# SocialSense: A Collaborative Mobile Platform for Speaker and Mood Identification

Mohsin Y Ahmed, Sean Kenkeremath, and John Stankovic

University of Virginia, Computer Science Department
Charlottesville, Virginia, USA
{mya5dm, stk4zn, stankovic}@virginia.edu

**Abstract.** We present *SocialSense*, a collaborative smartphone based speaker and mood identification and reporting system that uses a user's voice to detect and log his/her speaking and mood episodes. SocialSense collaboratively works with other phones that are running the app present in the vicinity to periodically send/receive speaking and mood vectors to/from other users present in a social interaction setting, thus keeping track of the global speaking episodes of all users with their mood. In addition, it utilizes a novel event-adaptive dynamic classification scheme for speaker identification which updates the speaker classification model every time one or more users enter or leave the scenario, ensuring a most updated classifier based on user presence. Evaluation of using dynamic classifiers shows that SocialSense improves speaker identification accuracy by 30% compared to traditional static speaker identification systems, and a 10% to 43% performance boost under various noisy environments. SocialSense also improves the mood classification accuracy by 4% to 20% compared to the baseline approaches. Energy consumption experiments show that its device daily lifetime is between 10-14 hours.

**Keywords:** social interaction, assisted living, depression, smartphone.

## 1 Introduction

Speaker identification systems based on in-home/on-body/smartphone microphones are used for various applications such as voice based authentication, home health care, security, and daily activity monitoring. With the pervasive usage of smartphones in everyday life, it is an exceptionally suitable unobtrusive platform for speaker identification reducing the overhead of on-body or contextual sensors. Besides speaker identification, speaker mood detection is another important problem in human interaction studies and social psychology research. The challenges of smartphone based speaker identification and mood detection include preserving user privacy, maintaining identification accuracy, accurate operation of the system irrespective of smartphone location, resilience against ambient noise and operating under energy constraints.

Both speaker and mood identification are part of a bigger and important health sensing problem, detection of human social interaction, which is an important indicator of mental and physical health in people of all ages. Regular good social interaction

brings many health benefits including reduced risk for cardiovascular and Alzheimer's disease, some cancers, osteoporosis and rheumatoid arthritis, steady blood pressure and reduced risk of depression and other mental disorders. On the other hand, social isolation culminates to loneliness and depression, physical inactivity and overall having a greater risk of death for older people. Therefore, a system able to detect people's social interactions and mood would be greatly beneficial for caregivers to more accurately diagnose and treat patients suffering from psychological disorders.

We present SocialSense, a collaborative smartphone based speaker and mood identification and reporting system which logs user speaking and mood episodes from his/her voice. A person can be uniquely identified by his smartphone Bluetooth ID. After SocialSense detects its user's speaking episode and mood, it broadcasts a message containing the user ID, the speaking episode timestamp, and corresponding mood to all neighboring phones via Bluetooth broadcasting. Thus every phone logs a global scenario of the social interaction environment. In a nutshell, SocialSense can answer the following questions:

- When is the phone user speaking?
- What is the mood of the user during speaking?
- With whom is the user speaking to? Who else are present around?
- When are the other persons in the environment speaking?
- What are the moods of other persons while they speak?

Besides detecting the smartphone user's mood, understanding the moods of other persons present in a social interaction is an important indicator of the global mood, hence the quality of the social interaction. Having this feature, SocialSense can potentially be used to demonstrate and verify the effect of mood contagion, i.e. how multiple individuals in a social interaction reach a mood convergence [1]. Using the idea of mood contagion, SocialSense can be used as a recommender system where it can recommend happy persons as potential conversation partners of sad persons to cheer them up. The actual use of SocialSense for mood contagion is outside the scope of this work.

Our prime target for the usage of SocialSense is in assisted living facilities for the elderly where the prevalence and magnitude of depression is of major concern. More than 1 million Americans reside in assisted livings presently. Studies found that, 20% to 24% of assisted living residents have symptoms of major or minor depression which is likely to cause physical, cognitive, and social impairment and delayed recovery from medical illness and surgery to these elderly. The scary fact is that, many depressive older adults end up committing suicide. Among men of age 75 and over, rate of suicide is 37.4 per 100,000 population. Several diagnostic barriers exist for the screening and treatment of depression in assisted livings which includes lack of regulatory requirements, privacy concerns, cost, and misinterpretation of depression. It is suggested that assisted living staff (nurse, therapist, medical director) should proactively assess for depressive syndromes instead of self-reporting of mood changes by the residents. SocialSense can be used as an automated diagnostic tool to monitor the mood and social interactions of the assisted living residents where each of the residents is provided with a smartphone with our system. [16]

Since SocialSense can capture the global scenario of a social interaction setting, it can be used as a data collection system for various social psychology and human interaction research. In addition, SocialSense incorporates a dynamic event-driven classification scheme for speaker identification. New people can enter into a social interaction while some people may leave at any time. SocialSense periodically refreshes its Bluetooth neighbor set and whenever it detects a change in the set, i.e., some people entered or left, it recreates the classification model based on the new neighbor set. For this purpose, it imports the user training feature files from the newly arrived phones to re-compute the classification model.

