

VibeFones: Socially Aware Mobile Phones

Anmol Madan, Alex “Sandy” Pentland
MIT Media Lab
 {anmol,pentland}@media.mit.edu

Abstract

Today's mobile phones are essentially off-the-shelf, mass-market wearable computers. In this paper, we describe mobile social software that uses location, proximity and tone of voice to create a sophisticated understanding of people's social lives, by mining their face-to-face and phone interactions. We describe several applications of our system – automatic characterization of social and workplace interactions, a courtesy reminder for phone conversations, and a personal trainer for dating encounters.

1. Introduction

The mobile phone will soon be the most popular consumer device on the planet. About half of the 800 million phones sold last year had significant processing power (more than a Pentium I), broadband data connectivity, external storage, integrated cameras and multimedia playback [12]. These are exciting times for wearable computing; for today's mass market mobile phone is really the researcher's wearable computer in disguise. People carry their mobile phones for most of their day, which makes them ubiquitous wearable sensors that can collect continuous, long-term, behavioral and social data, invaluable for pattern recognition and predictive modeling. Table 1 below compares our off-the-shelf mobile phones to leading wearable computing platforms of the last five years [28, 29].

Mobile social software can be defined as software that augments and mediates a user's social and collaborative abilities on a mobile phone or wearable device [8]. Most commercial applications are *social* in a very limited sense of the word – while they connect users and support sharing of information, they understand very little about the user or the nature of the interaction itself.

| Feature | ETH QBIC (2003-06) | MIT MiThril (2003-05) | VibeFone Motorola Phone |
|------------------------|-----------------------------|---------------------------------|------------------------------------|
| Processor and OS | 230-400 Mhz XScale, Linux | 230 Mhz StrongARM Zaurus, Linux | 312 Mhz XScale mobile phone, Linux |
| Connectivity | Wifi, optional cellular | wifi, optional cellular | Cellular, optional wifi |
| Peripheral Sensors | wired and Bluetooth sensors | wired and Bluetooth sensors | Supports Bluetooth sensors |
| Location and Proximity | GPS, bluetooth | IR tags, Bluetooth | Cell ID, bluetooth, optional GPS |
| Data storage | SD card > 4 GB | SD card > 4 GB | SD card > 4 GB |
| Video Capture | External | External webcam | Integrated 1.2 Mpx |

Table 1. Comparison of mobile phones with wearable computing platforms

We propose that mobile applications can become *socially aware* by measuring non-linguistic speech attributes (e.g., tone of voice, social signals) and interaction metadata (e.g. physical proximity). The juxtaposition of social network information with social signaling in voice creates compelling opportunities. An application that can gauge the user's interest in a conversation or experience can auto-annotate that interaction [4,18], make decisions about when to deliver relevant multimedia content [14] or share the interaction metadata with the user's friends or colleagues [10,17]. With real-time models and classification of the user's behavior, the mobile phone can even advise the user in specific contexts like dates or negotiations. Some of these ideas have been implemented on our system and are described in section 4.

2. The VibeFone Platform

2.1 Hardware and Software

The Motorola A780 and E680i are (proprietary) Linux-based mobile phones with 312Mhz XScale processors, touch-screen interface, quad-band GSM connectivity, Bluetooth radio, stereo audio out, Qtopia graphical UI, memory card slot and SMS / MMS



Figure 1. VibeFone application on an e680i mobile phone. The display fields from top to bottom are ‘mood’, ‘location’ and ‘buddies’.

functionality. One version of the A780 phone also has an in-built GPS receiver; although a long initial lock time (several minutes) limits its applicability.

We have added the following data collection and feedback enhancements to create the VibeFone application:

- Device-driver code to capture voice streams for the user (both during phone calls and off-line)
- Sense when a call is in progress, and device driver code to capture audio stream for the person on the other end
- Detect voiced and speaking regions in real-time using a multi-layer HMM, and extract activity and stress measures (from section 3) for the speaking segments
- Periodically scan for Bluetooth devices in proximity and current cell tower identifier; compare scan results to a list of buddies and known locations
- Automatically send SMS/MMS messages with known buddies to share the users’ social interaction metadata (if enabled with appropriate permissions)
- Graphical User Interface for feedback and

user interactions

- BlueZ (upgrade of native Bluetooth driver) and wireless 802.11 support (cross-compiled kernel to support SD cards)
- Capability to prompt the user to take pictures of an important event

Although the low level components of our platform are proprietary to Motorola, application layer software is available under the GNU GPL on our website [16].

