most common emotion in residential areas was "happy" (45%), whereas in the workplaces and city center "sad" was the mostly detected (54% and 49%, respectively). These results show the potential of a deeper analysis of correlation between emotions and location, which we plan to investigate further from the socio-psychological perspective with

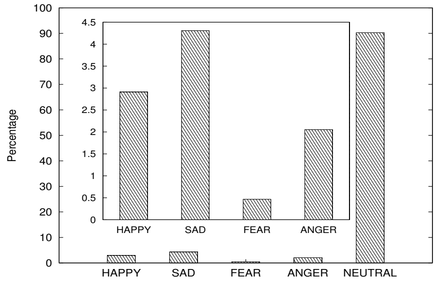

Figure 15. Distribution of broad emotions detected.

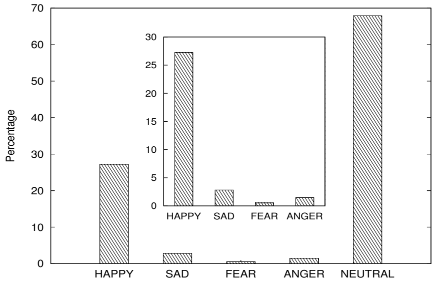

Figure 16. Distribution of broad emotions from daily dairies.

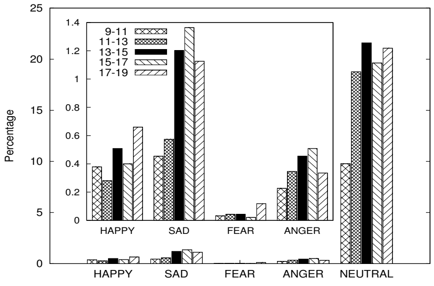

Figure 17. Distribution of broad emotions detected with respect to time of day.

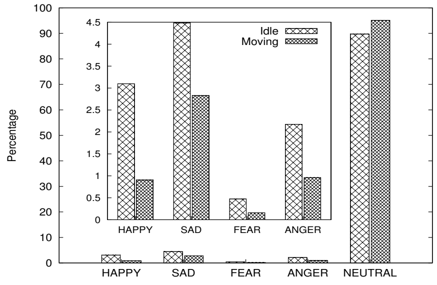

Figure 18. Distribution of broad emotions detected within a given physical state (idle/moving).

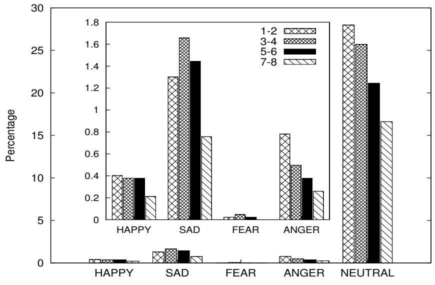

Figure 19. Distribution of broad emotions detected with respect to number of co-located participants.

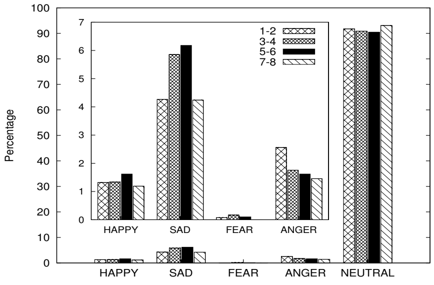

Figure 20. Distribution of broad emotions detected within a given number of co-located participants.

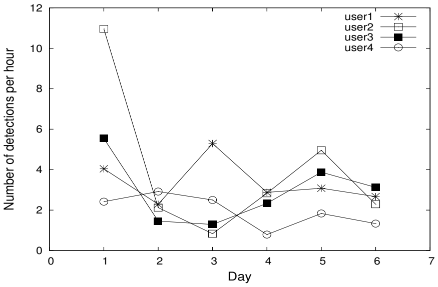

Figure 21. Variation of speech patterns over time.