The main contributions of SocialSense are:

- An unobtrusive voice based speaker identification and mood detection system using user's smartphone. It does not use any on-body or contextual sensors thus contributing to mobility and user-friendliness.
- A practical, easy, and short training scheme to train a phone to detect a person's own speaking episodes. One key novelty of SocialSense is that it avoids the need of exhaustive training by all users in a social interaction setting and still accurately detects all speakers by collaboration among the phones.
- SocialSense has privacy support for users. No audio samples are recorded or stored in the phone and features are extracted in real time and after classification they are removed from the system. There is no way to reconstruct the original audio signal at a later time from SocialSense.
- SocialSense's voice based mood detection module in every phone is conventional, however by collaboration among the phones, it can detect the mood of the members of a conversation group and the change of one's mood when he/she switches between conversation groups, i.e., demonstrate mood contagion. This novel idea hasn't been explored before and such a system would be invaluable for further experiments on social mood dynamics. Also, using a random forest classifier for mood detection compared to GMM and SVM classifiers used in baseline systems [7, 8], our system has a 4% to 20% increase in accuracy compared to the baselines.
- SocialSense supports real life environments where new people enter and existing people leave the social interaction environment. SocialSense periodically refreshes its Bluetooth neighbor set to detect such changes in the environment.
- Another novel feature of SocialSense is its dynamic event-driven classification scheme where it performs speaker identification using an up-to-moment classifier based on the current users present in the scenario. This yields an average 30% increase in classification accuracy compared to static classification.
- Evaluation with respect to noisy environments has been performed by injecting various artificial noise to simulate real life noisy environments and results demonstrates that SocialSense improves speaker identification accuracy in noise by 10%-43% based on different types of noise and mood detection accuracy in noise by 33% compared to the state-of-the-art systems. SocialSense has been evaluated by training with noise to yield these performance boosts, which hasn't been done in baseline approaches.

## 2 Related Work

Many of the existing speaker identification systems require the total number of speakers to be static, and they employ static classification schemes so that each speaker needs to train the system beforehand, which makes them less realistic [2, 3]. SocialWeaver [4] uses a multi-level classification for speaker identification. The first level uses energy histogram classifiers while the second level uses a GMM based classifier. Neary [5] uses similarity of sound environment to detect conversational fields. These energy and loudness based approaches have greater error in noisy conditions and they fail if there is a person present in the scenario without his phone. SpeakerSense [6] is a speaker identification platform built on a heterogeneous multiprocessor architecture. It attempts to reduce training overhead by training from real life events as phone calls and one-to-one conversations, but does not evaluate the system in noisy environments. Also, it requires the total number of speakers to be static and does not support realistic dynamic environments where speakers enter and leave on the fly.

There are a number of existing systems which detect user's mood from voice. EmotionSense [7] provides dual systems for speaker identification and emotion detection from user's voice using Gaussian mixture methods. [8] provides SVM based classifiers that recognize human emotional states from their utterances. However, these systems can only capture mood of a single person or entity and, therefore, are not suitable for social psychology experiments where a system would need to know moods of everyone in a social interaction. Also, there is no evidence that these systems would operate well under real life noisy environments.

Besides speaker identification and mood detection, there have been systems which detect other aspects of social interaction using different modalities. Some of the existing work on social interactions uses only on-body sensors such as accelerometers, gyroscopes, GPS, microphones, and cameras. Pierluigi et al. [9] built a badge having a triaxial accelerometer and a JPEG camera which is used to detect the presence of other people. Crowd++ [18] estimates the number of people talking in a certain place by unsupervised machine learning technique from smartphone audio inference. CenceMe [10] can automatically detect activities of individuals and share the sensing results through social networks.

Another type of work uses ambient sensors. [11] uses a sociometric badge equipped with infrared transmitter/receiver and microphone which senses and models human networks. In [12] four video cameras and audio collectors are placed in public areas such as the dining room, living room and hallway which can detect high-level social interactions among people such as greeting, standing conversation, and walking together.

We compare SocialSense with some state-of-the-art smartphone based sensing systems in table 1.

**Table 1.** Comparison of State-of-the-art

| System | Operations | Classifiers used | Results |
|---|---|---|---|
| EmotionSense [7] | Speaker identification, mood detection | GMM | 90% speaker ID accuracy, 70% mood detection accuracy |
| SpeakerSense [6] | Speaker identification | GMM | 95% speaker identification accuracy |
| Social Weaver [4] | Speaker identification, conversation group clustering | Loudness histogram, GMM | 90% speaker ID accuracy, 70-90% accuracy for conversation clustering |
| Neary [5] | Detect conversational fields i.e detecting multiple persons who are in a conversation | No classifier | 96.6% precision and 67.9% recall achieved in a controlled experiment |
| Qiang et al [13] | User activity, speaker ID, proximity, location | Naive Bayes, Discriminant, Boosted tree, Bagged tree | 92% accurate speech detection |
| Social Sense | Speaker identification, mood detection, mood contagion sensing | Logistic regression, Random forest | 94% speaker ID accuracy, 90% speaker ID accuracy in noise, 80% mood detection accuracy, 76% mood detection accuracy in noise |

## 3 SocialSense System Design and Operation

The assumption behind SocialSense is that every user in a social interaction setting carries his/her own phone with the SocialSense app running in it. However, if one or more persons is present without his phone, only his speaking and mood episodes will remain undetected and unreported, while all other users' speaking and mood episodes will be detected and broadcasted without any error.