2.2 Privacy and Feedback

The VibeFone application can access voice streams on the phone, but has been implemented such that only the statistics of various speech features are saved, for privacy reasons. Although we can quantify speech for the person at the other end of the line, it is essential that this done only with informed consent. We therefore do not analyze the other person’s speech because of these privacy concerns. The VibeFone application starts when the phone boots, but can be easily enabled or disabled from the main screen.

A complementary issue is the appropriateness of feedback. It is important to convert the feedback into human usable form, as well as deliver it such that it does not distract or interfere with the user’s current activity (for example, feedback of the form “your spectral entropy is rather high” is almost useless for the user). We have attempted to make the feedback messages simple and useful, and they can be displayed on the screen or played as audio on the headset.

2.3 User Scenarios

The VibeFone application starts when the mobile phone is switched on. Unless explicitly disabled, it periodically scans for other bluetooth devices and cell tower identifiers. The scan time is variable but is typically set at 5 minutes, as a compromise between granularity of interaction data and battery life.

There are two usage models for speech feature processing. The first model is designed to capture face-to-face interaction data. When the VibeFone application detects a buddy in the vicinity, it invokes the speech feature analysis engine and extracts and saves statistical measures of the users’ vocal activity and emphasis. The user is required to wear a noise-canceling headset (and is occasionally reminded by the VibeFone application to do so). During an initial training phase, the user is prompted for training labels

after predefined events and intervals. This generates rich interaction data which is maps measured proximity and voice features to the users' labels.

The second usage model is designed to capture phone interaction data. The application can automatically detect when the phone is being used and start the speech feature extraction engine. The caller ID is saved, and during training the user is prompted for a label at the end of the conversation. This usage model yields less rich data, but is marginally more convenient for the user as an external headset is not required.

The VibeFone also has two special applications – the Jerk-O-Meter and the Wingman3G. These have to be explicitly invoked by the user and are described in the next section. In these applications, the VibeFone evaluates the user's speech and provides real-time feedback, almost like a personal trainer or coach.

3. Background and Features

3.1 Social Proximity

Social contact with people and places is reflected on a mobile phone in many ways – co-location from bluetooth or short-range radio, precise GPS location, approximate location from cell tower identifiers and even call logs as metrics of social links and their strength. Eagle and Pentland used the Bluetooth radio functionality in mobile phones to recognize activity patterns, infer social relationships and model organizational rhythms [10]. Alternative methods of tracking co-location for social groups that have been used in the past include infra-red (IR) beacons to capture face-to-face interactions, short-range radios (same room or office) or GPS location-based services [17, 20, 24]. Gips and Pentland [14] used infra-red tags, voice features, and body motion to predict when people would exchange business cards or ask for project-related information.

In our system the basic element of social context is the identity of people in the users' immediate presence. We use bluetooth radios in mobile phones to measure proximity within a 10 meter range (e.g. same office or room). Bluetooth does not distinguish face-to-face interactions but is a good compromise for long-term user convenience and integrated functionality.

3.2 Social Signaling in Voice

The social and developmental psychology literature suggests that people sub-consciously use social signals to communicate [2,15,22]. Through this independent communication channel, a consistent *attitude* signal is expressed via various modalities; voice, facial expression and gesture. These social signals are intentional attitudes like: *I'm interested, friendly or attracted to you.*

To quantify these social signals Pentland [25] developed *texture* measures for four types of social signaling, which were designated activity level, engagement, emphasis¹, and mirroring. Audio features measuring these signals have been used to successfully predict outcomes of salary negotiations [5], effectiveness of business pitches [26], as early indicators of clinical depression [27], and to conversational interest and dating outcomes [21] (described here in section 3.2 and 3.3) with accuracy comparable to that of human experts in analogous situations.

Activity Measure

Calculation of the activity measure begins by using a two-level HMM to segment the speech stream of each person into voiced and non-voiced segments, and then group the voiced segments into speaking vs. non-speaking [3]. Conversational activity level is then measured by the z-scored percentage of speaking time.

Engagement Measure

Engagement is measured by the z-scored influence each person has on the other's turn-taking. When two people are interacting, their individual turn-taking dynamics influences each other and can be modeled as a Markov process [19]. By quantifying the influence each participant has on the other we obtain a measure of their engagement...popularly speaking, were they driving the conversation? To measure these influences we model their individual turn-taking by a Hidden Markov Model (HMM) and measure the coupling of these two dynamic systems to estimate the influence each has on the others' turn-taking dynamics [7].