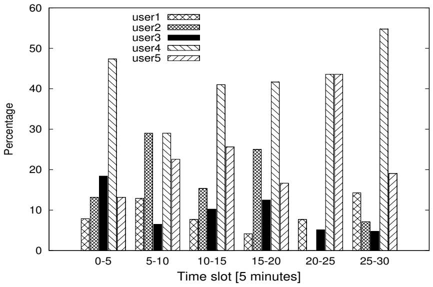

Figure 22. Speech time as percentage of the total speech time for all users per time slot during the meeting experiment.

more focused studies. We note that these findings might be specific to the cultural context of this study. EmotionSense can also be used to analyze the speech patterns of users with respect to time. Figure 21 shows the speech patterns of users over time, which reveal a considerable amount of consistency in most participants' verbal behavior across a 6-day period (except for the first day of "user2" that looks unique for him/her).

We also analyzed the data collected from a meeting which was held during the experiment when 11 of the participants sat together and talked for 30 minutes. We identified *conversation leaders* in each time slot of length 5 minutes. The analysis is shown in Figure 22. We considered only the top five most active speakers for this plot. We used an audio sample length of 3 seconds for this analysis as the difference between the speaker recognition accuracies for sample lengths of 3 and 5 is only 3% (Figure 4). We can observe that "user4" is the leader in almost all the slots, except in 2nd and 5th slots, where he/she was challenged by "user2" and "user5", respectively.

RELATED WORK

In the past years, we have witnessed an increasing interest in the use of ubiquitous technologies for measuring and monitoring user behavior [24]. Experience sampling [11, 23] is also used to evaluate human-computer interaction especially for mobile systems since the use of the devices is not restricted to indoor environments. The most interesting example of such systems is MyExperience [11], a system for feedback collection triggered periodically, partly based on the state of the on-board sensors. MyExperience does not include components for speaker, emotion recognition and energy saving mechanisms. In [23], the authors examined the use of a mobile phone based experience sampling application for cognitive behavioral therapy. The application collects the data about users emotions and their scales, however, this data has to be manually entered into the system. Recently, EEMSS, an energy efficient system for user activity recognition, has been presented in [30]. With respect to this work, EmotionSense allows for programmability and provides additional capabilities in terms of co-location detection and emotion and speaker recognition. The authors of [14] model dominance in small group meetings from audio and visual cues, however, they do not model emotions. Recent systems for quantitatively measuring aspects of human behavior using purpose-built devices include the Sociometer [8] and the Mobile Sensing Platform [31]. EmotionSense instead targets off-the-shelf devices that are already part of the everyday life of billions of individuals.

With respect to voice processing technologies, a survey of GMM-based speaker recognition technology can be found in [26]. SoundSense [19] is a system for recognizing sound types (music and voice) and situations based on mobile phones; similar algorithms that are complementary to ours may be added to EmotionSense to provide information about the social situations to psychologists. The emotion recognition system that we implemented is close to that devised by the Brno University of Technology team for the Interspeech 2009 Emotion challenge described in [15]; however, this system was not based on mobile phones. In addition to voice based emotion recognition systems, there are systems built using wearable emotion detectors [29].

CONCLUSIONS

In this paper, we have presented EmotionSense, a novel system for social psychology study of user emotion based on mobile phones. We have presented the design of novel components for emotion and speaker recognition based on Gaussian Mixture Models. We have discussed the results of the evaluation of the system by means of a series of benchmarks and a large-scale experiment that involved 18 participants. We have also shown how the information collected by EmotionSense can be used by social scientists in order to understand the patterns of interaction and the correlation of emotions with places, groups, and activity.

We plan to improve the emotion classifiers by optimizing the size of model and the PLP front-ends in order to obtain an optimal one for emotion recognition [27] and by connecting external sensors such Galvanic Skin Response device. We also plan to improve the noise robustness of the system by considering more realistic noise models. Finally, our long-term goal is to be able to provide real-time feedback and psychological help to users/patients in an interactive way, also by monitoring therapies day by day and modifying them if necessary.

ACKNOWLEDGMENTS

The authors would like to thank the participants of the social psychology experiment. This work was supported through Gates Cambridge Trust, and EPSRC grants EP/C544773, EP/F033176, and EP/D077273.