Figure 1 shows the system diagram. The SocialSense app runs continuously in each phone listening to audio streams. Silent frames are detected by comparing each frame's energy to a threshold, and filtered from further processing to save energy. Each phone periodically updates its phone-set within its Bluetooth proximity range (~10 m). It is required that the system meets the energy constrains of mobile devices in order to make it usable in realistic scenarios. SocialSense is capable of running for 10 to 14 hours continuously in smartphones and tablets which is good enough for its usage as a healthcare, research and data collection tool in assisted living. SocialSense is made up of a number of modules described in the following sections.

### 3.1 Phone-set Formation

A phone's phone-set is defined as the set of phones running the SocialSense app situated within the Bluetooth proximity range from that phone. This module running in every phone refreshes its phone-set periodically (generally every 30s) to keep the most recent neighboring phones in its phone-set. The periodic interval is set so that it is neither too short to trigger redundant phone-set discovery process nor too long to miss significant changes in the phone-set, considering realities of human social interaction. All members of a phone-set are assumed to be close enough to participate in a conversation. Conversely, phones not belonging to the phone set are assumed to be not participating in a conversation.

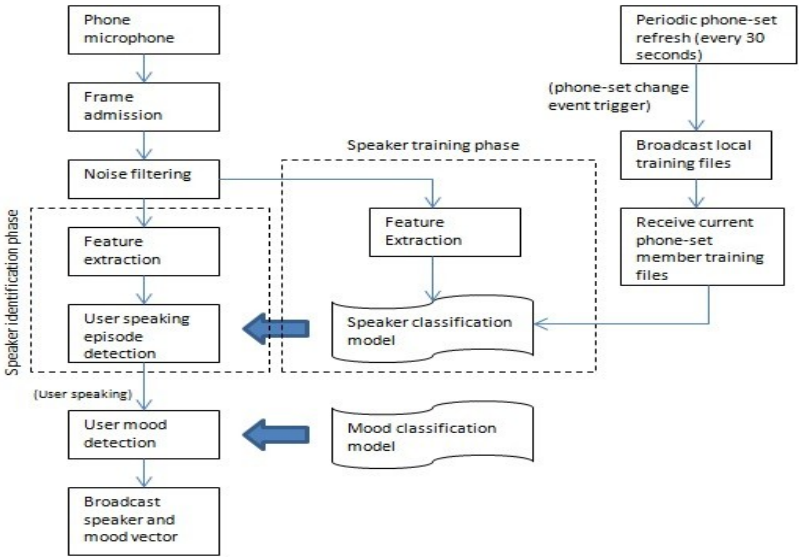



**Fig. 1.** SocialSense block diagram

### 3.2 Speaking Episode Detection Module

This module determines whether a voice segment belongs to the phone user or not. The speaker identification is a binary classification problem where every non-silent audio segment must be classified into one of two classes: "phone user's voice" or "anything else" (e.g. others' voice or ambient noise). It uses a dynamic logistic regression based classifier, which can be easily trained by the user (or support personnel in assisted living). The user trains the speaker classification system by speaking for 60 seconds in front of the phone in normal tone and loudness. This simple, easy-to-use and short self- training scheme allows the classifier being updated with the latest voice samples of the user. In assisted living facilities, this training will be done by the staff.

Some existing smartphone based speaker identification systems classify speakers based on the loudness of the perceived audio signal [4], [13]. The hypothesis behind

those works is that, a user's voice is loudest in his own phone in a particular time instant compared to any other neighboring phones at the same time (as the user is supposed to be the closest person to his phone). However, this scheme doesn't work well in noisy environments, and also in the situation where a person without any phone is talking with people having their phones. In the latter case, when the person not having his phone is talking, his voice will be loudest to the person's phone who is closest to him, so that phone will incorrectly assume that the person without his phone is its user and classify positively, which is incorrect. Other systems like SpeakerSense [6] require training a speaker model for each individual who needs to be recognized, thus incurring large training overhead and resulting in complex, power-hungry classifiers.

SocialSense, on the other hand, uses a simple logistic regression based binary classifier with very little training overhead using 39 MFCC (mel-frequency cepstral coefficient) features. The phone-user (or staff) can train the system easily by speaking for 60 seconds in front of the phone in normal tone and volume to create a speaker classification model. As human voice may occasionally change depending on his physiological state, using this easy-to-use training scheme, the system can detect when its user is speaking irrespective of his voice quality, in the presence of noise and even when a person without his phone is present in the scenario as well. Unlike volume based systems, SocialSense does not fail when a user is present without his phone. Only his speaking and mood episodes remain undetected, but the systems in other users' phones work fine. The presence of a user without his phone does not incur any error or failure in the overall system operation.