¹ In previous publications we have referred to this feature as stress rather than emphasis. This has proven to cause confusion with the features used in the stress measurement literature, and so we refer to emphasis instead.

Emphasis Measure

For audio data emphasis is measured by the variation in pitch and amplitude. For each voiced segment we extract the mean energy, frequency of the fundamental format, and the spectral entropy. Averaging over longer time periods provides estimates of the mean-scaled standard deviation of the formant frequency and spectral entropy. The z-scored sum of these standard deviations is taken as a measure speaker emphasis; such emphasis can be either purposeful (e.g., prosodic emphasis) or unintentional (e.g., due to physiological stress caused by discomfort).

Mirroring Measure

Mimicry, in which the gestures and prosody of one participant is 'mirrored' by the other, is considered a signal of empathy and has been shown to positively influence the outcome of a negotiation and other interpersonal interactions [5,6]. Mimicry is a complex behavior and therefore difficult to robustly measure using computer methods.

However in our experiments we found that the distribution of utterance length was often bimodal. Sentences and sentence fragments typically occurred at several-second and longer time scales. At time scales less than one second there are short interjections (e.g., "uh-huh"), but also back-and-forth exchanges typically consisting of single words (e.g., "OK?", "OK!", "done?", "done!"). Because these exchanges, which we refer to as *mirroring*, are an important part of mimicry and are easily measured, we have taken the z-scored frequency of these short utterance exchanges as a proxy measure of mimicry.

3.3 Measuring Interest in Voice

In order to understand which vocal signaling measures (if any) could predict *interest* in conversation we conducted a study with 20 participants, 10 males and 10 females, paired up with a same-sex conversation partner. Each pair participated in 10 successive short (3 minute) conversations on random topics; the total duration of each session was approximately 30 minutes. After each short conversation, participants provided training labels; their interest rating on a scale of 1-10 (10 being highest). The data was analyzed with a stepwise linear regression that compared each individual's interest ratings with both participant's four social signaling measures (for a total of eight social signaling measures). At each step of the regression the most significant measure with $p < 0.05$ was added to the regression; measures with $p > 0.10$ were dropped. The

stepwise regression continued until there were no more measures with $p < 0.05$.

For both men and women, the speech signals extracted from the audio of both speakers were found to be significantly correlated with the self-reported interest labels. For men, all features are correlated with the self-reported labels ($r=0.66$, $p=0.02$). The emphasis and activity measures for both men ($r=0.59$, $p=0.01$) and activity features alone ($r=0.52$, $p=0.007$) are also significantly correlated with the reported labels. The influence feature and back-and-forth features alone did not have statistically significant correlations with interest labels.

Similarly for women, the speech features are correlated with the self-reported labels as well ($r=0.67$, $p=0.006$). The stress and activity measures for both women also have a significant correlation with the labels ($r=0.6$, $p=0.004$). Only the speaker's emphasis and activity measures ($r=0.48$, $p=0.006$) also played an important role.

The overall distribution of interest ratings can be split into a two class model – those labeled as high interest (all ratings 8 and above) and low interest (all rating less than 8). If the activity and emphasis features are used in a linear classifier the cross-validated classification accuracy for judging between high and low interest conversations is 64%; using an SVM with linear kernel the accuracy rises to 75%.

3.4 Measuring Sexual Attraction in Voice

To determine which signaling measures correlated with sexual *attraction*, we analyzed 60 five-minute speed-dating conversations. Data was collected from several events held in 2004, and participants were singles from the graduate student community between the ages of 21 to 45. In addition to the standard 'romantically attracted' question, participants were also two other yes/no options: would they like to stay in touch just as friends, and would they like to stay in touch for a business relationship. If they both replied 'yes' to the same questions, contact information was shared between the two parties. These 'stay in touch' questions allowed us to explore whether romantic attraction could be differentiated from other factors. More details about the experiment are available in [21].

The four social signaling measures for both male and female were compared by linear regression to the question responses, and in each case the resulting predictor accounts for more than 1/3rd of the variance.

For the females responses, for instance, the correlation with the 'attracted' responses were $r=0.66$, $p=0.01$, for the 'friendship' responses $r=0.63$, $p=0.01$, and for the 'business' responses $r=0.7$, $p=0.01$. Corresponding values for the male responses were $r=0.59$, $r=0.62$, and $r=0.57$, each with $p=0.01$. The emphasis measure showed correlation with both participants saying 'yes' or both saying 'no' for the 'attraction' ($r=0.6, p=0.01$) and 'friendship' ($r=0.58, p=0.01$) questions.