REFERENCES

1. T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich. Mobiscopes for Human Spaces. *IEEE Pervasive Computing*, 6:20–29, 2007.
2. Audacity. http://audacity.sourceforge.net/.
3. R. Banse and K. R. Scherer. Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.
4. V. Bezooijen, Otto, and Heenan. Recognition of Vocal Expressions of Emotion: A Three-Nation Study to Identify Universal Characteristics. *Journal of Cross-Cultural Psychology*, 14:387–406, 1983.
5. C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.
6. N. Bolger, A. Davis, and E. Rafaeli. Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology*, 54(1):579–616, 2003.
7. W. Campbell, D. Sturim, and D. Reynolds. Support Vector Machines using GMM-Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13:308–311, 2006.
8. T. Choudhury and A. Pentland. Sensing and Modeling Human Networks using the Sociometer. In *Proceedings of ISWC '03*, pages 216–222, 2003.
9. L. A. Clark and D. Watson. Mood and the Mundane: Relations Between Daily Life Events and Self-Reported Mood. *Journal of Personality and Social Psychology*, 54.2:296–308, 1988.
10. L. Feldman Barrett and J. Russell. Independence and Bipolarity in the Structure of Current Affect. *Journal of Personality and Social Psychology*, 74:967–984, 1998.
11. J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay. MyExperience: A System for In situ Tracing and Capturing of User Feedback on Mobile Phones. In *Proceedings of MobiSys '07*, pages 57–70, 2007.
12. H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, 87(4), 1990.
13. Hidden Markov Model Toolkit. http://htk.eng.cam.ac.uk/.
14. D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling Dominance in Group Conversations Using Nonverbal Activity Cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.
15. M. Kockmann, L. Burget, and J. H. Cernocky. Brno University of Technology System for Interspeech 2009 Emotion Challenge. In *Proceedings of Interspeech '09*, 2009.
16. J. Kukkonen, E. Lagerspetz, P. Nurmi, and M. Andersson. BeTelGeuse: A Platform for Gathering and Processing Situational Data. *IEEE Pervasive Computing*, 8(2):49–56, 2009.
17. M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell. Emotional Prosody Speech and Transcripts, 2002.
18. lightblue. http://lightblue.sourceforge.net/.
19. H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell. SoundSense: Scalable Sound Sensing for People-centric Applications on Mobile Phones. In *Proceedings of MobiSys '09*, pages 165–178, 2009.
20. M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in Its Natural Habitat: Manifestations and Implicit Folk Theories of Personality in Daily Life. *Journal of Personality and Social Psychology*, 90(5):862–877, 2006.
21. M. R. Mehl and J. W. Pennebaker. The Sounds of Social Life: A Psychometric Analysis of Students' Daily Social Environments and Natural Conversations. *Journal of Personality and Social Psychology*, 84(4):857–870, 2003.
22. E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell. Sensing Meets Mobile Social Networks: The Design, Implementation and Evaluation of the CenceMe Application. In *Proceedings of SenSys '08*, pages 337–350, 2008.
23. E. M. Morris, Q. Kathawala, K. T. Leen, E. E. Gorenstein, F. Guilak, M. Labhard, and W. Deleeuw. Mobile Therapy: Case Study Evaluations of a Cell Phone Application for Emotional Self-Awareness. *Journal of Medical Internet Research*, 12(2):e10, 2010.
24. A. S. Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, 2008.
25. Pyke. http://pyke.sourceforge.net/.
26. D. A. Reynolds. An Overview of Automatic Speaker Recognition Technology. In *Proceedings of ICASSP'02*, pages 300–304, 2002.
27. B. Schuller, G. Rigoll, and M. Lang. Hidden Markov Model-based Speech Emotion Recognition. In *Proceedings of ICME '03*, pages 401–404, 2003.
28. A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman. MLLR Transforms as Features in Speaker Recognition. In *Proceedings of Interspeech'05*, pages 2425–2428, 2005.
29. D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troester, and C. Haring. Activity and Emotion Recognition to Support Early Diagnosis of Psychiatric Diseases. In *Proceedings of Pervasive Healthcare '08*, pages 100–102, 2008.
30. Y. Wang, J. Lin, M. Annavaram, Q. A. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh. A Framework of Energy Efficient Mobile Sensing for Automatic User State Recognition. In *Proceedings of MobiSys '09*, pages 179–192, 2009.
31. D. Wyatt, J. Bilmes, T. Choudhury, and J. A. Kitts. Towards the Automated Social Analysis of Situated Speech Data. In *Proceedings of UbiComp '08*, pages 168–171, 2008.