### 3.3 Mood Detection Module

Detecting speaker mood in a mobile platform is a major challenge in this work. If a voice segment has been classified as a user's voice by the speaking episode detection module, this module further determines the user's mood (happy, sad, angry, neutral) from his voice. Then it generates a speaking and mood vector consisting of the starting and stopping timestamp of the user's speaking episode and mood during that speaking episode. This module extracts 39 MFCC coefficients from each user utterance window and calculates 9 different statistics on each MFCC coefficient culminating to 351 audio features. These statistics are: geometric mean, harmonic mean, arithmetic mean, range, skewness, standard deviation, z-scored average, moment and kurtosis. The MFCC coefficients combined with these statistics carry a large amount of prosodic and energy based information correlated to emotion. It then uses these features to train a random forest classifier from the EMA emotional utterance dataset [15] for detecting mood.

### 3.4 Message-Exchange Module

SocialSense forms a Bluetooth network among all members of a phone-set. When a phone has a speaking and mood vector to send, it broadcasts the vector using flooding over the network. It has an incoming thread and an outgoing thread to handle incoming and outgoing messages, respectively. It maintains a message queue, new vectors

to be broadcasted are enqueued in the queue and the outgoing thread sends vectors one by one from the queue.

### 3.5 Dynamic Event-Driven Classification Module

For speaker identification, the logistic regression classifier uses a positive training file to keep training samples from the phone's user, and uses another negative training file to keep training samples from all other users. During startup of a conversation, SocialSense broadcasts its local positive training file to all neighbors which they use for their negative training. If there are 4 phones in the scenario, each phone uses its local file for positive training and 3 other files received from others for negative training. The phone-set discovery process triggers every 30 seconds to refresh the phone-set. If there is a change in the phone-set during a periodic phone-set refresh (an old user leaves or a new user enters), an event is triggered. When the event triggers, each phones broadcasts its positive training file over the network and updates its negative training using only the files received from phones present in the current scenario, and then rebuilds an updated classifier for speaker identification. This improves classification accuracy by 33% on average compared to static training and makes the training process for each user simple, which is shown in the evaluation section.

## 4 Evaluation

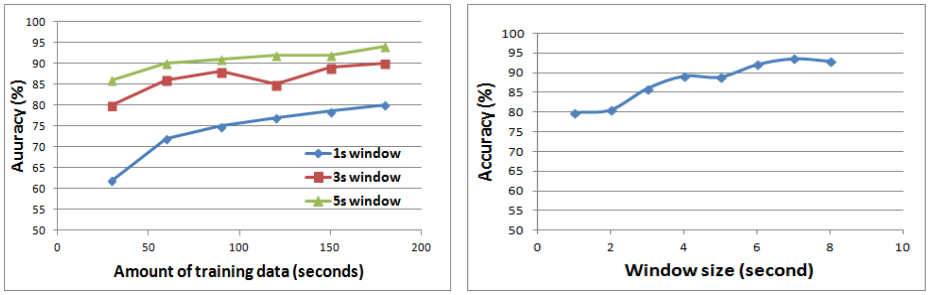
The evaluation consists of multiple parts. First, we evaluate how accurate SocialSense is in identifying speakers. Then we evaluate the effectiveness of mood identification. We have done these evaluations in quiet and noisy (artificially injected) environments, showing that training with noise in noisy environments yields good increase in performance. We have demonstrated the impact of window size, amount of training data, and dynamic classification for speaker identification. We also compare our results with some state-of-the-art solutions.

### 4.1 Speaker Identification Evaluation

We have evaluated the performance of SocialSense's speaker classification module in terms of the classification accuracy, which is the overall correctness of the model. The data for these evaluations are taken from voice segments collected from 7 persons. There were 4 females and 3 males among them. A 1.5 hour long conversation on various random topics between two of these females was recorded by us. Another 6 conversations, each around 5 minutes in length, between a male and a female, were collected from the internet. We collected 3 solo speech recordings from the remaining 3 persons for 10 minutes each. We extracted individual voice recordings from each of these 7 speakers separately from these recorded conversations and simulated 2, 3, 4 and 5 person conversations from these. We performed all the speaker identification experiments from these simulated conversations. For example, for simulating 3 person conversations, we trained the logistic regression classifier with one person as

positively trained, and the other two persons as negatively trained, with all 3 combinations of three persons, and all 35 possible selections of 3 persons from a set of 7 persons.

**Training Size.** Intuitively speaker identification accuracy increases with the increase of training data, as the classifier can encode more information with a longer training. This phenomenon is shown in figure 2. The training and testing data were taken from voice samples collected from 7 persons, with 2, 3, 4 and 5 person simulated conversations. The accuracy for 3 separate window sizes is shown for training up to 180 seconds.