An interesting observation was that for the 'attracted' question female features alone showed far more correlation with both male ($r=0.5, p=0.02$) and female ($r=0.48$, $p=0.03$) responses than male features (no significant correlation). In other words, female social signaling is more important in determining a couples 'attracted' response than male signaling. The most predictive individual feature was the female activity measure.

| Question | SVM Accuracy (Linear kernel) |
|--|------------------------------|
| Are you interested in this person? (asked of females) | 0.71 |
| Would you like to be friends with this person (asked of females) | 0.76 |
| Are you interested in this person? (asked of males) | 0.64 |
| Would you like to be friends with this person (asked of males) | 0.79 |

Table 2. SVM Classification Accuracy

4. Mobile Phones as Social Prosthesis

The VibeFone application augments traditional means of gathering social interaction data (surveys or ethnographic studies) and speech data. The mobile phone platform is highly conducive to collecting long-term continuous data and sampling the user for training labels.

Example 1: Characterizing Everyday Social Interactions

The following data was collected from three friends who spent half a day sightseeing in downtown Boston. The VibeFone application asked the users for an interest rating ten minutes after a new buddy was detected in his or her proximity. In this particular session, a total of 27 interest labels were collected

which have been used for model training. The training labels ranged from "1 – depressed" to "5 – excited".

The activity and emphasis features extracted from each person's voice were significantly correlated with their interest labels. If we consider the last five minutes of the user's interaction before the training label was assigned, features from the interest experiment in section 3.3 (emphasis and activity) are highly correlated with the self-reported labels ($r = 0.52$, $p = 0.05$).

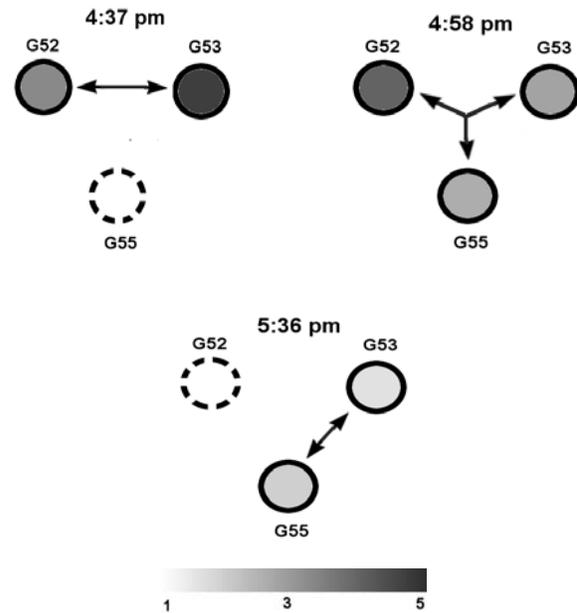


Figure 2. Predicted interest for 3 friends in downtown Boston for an hour. The circles represent people, fill-in color represents our predictions of their interest level (1-5) and a connecting line represents (bluetooth) co-location.

Fig 2 above summarizes the interactions for the same three friends over the course of an hour at a popular tourist site. Each circle represents a person, and participants in proximity (detected using Bluetooth) are represented by a connecting line. The shade of gray used to fill the circle represents the interest or activation level generated from our interest measure. At 4:37 pm, G52 and G53 are co-located and seem interested in their sightseeing activities. Twenty minutes later, when G55 joins them, G52 is more "activated" and a decreasing trend is observed forty minutes later after G52 leaves the group. Our automated measures for this session had a mean

difference of 0.7 (<15% error) from the reported labels.

This illustrates that by using measures of social signaling in voice with traditional social network analysis, it is possible to build a more complex understanding of social encounters. With reasonably accurate real-time predictive modeling on a mobile phone, it is possible to record relevant interactions, take pictures, or share this interaction metadata across friends. SMS is extremely popular with youth in Asia and Europe, and in our system the user's 'interest or activation' information can be shared with absent buddies, with appropriate privacy restrictions.