(a) Amount of training data (b) Window size

**Fig. 2.** Speaker identification accuracy vs. (a) Amount of training data; (b) Window size

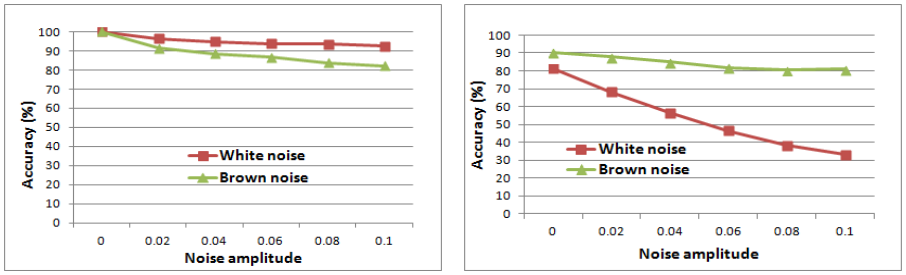
As we can see from figure 2(a), there is a sharp increase in accuracy between 30s and 60s of training, and beyond that the accuracy increases slowly. Also the accuracy is highest for a 5s window size. These values are the lower bounds on the training data needed to accurately identify the speaker on the phones, i.e. a minimum of 60 seconds of training is required with a minimum of a 5 second window size.

**Window Size.** This test was conducted on 2 person conversations. One person was trained as positive while the other was trained as negative for 60 seconds. The window size was varied from 1 to 8 seconds and each window was classified using the logistic regression classifier.

Results from figure 2(a) and 2(b) suggest that, a window size between 5 to 7 seconds is optimal for speaker identification. 3-4 second long window sizes yield accuracy of 86-89% which is acceptable. 5-7 second long window sizes can be used for warm conversations where each speaker talks for a long time before switching turns, while a 3-4 second window can be used for cold conversations with frequent turn-takings with short speaking duration in each turn.

**Effect of Noise.** Noise is a very important and realistic issue to consider to evaluate a smartphone based speaker identification system. It is very likely that users will move with their phones to different places (both indoor and outdoor) engaging in social interactions. Therefore, the system must be able to correctly identify its speaker under various types of noise.

Evaluation has been done to test the effect of artificial noise on speaker recognition accuracy. These tests were also done using 2 person conversations collected from 7 speakers. We used Audacity [14], an open-source sound editing software to inject artificial white and Brown noise into voice samples, and observed classification accuracy under different levels of noise. White noise is quite similar to television static or the humming of an air conditioner and Brown noise is similar to gusty wind. Therefore, these artificial noises can simulate real indoor and outdoor noisy environments.




(a) Similar train-test (b) Different train-test

**Fig. 3.** Effect of noise on speaker recognition accuracy, (a) With similar train-test set; (b) With different train-test

Figure 3(a) shows the effect of white and Brown noise on SocialSense speaker recognition accuracy with similar train test set. During no noise, the accuracy is best at 100%, while during maximum noise, the accuracy degrades to 82%, which is an 18% drop. However, because the train and test sets are similar in this case, this is not a realistic scenario. Figure 3(b) shows the effect of noise under different train-test sets. Here, the best accuracy during no noise is 90.2% and the worst accuracy is only 33%, which is a shocking 57% drop, and demonstrates how the system will fail in presence of noise, if no measure is taken.

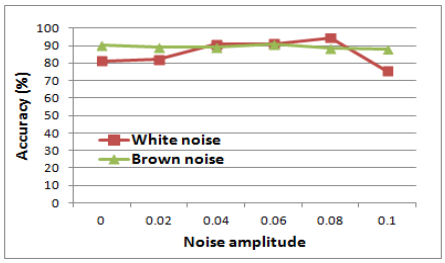



**Fig. 4.** Effect of noise on speaker recognition accuracy, with training in noise

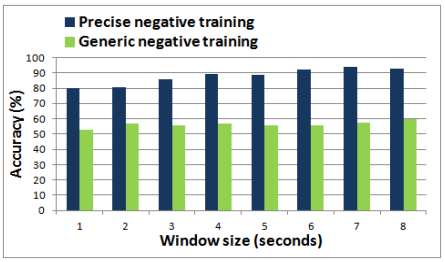



**Fig. 5.** Effect of dynamic training vs. generic training

It is a design characteristic of SocialSense that a phone user can have his own phone trained for detecting his speaking episodes. This adds a lot of flexibility to the system. In noisy environments, the user can have his phone trained in noise to en-

hance speaker identification accuracy. Because of the short training session and each user needing to train only himself (as opposed to other systems where all user need to train every phone), the training overhead is low.

Figure 4 shows the effect of noise when the training is done in noise as well. It shows that even in worst noisy conditions the accuracy drops to 76% for white noise and 89% for Brown noise, which is a 43% performance boost for white noise and 10% boost for Brown noise. We have limited the noise amplitude to 0.1 in these experiments as this level is commensurate to real life extremely noisy environments.

**Effect of Dynamic Classification.** Because of the dynamic event-driven classification scheme in SocialSense, every phone is trained with a precise negative training set comprised of the voices of all other persons present in the social interaction setting. The phones update their training files by message exchange whenever a new person enters or leaves the Bluetooth range. Without the dynamic classification, every phone had to use a generic negative training comprising of generic voices from arbitrary persons, since no apriori knowledge of the users is available.

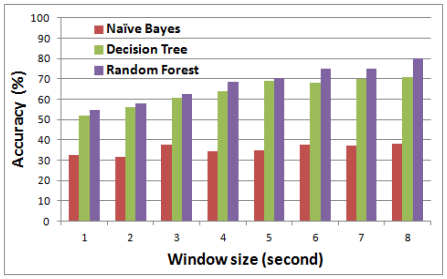
We used voice samples from 5 persons for precise training, with 1 trained as positive and other 4 trained as negative. The testing samples had voices from all 5 persons. For generic training, we used a separate voice collection from 3 people (The EMA dataset [15]) for negative training, and used the same test set as precise training.

The performance comparison of precise and generic training is shown in figure 5, which shows a significant classification improvement (30% on average) due to dynamic training. Consequently, this novel aspect of our solution results in a major performance improvement.

**Worst Case Analysis of Dynamic Speaker Classification.** The phone-set refresh process triggers once in every 30 seconds. If there is a change in the phone-set immediately after a refresh process, all the phones will stay with outdated classifiers for 30 seconds in the worst case. There are 2 cases to consider: i) some new phones arriving, ii) some existing phones leaving. In the first case, if some new phones arrive right after the refresh process, they will remain unknown to the existing phones for 30 seconds until the next refresh process. In this time, the newly arrived phones cannot send or receive any vectors, so their social interactions will not be logged. Also, during this time, the newly arrived phones will have a blank negative set, so all persons' speaking episodes will be considered as positive in these phones. To avoid classification errors due to the initialization in the newly arrived phones, voice segments during this initialization window are ignored. The second case, where some existing phones leave right after the refresh process, is less complicated than the first case. In this case, all the remaining phones will stay with redundant negative sets, containing trainings from people who doesn't exist anymore. But, there will be no classification error in these phones unlike the first case. For both cases, situation comes back to normal in at most 30 seconds, after the immediate next phone-set refresh.

### 4.2 Mood Detection Evaluation

Because of the difficulty associated to get real life data for mood evaluation, we performed both training and testing from the Electromagnetic Articulography dataset [15], which contains 680 acted utterances of a number of sentences in 4 different emotions (anger, happiness, sadness and neutrality) by 3 speakers. We used 3 different classifiers to model each mood using MFCC features with 9 statistics (total 351 features), naive Bayes, random forest, and decision tree. We also varied the acoustic window size from 1 to 10 seconds.




(a)

| Real mood | Classified as | | | |
|---|---|---|---|---|
| | Angry | Happy | Neutral | Sad |
| Angry | 144 | 19 | 1 | 2 |
| Happy | 32 | 110 | 4 | 18 |
| Neutral | 4 | 11 | 125 | 16 |
| Sad | 1 | 4 | 23 | 154 |

(b)

**Fig. 6.** (a) Effect of audio sample length on emotion recognition accuracy with various classifiers; (b) Confusion matrix for emotion recognition with random forest classifier

The random forest classifier yielded best cross classification results for 10 folds, as shown in figure 6(a). This classifier resulted in a 4% to 10% increase in accuracy compared to the baseline EmotionSense [7] with varying window size, and a 20% increase for speaker independent model compared to [8]. The results for the baselines are taken from corresponding existing works. The figure demonstrates that mood classification accuracy increases with increasing window sizes, however beyond 6s window size it becomes stable and doesn't change much. The confusion matrix for the random forest classifier is shown in figure 6(b).

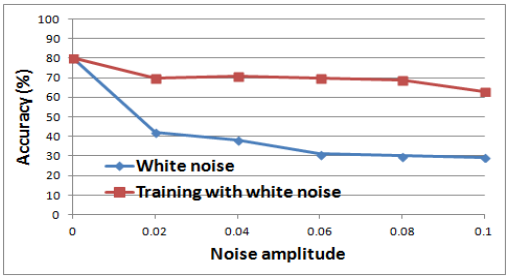



**Fig. 7.** Effect of noise and improvement with training with noise for mood classification

Similar to the speaker identification module, we evaluated the performance of mood detection module under noise. The effect of white noise has been noticed as to

be more detrimental than brown noise, so we have done this experiment for white noise only. We injected white noise with amplitudes varying from 0.02 up to 0.1 into the mood dataset. We trained with mood utterances without noise and tested with utterances in noise.

As expected, the performance dropped drastically, as shown in figure 7. However, similar to the speaker identification module, we trained the mood dataset in noise and performed a 10 fold cross validation, which yielded a 33% performance boost in the worst 0.1 noise amplitude, as shown in figure 7. The existing mood detection systems hasn't evaluated the possibility of training with noise, which in our case, yielded a significant increase in performance.