Example 2: Personal Relationship Management at the Workplace

In a corporate environment, the value of improving personal networks and interactions is evident. Organizational hierarchy does not map to internal social and collaboration networks, and even less with the *quality* or *strength* of the network links. On the other hand, these very networks are the arteries of communication and productivity. The development of a network of interpersonal relationships has been found to be critical to a general manager's ability to execute his or her agenda, and the quality of these relationships has been found to be a key determinant of managerial effectiveness [11]. Similarly, it has been shown that subordinates reporting good relationships with superiors are better performers assume more responsibilities and have greater contributions to their units than those reporting poor relationships.

Researchers have used several approaches to map internal social networks – surveys [9], mining email headers [1], or measuring co-location using both IR and Bluetooth radios. We propose that the traditional features, frequency and length of interactions, could be augmented with data about the *quality* of workplace interactions. The VibeFone application can be used to build a sophisticated understanding of interactions within workgroups and even manage personal relationships, by analyzing workplace interactions on the order of a few minutes. A cumulative history of the *quality* of interactions with a particular individual is a proxy for the relationship shared with that person. Our current workplace interaction miner application is designed for long-term use, and captures speech data during phone calls and Bluetooth proximity information.

Example 3: A Real-Time Interest Reminder

The Jerk-O-Meter is a special mode in the VibeFone application. It is a real-time conversational interest meter intended to detect if a user is distracted during phone calls and display an appropriate reminder. The Jerk-O-Meter is based on the results of the conversational interest study in section 3.3. Stress and activity measures extracted from the user's voice are used to predict how *interested* he/she is in the conversation.

In order to use the results from the interest study in this application, some simplifying assumptions were made. For privacy reasons, only the users' features are extracted. The engagement and mirroring features require synchronized audio from both people, and hence are not calculated in the real-time application. For voicing / speaking detection, HMM parameters are derived from off-line MATLAB analysis of training data rather than computed on the fly. The interest rating is predicted every 30 seconds.

When invoked, the application displays a message that corresponds to the predicted value which can be:

“Stop being a Jerk!”
“You could do better”
“Now we're getting somewhere”
“Wow you're a smooth talker”

The message or a warning bell can also be played over the ear-piece. The user controls when the Jerk-O-Meter application mode is invoked.

Example 4: A Personal Trainer for Dating Encounters



Figure 3. Wingman3G mode on an A780 mobile phone. The three progress bars are for speaking time, voice rate and vocal emphasis. The feedback message is at the bottom.

conversation in a dating situation. The Wingman3G mode in our application is directly based on results of the attraction study. The user can activate this mode when needed,

and can switch between the male and female models. The female model uses coefficients derived from the study. For the male model, the application encourages specific behavior that improves chances of a positive outcome. For example, low-mid speaking time with short back-and-forth interactions increased the probability of a positive response in our study.

Wingman3G is designed as a personal training tool, and the feedback messages include "Maybe you could speak a little slower?" (if the voiced rate is very high) or "You're getting there, maybe you could relax a little? " (when the emphasis measure exceeds the optimal range).

5. Related Work

Jabberwocky, HummingBird, and Proem [17,20 and 24] are a few of many mobile social networking and collaboration projects that connect users to their social networks or facilitate introductions between strangers with similar interests or profiles. The Reality Mining project uses social proximity to identify social links and relationships and facilitate introductions [10]. The Uberbadge project is the most closely related to our work, since speech amplitude features and face-to-face proximity information is used to identify relevant bookmarks [14].

In the affective computing literature there are several projects which use non-linguistic speech features (energy, pitch, etc.) over smaller time windows (few seconds as compared to minutes in our case) to quantize conversational interest and engagement [13,30]. The link between our social signals of interest and the arousal axis of affect, described in their work, needs further exploration.

6. Future Work and Discussion

In this paper, we describe how mobile phones can automatically mine meaning from social and workplace interactions in real-time. For example, a system that can predict when users are *interested* in a particular experience could suggest audio or video content.

We are currently building such a system using VibeFones for tourists, in collaboration with the Scottish government. We are currently collecting tourist interaction data using our platform, and analyzing it to design more accurate models of the tourists' experience.

We have also introduced the idea of using mobile devices as social coaches or personal trainers, for phone conversations and dating encounters. Following our initial small-scale successes we are now addressing how to improve the accuracy of our predictions, and the questions of how appropriate and useful is this type of feedback?