### 4.3 Energy

SocialSense consumes energy in two ways: (i) idle listening, and (ii) once a speech episode is identified, it runs various modules and classifiers. Experiments were run to determine the lifetime of tablets and smartphones running SocialSense. The least energy cost for SocialSense is if it is idle listening and there are no speech episodes to process. Our experiments showed that Nexus 7 tablet ran for 14 hours and the HTC one smartphone ran for 12 hours for this *best* situation. When SocialSense is actively processing speech episodes there are 5 modules in the system which consume the majority of energy: i) acoustic processing and feature extraction, ii) logistic regression speaker identification classifier, iii) random forest mood detection classifier, iv) speech and mood vector transmit/receive, and v) periodic phone-set refresh, training file exchange and classification model file recreation. In a second set of experiments we modified the system to run all these modules continuously as if there was continuous speech. This is the worst case in regards to energy costs. In these experiments the Nexus 7 tablet ran for 12 hours (down from 14 hours) and HTC one smartphone ran for 10 hours (down from 12 hours). Consequently, SocialSense can operate between 12-14 hours on a tablet, and between 10-12 hours on a smartphone. This demonstrates that SocialSense can indeed be used as a healthcare device in assisted living since such devices can be charged over night.

## 5 Discussion

SocialSense detects speaking episodes and the mood of a user, and by collaboration it imports the speaking episodes and moods of the neighboring users as well. A user interface can be built upon this fine grained information showing the social interaction history of a user within a particular time-frame. Such a user interface will be able to display a user's common conversation partners, his amount of participation and engagement during a conversation with a particular partner, his mood during a conversation and hence mood during that time of that particular day, change of his mood with time or change of conversation partner and so on. Many of these quantities are of interest to psychologists when they treat a potentially depressive patient, and hence ask him relevant questions. The patients' answers are often vague, confusing and er-

roneous because most of the time they do not remember their social interaction history and mood for a very long time. SocialSense can eliminate the need for these oral questionnaires and hence avoid all the errors as it logs the social interaction data of a user with his moods. Therefore, this system can be used in places like an assisted living facility where depression and related psychological disorders are common among the occupants.

**Robustness.** As we argue that SocialSense is usable among the elderly in assisted living facilities, we are aware of the fact that the elderly are prone to forgetfulness, and it is very likely that they may sometimes forget to carry their phones during a social interaction. Though SocialSense is most accurate if every person carries his smartphone in order to detect everybody's speaking episodes and moods, the system does not break down if such assumption is violated. If a person does not carry his phone during a social interaction, his own speaking and mood episodes will remain undetected and unreported and others will not have his information for complete mood contagion. All the other persons' speaking and mood episodes will be detected and reported correctly. This is a major system design enhancement compared to volume based systems [4, 13] which fail when one or more persons forget to carry their phones. It is also important to note that overall diagnosis involves many conversations over multiple days and some missing information when smartphones are forgotten or turned off does not necessarily cause problems.

**Training in Assisted Livings.** SocialSense's easy to use individual training scheme and adaptability to noisy environments is very suitable for its usage in assisted living. We have shown in the evaluation section that it only takes 60 seconds of training for the system to work in any particular environment. Assisted living residents generally pass specific time of their days in specific locations (e.g., mornings in the hall room, noon at lunch room, afternoon in the garden). The assisted living support person can train the smartphone for each of these common environments. If a resident moves to a new location where the system needs to be retrained because of different noise levels, the support person can do the training very easily with 60 seconds of data.

**Mood Contagion.** Using SocialSense it is possible to detect not only the mood of an individual user, but also the moods of others present in the social interaction setting. According to the best of our knowledge, no such system has been built yet which can detect such a global mood. Thus, SocialSense can be used as a platform to verify and conduct experiments on *mood contagion* which is a psychological process by which a group of people engaged in a social interaction reaches emotional convergence, i.e. they all have similar feelings after a certain time though their initial feelings may be different. It is hypothesized that interventions based on knowledge of mood contagion can be used to help treat depression in the elderly.

**"In-Phone" vs. "In-Cloud" Scheme.** We adopted an "in-phone" processing scheme as opposed to "in-cloud" processing as in [17]. The term "in-phone" means that all data acquisition, feature extraction, and classification are performed in the phone

itself. A reasonable alternative to or solution is an "in-cloud" solution, where unprocessed raw data (conversation recordings) or semi-processed data (features) are sent to a central server where a web service performs further processing and classification. However, the "in-cloud" approach requires connectivity to the internet by wi-fi or 3G which is not always available or is sometimes unreliable. To handle the unreliability of connections various buffering and upload schemes have to be developed. A high-speed 3G/4G connection also imposes additional operating cost for each phone. The "in-phone" approach is cheaper and better supports mobility and could be used even when residents are away from the assisted living facilities.