7. References

- [1] L. Adamar and B. Huberman, "Information Dynamics in a Networked World", *Complex Networks, Lecture Notes in Physics*, 2003.
- [2] Ambady, N. and Rosenthal R., "Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta-analysis", *Psychological Bulletin*, 111(2), 256-274, 1992
- [3] Basu, B., "Conversational Scene Analysis", PhD Thesis in the Dept. of Electrical Engineering and Computer Science, MIT, 2002.
- [4] M. Blum, A. Pentland, "InSensed: Wearable Collection of Multimedia based on Interest", *Master's Thesis* ETH Zurich.
- [5] R. Caneel, "Social Signaling in Decision Making", Master's Thesis in Media Arts and Sciences, MIT 2005.
- [6] Chartrand, T., and Bargh, J., "The Chameleon Effect: The Perception-Behavior Link and Social Interaction", *J. Personality and Social Psychology*, Vo. 76, No. 6, 893-910
- [7] Choudhury, T., and Pentland, A., "NAASCOS, June 27-29, Pittsburgh, PA. PDF available at <http://hd.media.mit.edu>
- [8] T. Coates, "My working definition of Social Software", http://www.plasticbag.org/archives/2003/05/my_working_definition_of_social_software.shtml
- [9] A. Davis, B. Gardner, and M. Gardner, "Deep South", *University of Chicago Press*, Chicago, 1941.
- [10] N. Eagle, "Machine Perception and Learning of Complex Social Systems", *Ph.D. Thesis*, Program in Media Arts and Sciences, MIT, June 2005.
- [11] J. Gabarro, R. Kraut and C Egidio, "The Development of Working Relationships", *Intellectual Teamwork: Social and Technological Foundations of Co-operative work*, Eds. Lawrence Erlbaum Associates, NJ 1990.
- [12] Gartner Mobile Industry Report 2005
- [13] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting Group Interest-level in Meetings", *IEEE Int.*

Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2005.

[14] J. Gips and A. Pentland, "Mapping Human Networks", *IEEE Int'l Conference on Pervasive Computing and Communications*, Pisa, Italy, March 2006.

[15] M. Gladwell, "Blink", *Little Brown and Co.* NY 2004

[16] GroupMedia webpage, Human Dynamics Group, MIT Media Lab, <http://groupmedia.media.mit.edu>.

[17] Holmquist, L.E., Falk, J. and Wigstrm, J. "Supporting Group Awareness with Inter-Personal Awareness Devices", *Journal of Personal Technologies*, Vol. 3(1-2), Springer Verlag, 1999. pp.1321

[18] J. Healey and R. W. Picard, "StartleCam: A Cybernetic Wearable Camera", *Proc. of the International Symposium on Wearable Computers*, Pittsburgh, PA, October 1998.

[19] Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., & Jasnow, M. (2001). Rhythms of dialogue in early infancy. *Monographs of the Society for Research in Child Development*, 66(2), No. 264.

[20] Kortuem, G., Segall, Z., Cowan Thompson, "Close Encounters: Supporting Mobile Collaboration through Interchange of User Profiles", *1st International Symposium on Handheld and Ubiquitous Computing*, September 1999, Karlsruhe, Germany

[21] A. Madan, "Thin Slices of Interest", *Master's Thesis in Media Arts and Sciences*, MIT 2005.

[22] Nass, C., and Brave, S., "Voice Activated: How People Are Wired for Speech and How Computers Will Speak with Us", *MIT Press*, 2004

[23] Oudeyer, P-Y., "The production and recognition of emotions in speech: features and algorithms", *International Journal of Human Computer Interaction*, 59(1-2), pp 157-183.

[24] E. Paulos, E. Goodman, "The familiar stranger: anxiety, comfort and play in public places", *CHI*, 2004.

[25] A. Pentland, "Social Dynamics: Signals and Behavior", *IEEE Int'l Conf. Developmental Learning (ICDL)*, San Diego, CA, Oct 2004

[26] W. Stollzman, A. Pentland, "Elevator Rator", draft in submission.
<http://groupmedia.media.mit.edu/elev.php>

[27] M. Sung, C. Marci and A. Pentland, "Objective Physiological and Behavioral Measures for Tracking Depression", *excerpt from PhD Thesis*.

http://cgi.media.mit.edu/vismod/tr_pagemaker.cgi

[28] Wearable Computing Lab QBIC webpage, ETH Zurich
<http://www.wearable.ethz.ch/qbic.0.html>

[29] Wearable Computing Group platform webpage, MIT Media Laboratory
<http://www.media.mit.edu/wearables/platforms.html>

[30] C. Yu, P. Aoki, A. Woodruff, "Detecting User Engagement in Everyday Conversations", *ICSLP*, 2004