**Concurrent Speaking Episodes.** In our experiments described above, we assumed that users did not speak concurrently. In reality, speakers do speak concurrently on some occasions. So we also evaluated our system to test how it performs when users speak concurrently. Ideally, when two or more users are speaking concurrently, each of their systems should detect their own speaking episodes and log them as "speaking" in their individual phones. We performed experiments with 4 speakers (2 male, 2 female), with two concurrent speakers at a time for all 6 possible pairs of conversations. As expected, the system performance degraded. On average, SocialSense was 55% accurate in detecting a particular user's speaking episode when 2 concurrent users were speaking. While this sounds low, this result only applies to the portion of the speaking episode when there is actual concurrency, e.g., when two people first both start speaking (but then one usually backs off) or when someone interrupts a speaker.

## 6 Conclusion

This paper presents the design, implementation, and evaluation of SocialSense which is a collaborative mobile platform for speaker identification and mood and mood contagion detection from users' voice. Aside from its ability to recognize speaker and mood with significant accuracy, we have demonstrated its performance relative to the amount of training data and length of window size, culminating in an optimal benchmarking of these parameters. We provide empirical evidence that SocialSense performs well under various noisy environments when trained with noise, with an easy-to-use training scheme. Also, with a dynamic classification scheme, SocialSense is 30% more accurate in speaker identification compared to generic training with static classification. SocialSense is 4%-20% more accurate in speaker independent mood sensing compared to the baseline state-of-the-art mood sensing systems. It was also shown that SocialSense lifetime on various devices is between 10 to 14 hours.

**Acknowledgments.** This work was supported, in part, by NSF Grants CNS-1319302 and CNS-1239483, and a gift from PARC, Palo Alto. We cordially thank the reviewers for their insightful comments and suggestions.

## References


1. Neumann, R., Strack, F. (2000) Mood Contagion: The automatic transfer of mood between persons. Journal of Personality and Social Psychology, Vol. 79, No. 2, pp. 211-223.
2. Reynolds, D. A. (1995) Speaker identification and verification using gaussian mixture speaker models. Speech Communication, Vol. 17, Issue 1-2, pp. 91 - 108.
3. Reynolds, D. A., Rose, R. C. (1995) Robust text-independent speaker identification using gaussian mixture speaker models. Transactions on Speech and Audio Processing, Vol. 3, Issue 1, pp. 72 - 83.
4. Luo, C., Chan, M. C. (2013) SocialWeaver: collaborative inference of human conversation networks using smartphones. 11th ACM Conference on Embedded Networked Sensor Systems (SenSys), Roma, Italy.
5. Nakakura, T., Sumi, Y., Nishida, T. (2009) Neary: conversation field detection based on similarity of auditory situation. 10th Workshop on Mobile Computing Systems and Applications (HotMobile), Santa Cruz, California, USA.
6. Lu, H., Brush, B., Priyantha, B., Karlson, A. K., Liu, J. (2011) SpeakerSense: Energy efficient unobtrusive speaker identification on mobile phones. IEEE Pervasive Computing and Communication (PerCom), Seattle, Washington, USA.
7. Rachuri, K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., Aucinas, A. (2010) EmotionSense: A mobile phone based adaptive platform for experimental social psychology research. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), Copenhagen, Denmark.
8. Yu, C., Aoki, P. M., Woodruff, A. (2004) Detecting user engagement in everyday conversations. 8th International Conference on Spoken Language Processing, South Korea.
9. Casale, P., Pujol, O., Radeva, P. (2009) Face-to-face social activity detection using data collected with a wearable device. 4th Iberian Conference on Pattern Recognition, Portugal.
10. Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., Campbell, A. T. (2008) Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. 6th ACM Conference on Embedded Networked Sensor Systems (SenSys), Raleigh, North Carolina, USA.
11. Choudhury, T. (2004) Sensing and modeling human networks. Ph. D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology.
12. Chen, D., Yang, J., Malkin, R., Wactlar, H. D. (2007) Detecting social interactions of the elderly in a nursing home environment. ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 3, No. 1, pp. 1–22.
13. Li, Q., Chen, S., Stankovic, J. A. (2013) Multi-modal in-person interaction monitoring using smartphone and on-body sensors, IEEE International Conference on Body Sensor Networks, Cambridge, MA, USA.
14. Audacity. http://audacity.sourceforge.net/.
15. Kim, J., Lee, S., Narayan, S. S. (2010) An exploratory study of manifolds of emotional speech. Acoustics Speech and Signal Processing, Dallas, TX, USA.
16. Stefanacci, R. G. (2008) How big an issue is depression in assisted living? Assisted Living Consult, Vol 4, No. 4, pp 30-35.
17. Miluzzo, E., Cornelius, C. T., Ramaswamy, A., Choudhury, T., Liu, Z., Campbell, A. T. (2010) Darwin phones: the evolution of sensing and inference on mobile phones. 8th International Conference on Mobile Systems, Applications, and Services (MobiSys), San Francisco, California, USA.
18. Xu, C. et al (2013) Crowd++: unsupervised speaker count with smartphones. ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland.