# Thin Slices of Interest

by

## Anmol P. Madan

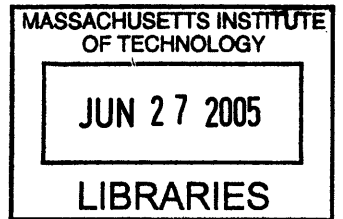B.E., Maharashtra Institute of Technology, Pune University (2003)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

Author_____
Program in Media Arts and Sciences
May 5th, 2005

Certified by_____
Dr. Alex Pentland
Toshiba Professor of Media Arts and Sciences
MIT Media Laboratory
Thesis Supervisor

Accepted by_____
Andrew B. Lippman
Chair, Department Committee on Graduate Students
Program in Media Arts and Sciences

# Thin Slices of Interest

by

Anmol P. Madan

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on May 5th, 2005, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

## Abstract

In this thesis we describe an automatic *human interest detector* that uses speech, physiology, body movement, location and proximity information. The speech features, consisting of activity, stress, empathy and engagement measures are used in three large experimental evaluations; measuring interest in short conversations, attraction in speed dating, and understanding the interactions within a focus group, all within a few minutes.

In the conversational interest experiment, the speech features predict about 45% of the variance in self-reported interest ratings for 20 male and female participants. Stress and activity measures play the most important role, and a simple activity-based classifier predicts low or high interest with 74% accuracy (for men).

In the speed-dating study, we use the speech features measured from five minutes of conversation to predict attraction between people. The features predict 40% of the variance in outcomes for attraction, friendship and business relationships. Speech features are used in an SVM classifier that is 75%-80% accurate in predicting outcomes based on speaking style.

In the context of measuring consumer interest in focus groups, the speech features help to identify a pattern of behavior where subjects changed their opinions after discussion.

Finally, we propose a prototype wearable 'interest meter' and various application scenarios. We portray a world where cell phones can automatically measure interest and engagement, and share this information between families and workgroups.

Thesis Supervisor: Dr. Alex Pentland
Title: Toshiba Professor of Media Arts and Sciences, MIT Media Laboratory

# Thin Slices of Interest

by

Anmol P. Madan

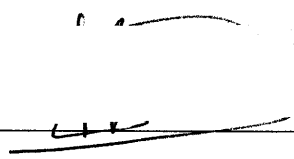The following people served as readers for this thesis:

Thesis Reader_____

Dr. Alex P. Pentland
Toshiba Professor of Media Arts and Sciences
MIT Media Laboratory

Thesis Reader_____

Dr. Carl Marci
Director of Social Neuroscience, Department of Psychiatry
Massachusetts General Hospital and Harvard Medical School

Thesis Reader_____

Dr. Pattie Maes
Professor of Media Arts and Sciences
MIT Media Laboratory

## Acknowledgements

"One has the ability and privilege to do one's respective duty, but has no control over the results. The fruits of work should not be the motive, for they shall follow one's actions"
Bhagvad Gita, (2.47)

# Contents

# List of Figures

# Chapter 1

# Introduction

Non-linguistic social signals (e.g., 'tone of voice', body language) are often as important as linguistic or affective content in predicting behavioral outcomes and can be regarded as a form of social signaling. There is reason to believe that we can make good estimates of peoples' interest and engagement level from sensor data without requiring explicit interaction.

In Malcolm Gladwell's popular book, Blink, he describes the surprising power of "thin slicing," defined as "the ability of our unconscious to find patterns in situations and people based on very narrow 'slices' of experience" [30] [p. 23]. Gladwell writes, "there are lots of situations where careful attention to the details of a very thin slice, even for no more than a second or two, can tell us an awful lot" [30][p. 47].

Gladwell's observations reflect decades of research in social psychology, and the term "thin slice" comes from a frequently cited study by Nalani Ambady and Robert Rosenthal [1]. They have shown that observers can accurately classify human attitudes (such as interest) with from non-verbal behavior using observations as short as six seconds. The accuracy of such 'thin slice classifications are typically around 70%, corresponding to a correlation between observer prediction and measured response of about r=0.40. In wired for Speech, Nass and Brave [44] take this further and examine how people are "voice-activated"; not only is it possible predict social outcomes from verbal and behavioral signals, but people

also respond to voice technologies as they respond to other people and behave as they would in any social situation.

Throughout history, people have attempted to accurately measure interest and engagement. A quick Google search reveals about a million websites that discuss displays of interest or romantic attraction, many cite voice and body signals. The science of measuring interest has evolved over the centuries from heresy to surveys and feedback forms and human observers to A.C. Nielsens 'people meters'.

In this thesis, we describe an automatic *human interest detector* that uses speech, physiology, location, proximity and other features. This detector uses advanced machine perception techniques and can run on a variety of platforms - computers, cell phones and proprietary hardware [27]. The interest detector can run locally on the chosen hardware platform, or can evaluate data sent over a network connection to measure interest. The ability to understand these social signals in an automated fashion allows us to make better predictions about the outcomes of interactions - for example speed-dating, conversations, movies, maybe even interests in products. Our reasoning is that people actively construct these social signals with a communicative intent, and we can train computers to measure them. We find that this approach allows us to make surprisingly accurate predictions about the significant outcomes of social situations.

There are many instances in day-to-day life where being able to measure human interest automatically can have significant impact. Movie and television audiences, product marketing, qualitative market research, focus groups, amusement parks, music and entertainment, sales training and many other commercial industries spend roughly $16 billion every year trying to understand consumer feedback and interest. Current methods of qualitative analysis lack implicitness and require some element of human involvement to generate this data - filling out a feedback form or observing a customer. One can imagine that a future where pervasive social intelligence in cell phones and computers may change the way we interact and handle information.

16

## 1.1 Defining Interest

'Interest' and 'engagement' are commonplace words with many contextual definitions. Merriam-Webster's defines interest as "a feeling that accompanies or causes special attention to an object or class of objects; something that arouses such attention". An interesting perspective is Csikszentmihalyi's flow theory [17], described in his own words as "being completely involved in an activity for its own sake. The ego falls away. Time flies. Every action, movement, and thought follows inevitably from the previous one, like playing jazz. Your whole being is involved, and you're using your skills to the utmost."

On a more physiological level, engagement can be related to the arousal axis of the arousal/valence model [41]. Valence is described as the subjective feeling of pleasantness or unpleasantness, and arousal as the subjective feeling of being activated or inactivated. There are known sympathetic and parasympathetic physiological effects associated with arousal, and have been extensively measured using facial expressions, facial EMG, heart rate, and skin conductance data [56]. The second chapter on interest features describes the physiological effects in more details.

In contrast, people actively use nonverbal and behavioral social signals to communicate social variables such as interest, status, determination, or cooperation. These signals mediate social communication and arise from the interaction of two or more people rather than being a property of a single speaker. The social signaling that we measure seems to be a sort of 'vocal body language' that operates relatively independently of linguistic or affective communication channels, and is strongly predictive of the behavioral outcome of in social interactions.

Replacing human perception by machine perception has both advantages and disadvantages. It is hard to give computers the sub-conscious perceptual ability that we as humans possess, when we don't understand it completely ourselves. On the other hand, computers are far more capable of evaluating low-level physiological effects like skin conductivity (SC) or vocal stress. The right approach from our perspective, and the one we have tried to follow in this work, is to build computational and machine perception models that are based on current

17

understanding of the sciences of physiology, psychology and sociology.

## 1.2  Contributions of This Thesis

In this thesis we attempt to build a *human interest detector* using speech features, physiological features, body movement, location and proximity. We explain these features in detail and justify as to why they should be considered. The speech features, consisting of activity, stress, empathy and engagement measures are used in three large experimental evaluations measuring interest and engagement in conversation, attraction in speed-dating, and understanding interactions in focus groups.

In the conversational interest experiment, the speech features predict about 45% of the variance in self-reported interest ratings for 20 male and female participants. Stress and activity measures play the most important role, and a simple activity-based classifier predicts low or high interest with 74% accuracy (for men). This study provides evidence that our chosen features capture some element of the non-verbal body language and social signaling that is indicative of human interest, within a few minutes.

In the speed-dating study, we use the speech features measured from five minutes of conversation to predict attraction between people. The features play an important role in predicting the outcome of these dyadic interactions. They predict 40% of the variance in outcomes for attraction, friendship and business relationships. Speech features are used in an SVM classifier that is 75%-80% accurate in predicting outcomes based on speaking style. We use these results to build a prototype real-time dating meter on a PDA.

We then evaluate how speech features may offer insight when measuring consumer interest in focus group study. Although our experiment and analysis are exploratory, the speech features help to identify a pattern of behavior where subjects changed their opinions after discussion.

Finally, we propose a prototype wearable interest meter and various application scenarios. We portray a world where cell phones can automatically measure interest and engagement,

and share this information between families and workgroups. Such a system could automatically patch in pre-approved team-members into the conversation, or selectively record and annotate discussions. We have also received commercial interest in using it as a wearable tool for gather qualitative consumer research data.

## 1.3   Structure

The second chapter explains the various features (speech, physiology, body movement and location/proximity) in more detail and justifies as to why there are important to consider. We also show results from two pilot experiments. The first experiment shows how concordance in skin conductivity for movie or television commercial audiences may identify areas of excitement and provide creative feedback. The second shows how head nodding may be an indicator of changes in interest in a group discussion.

The third chapter describes the GroupMedia hardware and software infrastructure, which is used for data collection and real-time feedback in the following studies. The GroupMedia system is a wireless, distributed wearable system based on the Sharp Zaurus SL 5500/6000 that is capable of real-time data collection and inference for speech features, physiology, body motion and location and proximity information. The system is designed to support group interactions and features advanced networking and communications abilities

Chapters four, five and six describe three larger studies that help us understand the role of speech features in interest, engagement and attraction. We to predict answers to questions like "Do you find this conversation interesting?, "Does she like me? and even "Are you going to buy that product?. We show that our chosen speech features and social signaling measures capture the underlying behavior and can predict outcomes with accuracies of 70-80% within minutes.

Finally in chapter seven, we propose a wearable, distributed interest meter termed InterestNetworks. We describe several potential applications in social networking, collaborative filtering, selective recording and annotation and qualitative market research.

# Chapter 2

# Interest Features

This thesis uses multi-modal features, primarily from signaling theory and physiology that can be measured non-invasively and unobtrusively on a mobile device. Our motivation was to see this system in broad application, by using features that offer meaningful insight about the person's behavior, yet could be implemented on next-generation mobile devices without taking away from the user experience. We describe four sets of features that satisfy these criteria; speaking style, autonomous physiological reactions, body movements and location and proximity data.

## 2.1 Speech Features

### 2.1.1 Speech Production and Autonomous Arousal

It is important to provide some background about the process of speech production before diving into the relevant speech features. According to O'Shaughnessy [46], audible speech can be divided into sound segments, which may share common acoustical and articulator properties. For each sound, there is a positioning for each of the vocal tract articulators: vocal folds (or cords), tongue, lips, teeth, velum, and jaw. Sounds are divided into vowels

(unrestricted airflow in the vocal tract) and consonants (restricted airflow at some point and have weaker intensity).
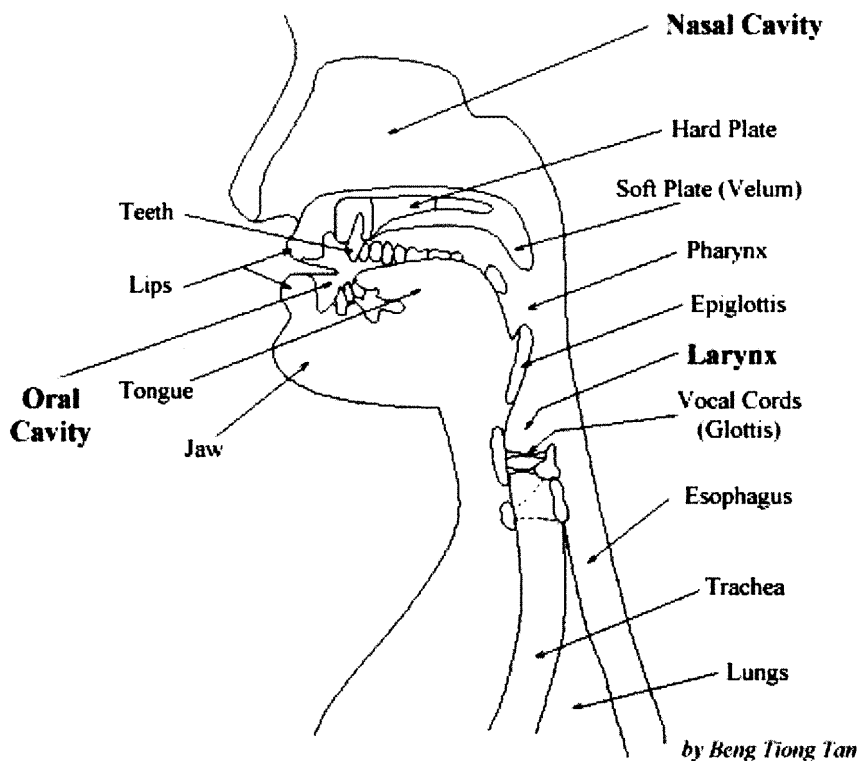


Figure 2-1: Speech Production Organs

Speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter; the source may be periodic, resulting in *voiced* speech or noisy and aperiodic, resulting in *unvoiced* speech. The voicing source occurs in the larynx, at the base of the vocal tract, where airflow can be interrupted periodically by vibrating vocal folds. The pulses of air are produced by opening and closing of the folds, and the pulse volume (vs. time) represents half a sine wave. For both voiced and unvoiced excitation, the vocal tract acts a filter and amplifies certain frequencies while attenuating others.

All the speech features used in this work are based on voiced speech. As a periodic signal, voiced speech has spectra consisting of *harmonics* of the *fundamental frequency F0* of the vocal fold vibration; this frequency is the physical aspect of speech corresponding to the

perceived pitch. Since the vocal tract changes shape almost continually, voiced sounds are instead only locally quasi-periodic (or almost periodic). Unvoiced sounds have no pitch and include lip bursts (like /p) or fricatives (like /s/sh).

Current literature in affective computing (Picard [26, 50], Oudeyer [47], Breazeal[6]) also links certain emotional states with sympathetic or parasympathetic physiological states, which in turn have quite mechanical and thus predictable effects on speech, especially on pitch, (fundamental frequency F0) timing and voice quality. For instance, when the sympathetic nervous system is aroused (anger, fear or joy), heart rate and blood pressure increase, the mouth becomes dry and there are occasional muscle tremors. *Speech becomes loud, fast, and enunciated with strong high frequency energy.* When the parasympathetic nervous system is aroused (bored, sad emotional states), heart rate and blood pressure decrease, salivation increases and the resulting *speech is slow, low-pitched with little high-frequency energy.* With reference to the arousal/valence graph, physiological and speech effects (prosody, rate, pitch, loudness) that map onto the arousal axis are generally better understood than the ones that map onto the valence axis.

Current research in affective computing uses speech to identify particular emotional states, with diverse approaches being applied. For example, Dallaet et. al. [18] implemented a method based on majority voting of subspace specialists to classify acted spoken utterances into four emotion types; Batliner et. al [8] provided a comparative study of recognizing two emotion types, "neutral" and "anger", expressed by actors and subjects ; Schroder et. al. [54] analyzed the correlation between various acoustic features and three abstract emotional characteristics in a collection of broadcast media recordings. A more detailed review of automatic emotion recognition studies in speech can be found in Cowie et. al. [15]

In contrast, we propose that minute-long averages of audio features often used to measure affect (e.g., variation in pitch, intensity, etc) taken together with conversational interaction features (turn-taking, interrupting, making sounds that indicate agreement like 'uh-huh') are more closely related to social signaling theory rather than to an individual's affect. Animals communicate and negotiate their position within a social hierarchy in many ways,

23

including dominance displays, relative positioning, and access to resources. Humans add to that repertoire a wide variety of cultural mechanisms such as clothing, seating arrangements, and name-dropping (Dunbar [19]).Pentland [49] constructed measures for four types of vocal social signaling, designated activity level, engagement, stress, and mirroring. These four measures were extrapolated from a broad reading of the voice analysis and social science literature, and some of the results included in this thesis establish their general validity.

## 2.1.2 Activity

Calculation of the activity measure begins by using Basu's two-level HMM [7] to segment the speech stream of each person into speaking and not-speaking segments. The first level of the cross-linked HMM uses noisy autocorrelation, number of autocorrelation peaks and the spectral entropy to calculate the voiced and non-voiced segments. For a voiced segment, the (normalized) autocorrelation peaks are fewer and larger in amplitude for the periodic component, and the spectral entropy is lower than that of an unvoiced segment. The voiced/unvoiced states are then used to compute the speaking/not-speaking segments using a linked HMM. The performance of a linked HMM has been shown to be better than a simple HMM as knowledge about the speaking states can reinforce estimation of the voicing states in noisy conditions.

The speaking/not-speaking states are then converted to conversation activity measures as z-scored percentage of speaking time and the frequency of voiced segments (voicing rate).

## 2.1.3 Stress

Stress is measured by the variation in prosodic emphasis. Zhou et.al. [61] examined the relationship between five linear features (duration, pitch, intensity, glottal source, vocal tract spectrum) and non-linear features (derived from the Teager Energy Operator) and reported that voiced pitch was the best individual linear predictor of stress. We use three linear features to measure stress, which have been derived from a broader reading of the literature, and capture the essence of Zhou's individual linear features. In [62] it has been

24

shown that a similar combination of (functions of) intensity, pitch and duration have the lowest error rate for stress classification, in comparison with Teager and MFCC based features.

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

### 2.1.4 Engagement

Jaffe et.al.[33] show that when two people are interacting, their individual turn-taking dynamics influence each other and can be modeled as a Markov process. By quantifying the influence each participant has on the other we obtain a measure of their engagement...popularly speaking, were they driving the conversation? Choudhary and Pentland [13], motivated by a need for a minimal parameterization, developed generalized Coupled Hidden Markov Models (CHMMs) to describe interactions between two people, where the interaction parameters are limited to the inner products of the individual Markov chains. When two people are interacting, their individual turn-taking dynamics influence each other and can be modeled as a Markov process, in terms of the influence each person has on the other. This influence parameter expresses then how strong the overall state for an actor A is depending on the state of actor B.

Engagement is measured by the z-scored influence each person has on the other's turn-taking by using this implementation [13]. Our method is similar to the classic method of Jaffe et al. [33], but with a simpler parameterization that permits the direction of influence to be calculated and permits analysis of conversations involving many participants. Our engagement/influence measure was shown to have an extremely high correlation with one measure of the social novelty (and thus presumably the interestingness) of the information

being presented [13]. Pentland et. al.[48] also have shown that the influence parameter plays a significant role in predicting the outcome of salary negotiations between middle managers and vice presidents.

There is some literature to indicate that this form of turn taking is related to vocal engagement in conversation. Woodruff [2] describes how people change thier conversational style to one with increased turn-taking (measured more accurately by the influence model) if they are engaged in a conversation. Yu et. al [59] use a more advanced model based on coupled-HMMs to measure turn-taking in vocal engagement.

### 2.1.5 Mirroring

Chartrand and Bargh[12] show in great detail how people unconsciously 'mirror' the mannerisms and expressions of their interaction partners, and unconsciously and passively match others in their environment. They demonstrate that such mimicry not only facilitates smoother interactions and increases likeability, but it can also signal empathy. Kimbara and Parrill [36] state that mirroring in speech may be correlated with the perception of *social resonance*. Our measurement of non-linguistic empathy has been shown to positively influence the outcome of a salary negotiation [48].

In our experiments the distribution of utterance length is 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?', 'yup.'). The z-scored frequency of these short utterance exchanges is taken as a measure of mirroring. In our data these short utterance exchanges were also periods of tension release.

## 2.2 Physiological Features

### 2.2.1 The Autonomous Nervous System

The autonomic nervous system consists of sensory neurons and motor neurons that run between the central nervous system and various internal organs such as the heart and lungs. It is responsible for monitoring conditions in the environment and bringing about appropriate changes in the body. The autonomic nervous system has two subdivisions, the sympathetic nervous system and the parasympathetic nervous system.
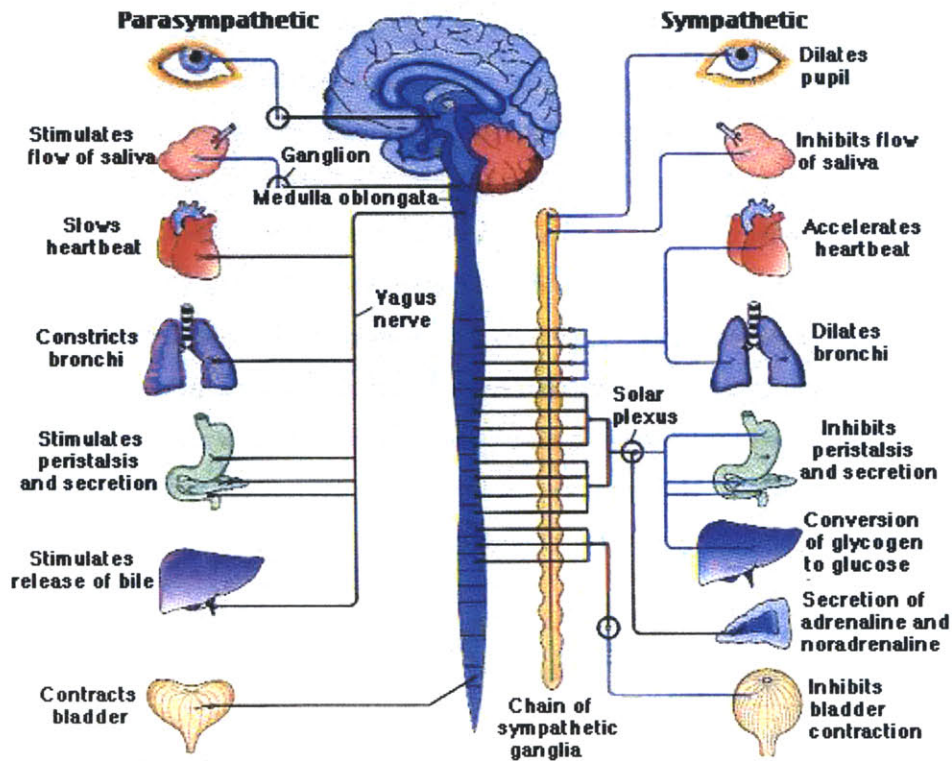


Figure 2-2: Autonomic Nervous System

Stimulation of the sympathetic branch of the autonomic nervous system prepares the body for emergencies or 'fight or flight'. The observed effects include an increase in heart rate, increase in blood pressure, dilation of pupils, and diversion of sugar rich blood from the skin and viscera (stomach) to the brain and heart. On the other hand, parasympathetic

stimulation causes slowing down of the heartbeat, lowering of blood pressure, constriction of the pupils and increased blood flow to the viscera, amongst other effects. In times of danger, the sympathetic system prepares the body for violent activity. The parasympathetic system reverses these changes when the danger is over. Although the autonomic nervous system is considered to be involuntary, this is not entirely true. A certain amount of conscious control can be exerted over it as has long been demonstrated by practitioners of Yoga and Zen Buddhism.

Skin conductivity (SC) has been used as a measurable parameter of a person's internal 'state' [50]. In terms of physiology, SC reflects sweat gland activity and changes in the sympathetic nervous system and measurement variables. It is measured as change in the relative conductance of a small electrical current between the electrodes placed at the fingertips (or the palm). The activity of the sweat glands in response to sympathetic nervous stimulation results in an increase in the level of conductance. According to Fuller [25], there is a relationship between sympathetic activity and emotional and cognitive arousal, although one cannot identify the specific emotion being elicited. Fear, anger, startle response, orienting response and sexual feelings are all among the emotions that may produce similar skin conductivity responses. Similarly, instantaneous heart-rate and EKG have also been successfully used to measure sympathetic nervous system arousal [56]. The effects of the autonomous arousal on heart rate are more complex, but heart-rate variability can be used to tease out influences on skin conductivity.

## 2.3  Skin Conductivity(SC)

As mentioned previously, SC is an indicator of sympathetic nervous stimulation. Studies in psychophysiology have used skin conductivity (or electrodermal activity) as a physiological index for multiple individuals exposed to the same cognitive or emotional events. They have gone to the extent of looking at correlations in skin conductivity and using it as an objective measure of relationships between married couples. Marci [43] has shown how skin conductivity correlates are indicative of laughter and counseling empathy between patients

and therapists during psychotherapy.

Traditionally, skin conductivity alone has not been very accurate because of its noisy nature. A sympathetic response may cause a skin conductivity rise in an individual due to many different causes, like opening eyes, standing up, drinking coffee, etc. physiological arousal which may not be related with cognitive or emotional arousal. Hence, in the following studies, we measure skin conductivity as a group index, by seeking similar SC responses within the dynamic set of people. We propose that SC trends experienced by the entire group will have higher correlation with cognitive or emotional events that the group has experienced. In addition, SC trends can also be weighted with speech features explained in the previous section to increase accuracy.

Skin conductivity response signals essentially contain two types of information. Instantaneous spikes have been correlated with startle and strong emotional physiological responses [50, 31]. On the other hand, longer duration SC rise and fall trends have been related to relatively longer term changes in empathy [43, 50] between individuals.

The raw skin conductivity signal is pre-filtered and smoothed using a median average to eliminate high-frequency noise in measuring equipment. In order to measure the extent to which SC responses in several people are correlated to cognitive or emotional events that they all experienced, we calculate the cross-correlation co-efficient on adjusted magnitude over a window of 5 seconds. This allows us to capture longer-term trends as well as short-term responses. We can isolate instantaneous peaks by differentiating the signals or doing matched filter correlations. To analyze long-term trends, we also measure correlation between the slopes for skin conductivity signals.

### 2.3.1 Pilot Study - Movie and Television Audiences

*Scenario : A few friends are intently watching an exciting movie together. The scene looks perfectly normal, until one notices that their baseball hats have accelerometers on them, and they are wearing small SC leads on their fingertips. Their animated reactions to various*

29

*scenes in the movie, reflected in their physiology could be invaluable information for the movie's producers and editors.*



Figure 2-3: (left) Three people watching a short movie while wearing skin conductivity leads, accelerometer hats and microphones. (right) Close up of the GSR leads

The GroupMedia system was used to measure audience behavior and reactions to short movies and commercials, in the form of their skin conductivity response, head-movements and speech features (intermittent conversation, laughing, and conversation after the clip). In our experimental protocol for short films, a total of 15 subjects (in groups of 3-4) were shown a three-minute short film, followed by five minutes of conversation, and then by another nine-minute film. The subjects used a real-time touch-screen application on their PDAs to give a moving scale of interest that could be correlated with their features.

Figure 2.3 shows skin conductivity responses for two individuals, measured for about twenty-five minutes while the subjects watched the two short films. The first short film was generic, and the second was specifically related to the academic institution of one of the two subjects. The vertical lines in figure 7 indicate changes in activity. Section 1 was when subjects were watching a short film, section 2 was when they engaged in conversation, sections 3 to 6 was when they watched another short movie and during section 7 they filled out final survey forms and spoke informally.

As seen, there is a sharp rise in SC values at the start of section 2 for both subjects; this is when they started conversing at the end of the first short film, and one of them asked a question. Points A [during conversation] and B [during the second short film] reflect typical
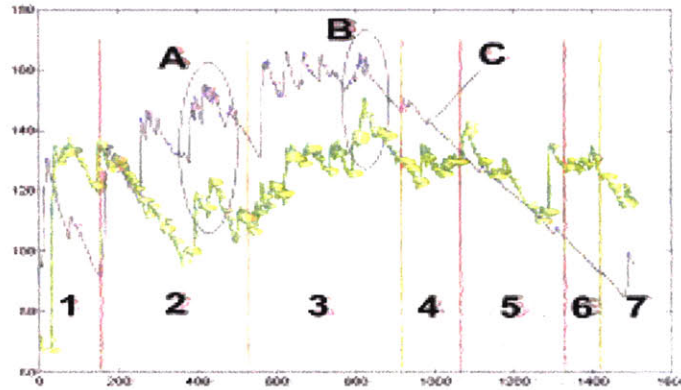
Figure 2-4: Skin conductivity signals for two people, where the vertical lines indicate contextual events in the epoch, like watching a short film(1), conversation (2), watching another short film(3-6) and filling out survey forms (7).It is seen that several regions (e.g. A, B) are highly similar while others (e.g. C) are not.

areas of strong correlation in SC responses of both individuals experiencing identical startle, cognitive or emotional events. On the other hand, C represents the start of a long-term drop in SC response, for a subject who was not related to the academic institution that the second film was about. By using the survey forms and the interest-ratings, we were able to correlate these strong physiological reactions with particular scenes and effects in the movie, and this feedback was provided to the movie's creators.

In a related study, we evaluated the reactions of 12 subjects (in groups of 2-4) to short (30 to 90 second) commercials and clips. One of these was expected to be boring, another was expected to be extremely funny and amusing, and the third had an unexpected shocking ending. The graph above shows the SC spikes for the clip with an unexpected shocking ending for 4 subjects viewing it (S1 and S2 together, S3 and S4 together) - prominent spikes at the surprise ending can be detected with slope detection on the SC signal.

This form of physiological "annotation" across multiple people illustrates the concordance in SC that may be useful towards identifying external cognitive or emotional events. If the link between behavior, physiology and interest for movie audiences is better understood it could also provide creative feedback for movies, commercials and other experiences.

31
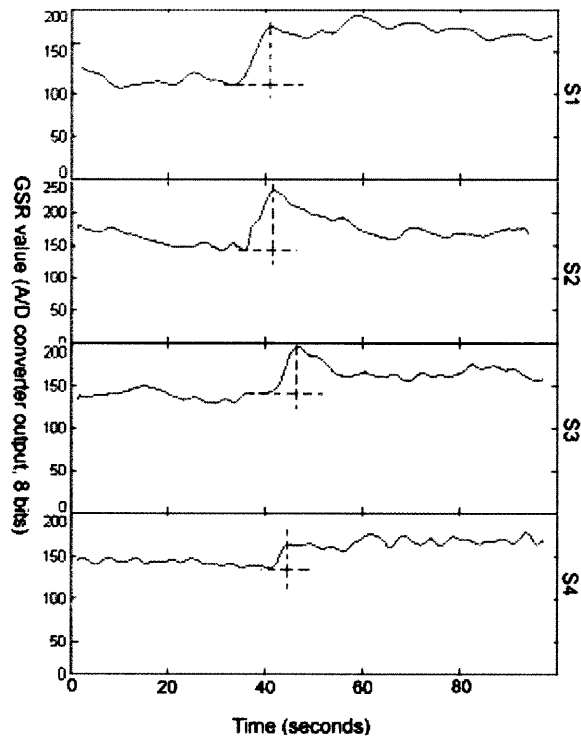
Figure 2-5: Almost simultaneous SC spikes observed for 4 different subjects (across 2 sessions) watching a MasterCard commercial with an unexpected ending

## 2.4   Body Movement and Head Nodding

There is extensive evidence about the role of gestures and body motion as non-verbal behavioral cues (Burgoon [4],Henley [32],Knapp[37], Leathers [40], Ambady [1]). Although speech and gestures are different modalities, they are certainly linked on several levels and work together to present the same semantic idea. Valbonesi et.al. [57] observe that emphasis in multi-modal communication can be estimated by combining vocal prosody and hand motion data measurements.

Body motion can be converted into many different usable features. These include but are not limited to the head nodding energy, head nodding frequency, concurrent head-nodding in a group, generic body motion derived from an accelerometer signal, motion from specific activities such as walking, standing, running, etc. An interesting notion is that it may

32

not be necessary to completely understand gestures to derive clues about the underlying behavior. A related study [27] shows how accelerometer motion at the mid-torso level (using an electronic name badge platform) can be used as a feature that predicts interest and engagement, and even the possibility of exchanging business cards.

### 2.4.1 Head Nodding

Across cultures head nodding has been observed as an indicator of agreement or understanding while a headshake has indicated a negative signal for disbelief or disapproval. Kapoor [35] mentions that head nodding also plays a role in conversional feedback where both the speaker and the listener nod synchronously, to keep the conversation flowing. At the most basic level, this behavior can be compared to the reptilian principle *isopraxism*, where one animal imitates the behavior of the other.

Another possible role of head nodding may be derived from the chameleon effect as described by Chartrand and Bargh [12]. They observed that people mimic body movements of their conversational partners and that this is reflected more in the behavior of a seemingly more empathetic person. In a recent study Briol and Petty [3] showed that head nodding could affect the attitude in the person itself. All these findings can explain how unconscious head nodding could be an index of or even affect the interest and dynamics between both the speaker and the listener.

### 2.4.2 Pilot Study - Conversational Interest and Head Nodding

*Scenario : Group deliberations and decision-making are an integral aspect of Sloan Business School. Four business school students are keenly involved in an animated discussion to find a class project they have to execute as a team. Individuals are using a real-time touchscreen application on their Zaurus PDAs to give an objective rating of how interesting they find the ideas. This can be correlated with their head-movement and nodding, speech features and physiology, to understand the socio-physiometrics of brainstorming and idea-generation.*

The GroupMedia system has been used to measure conversational interest in ten sessions of an MIT class called Digital Anthropology. Each session involved a group of 3-4 people, for durations ranging from ten minutes to an hour. The students engaged in conversation and brainstorming sessions, while we evaluated their physiometrics and subjective interest ratings. Head movement and nodding were analyzed for the group as an aggregated group statistic. Various characteristics like group head nodding energy, variance and means were calculated and compared to the overall interest ratings for the group. In addition, results of the head-nodding classifier (described in the next chapter) were also compared to the interest ratings.
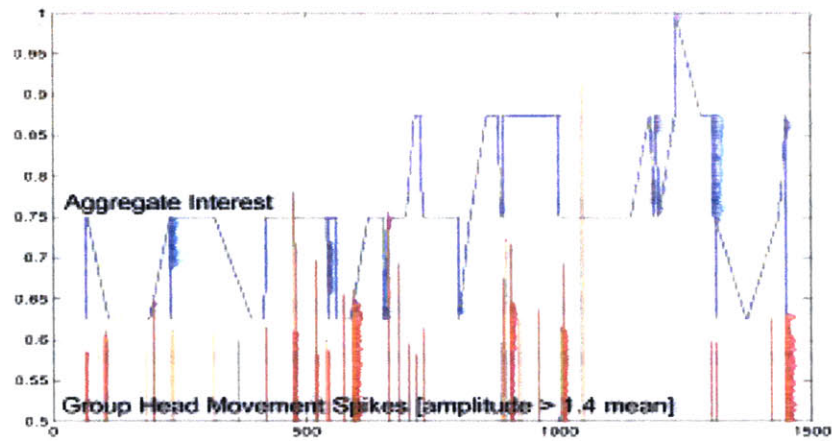


Figure 2-6: Plot of real-time interest ratings and corresponding overall group head-nodding energy

There was some correlation observed between the overall head-movement in the group and objective interest. The figure shows clusters of head nodding behavior when there are changes in interest ratings. Bursts of group head nodding correctly identify 80% of the changes in group interest level, with a 1/3 false alarm rate. Head nodding is not a perfect indicator of interest change, nor does it give the sign of the change, but it does provide the very useful information that we should look for changes in participant behavior and interest.

34

## 2.5 Location and Proximity

On a longer time scale, patterns of association between people and places are indicators of interest. Eagle and Pentland [21] have shown that coarse location data (from cell tower IDs) and proximity data (from Bluetooth IDs) for over 100 participants have significant correlation with relationship types (friends, work colleagues, etc.) and can be used to infer relationships with over 90% accuracy. Terry et.al.[55]'s Social Net is a similar system where patterns of collocation over time are used to infer possibly shared interests and then recommend introductions. In [27] proximity to other people and projects was used in a slightly different manner, as a feature to predict interest and objective outcomes.

Coarse location and proximity information can be easily measured in modern day handhelds and smartphones (using either wifi access point IDs or cell tower IDs for location and blue-tooth for proximity) without adding additional hardware in the environment. We believe that this information can add substantial power of context and increase the accuracy of predictions based on other features.

## 2.6 Features Summary

In this chapter we explain why four sets of features measurements of speaking style, physiology, body movements and location and proximity play an important role in measuring and predicting interest and engagement.

Speaking style is measured using the activity, stress, empathy and mirroring measures. The stress and activity features are derived from affective computing literature, and the empathy and mirroring features are conversation features. Skin conductivity, when measured as a group effect, can be a good indicator of sympathetic arousal. Body movements, and especially head nodding are channels of non-verbal social signaling. Over a longer time scale, patterns of association between people and places are indicators of interest. These can be inferred from location and proximity measurements. We also use these features in two pilot experiments. The first experiment shows how concordance in skin conductivity

for movie or television commercial audiences may identify areas of excitement and provide creative feedback. The second shows how head-nodding may be an indicator of changes in interest ina group discussion.

We have evaluated these features on the GroupMedia system, and believe that they can be implemented in real-time on a smartphone (with supporting hardware for measuring physiology). The experimental results in the next proceeding chapters are largely based on the speech features.

# Chapter 3

# GroupMedia system

In the Wearable Computing field there are many mobile platforms devised to measure physiology and body motion and do inference, most notably at GaTech[16], ETH Zurich [24] and as commercial products like BodyMedia [9] and LifeShirt [42]. Some of these have been designed to support specific applications like predictive healthcare, intelligent assistants, facilitation of social interaction, etc. Speech recognition is commonly used as input on cell phones, wearable computers, PDAs, etc. However, to our knowledge, there are not many mobile platforms that use non-linguistic speech features to predict behaviour and social outcomes as we have done.

We consider it important that the system embodies some basic design principles. It needs to be a clean design and reliable hardware/software that can collect continuous data and also provide feedback to the user. We wanted the system to be mobile, unobtrusive and certainly noninvasive. Measuring additional features like speech should not take anything away from the user experience of the PDA or smartphone.

Our hardware and software infrastructure is an extension of DeVaul's MiThril wearable computer [20], a proven accessible architecture that combines inexpensive commodity hardware, a flexible sensor/peripheral interconnection bus, and a powerful light-weight distributed sensing, classification, and inter-process communications software layer to facilitate the

development of distributed real-time multimode and context-aware application. Our development efforts have been towards adding speech feature calculation and real-time processing capabilities and converting the individual-based MiThril into a multiple-user, group-centric system. We call this platform the GroupMedia system.

The major components of the GroupMedia-MIThril architecture are: a PDA-centric hardware configuration, the Enchantment software network and resource discovery API, the real-time machine learning inference infrastructure, and secure ad-hoc networking over 802.11 wireless. Each of these components is briefly described in the following sections.

## 3.1   Hardware Platform

The GroupMedia system is built around the Zaurus SL 5500/6000 series PDA-based mobile device. This mobile platform is capable of real-time data analysis, peer-to-peer wireless networking, full-duplex audio, local data storage, graphical interaction, and keyboard/touch-screen input. The SL6000 also features bluetooth connectivity via an external CF card.

The system uses the Hoarder/SAK 2 sensor hub that to connects to the Zaurus over serial RS232. The SAK2 itself uses the i2c bus to connect to multiple sensors, which can include one or more accelerometer(s) (tri-axial Analog Devices ADXL) and up to 2 channels of skin conductivity(SC) leads. The analog SC signal is amplified and digitized on a daughter biometric(Bio) board. The SAK2/Bio board combination also supports other physiological measurements like heart-rate (using a Polar chest heart-rate monitor), Electro-Cardiogram (ECG), respiration (using a chest strain-gauge), and skin and core body temperatures. The SAK2/Hoarder board can function independently as a compact flash based data acquisition platform. The Zaurus/SAK2 combination is powered from external InfoLithium batteries that can support sensing and wireless transmission for up to 8-12 hours without being recharged.

DeVaul's core system is almost 2 years old, but even with the explosive growth in off-the-shelf smartphone technology, it has been hard to replace. We have built a next-generation
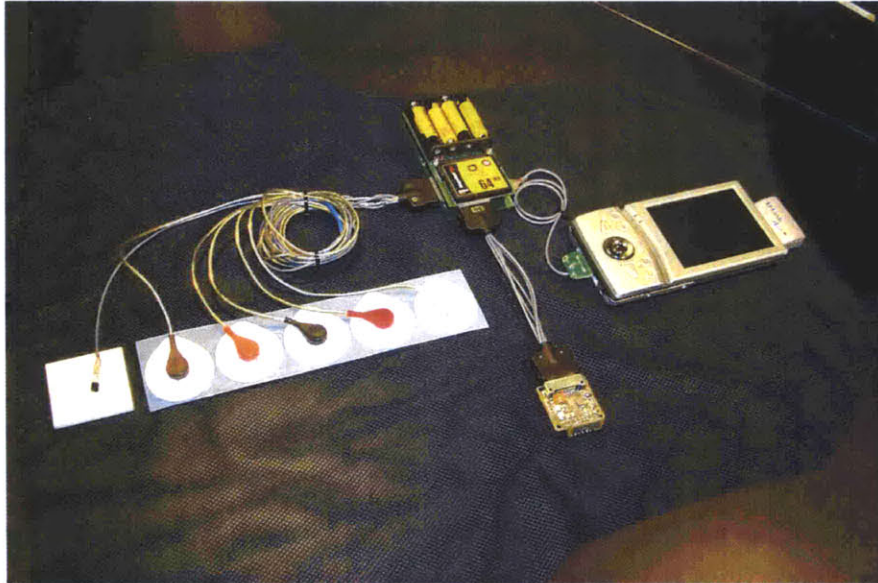
Figure 3-1: GroupMedia-MIThril 2003 system, comprised of the Zaurus SL 5500 PDA (right) with SAK2 sensor hub and physiological sensing board (top), EKG/EMG/SC/temperature electrodes and sensors (left) and combined three-axis accelerometer, IR tag reader, and RS232/I2C bridge (bottom)



Figure 3-2: Groupmedia bluetooth system, comprised of the Zaurus SL 6000 PDA, bluetooth skin conductivity sensor [52] and prototype bluetooth accelerometer (right)

version based around the SL6000 Zaurus and uses wireless bluetooth sensors. It has been successfully used with AC bluetooth skin conductivity sensors [52] and with bluetooth accelerometers [10]. At the time of writing of this thesis, we were also working on our own bluetooth accelerometer design, based on an AVR microcontroller, ADXL sensors and Taiyo Yuden low-power bluetooth transceiver. Although a bluetooth-based wireless system offers great flexibility (especially with smartphones), it is also more power-hungry and requires distributed power sources. With bluetooth sensors, the expected battery life reduces by about 50%. The Zaurus is a very versatile platform, and the latest version (CL3000) has a 4 GB hard-drive, amongst other features.

Ergonomic considerations for certain studies (like speed-dating) require wireless sensing and a smaller form-factor, which led us to use BodyMedia physiology measurements. Our system measures SC at the palm while BodyMedia devices measure it at the upper arm (fewer perspiration glands).We discovered that our system was more sensitive for SC measurements, as would be expected.

## 3.2   Software and Networking

The Enchantment whiteboard and signaling software is a light weight low cost means of routing information transparently across processes or distributed devices, and allows for easy routing and storing of raw signals and machine-learning classification results. Enchantment is intended to act as a streaming database, capturing the current state of a system, person or and on modest embedded hardware can support many simultaneous clients distributed across a network and hundreds of updates a second. Information transfer has SSL security and can be used by the real-time context classification engine or saved for offline analysis. The GroupMedia systems were modified to use SQLite on the Zaurus to build a long-term history of interaction, for example, over a few days, on a 1 Gigibyte capacity SD card.

The GroupMedia system also implements an advanced networking infrastructure for proximity and location detection and dynamic streaming of data. The system can switch easily from client-server mode to a peer-to-peer communication mode (over ad-hoc wifi), to share

information with other Zaurii in the vicinity. This is important because several studies require ad-hoc synchronization between multiple users for accurate sensing and timestamps. In addition, running kismet in warwalk mode allows us to measure approximate location (home, work, lab) and bluetooth scans allow us to measure people in proximity.

## 3.3   Speech Feature Processing

Significant modifications were made to the MiThril system to enable speech feature sensing and processing. Our first version allows for storage of audio in a custom format on the Zaurus, which is converted to speech-features off-line. The custom format is a stripped down version of wav files (for efficient usage of SD card space) and is available only to the authors (in consideration of privacy guidelines for human experimental subjects). Our more recent version [14] allows for calculation of activity (speaking time, voicing rate), stress (spectral entropy, standard deviation of pitch and std deviation of energy) and empathy (back and forth interactions) in real-time on a per-minute basis on the Zaurus. The spectral components use a fixed-point FFT (as the Zaurus does not have an FPU), yet the real-time engine has been optimized to consume less than 30% of processor resources at runtime.

## 3.4   Real-Time Head-Nodding classifier

The real-time head-nodding classifier is adapted from a Gaussian mixture model for motion classification (based on frequency of motion) devised by DeVaul [20]. The classifier performs an FFT transformation on the original 3-axis accelerometer signal and uses the resulting 31-dimensional features in a 2 class/2 component model. The system can accurately detect (vertical) head nodding. The model is trained off-line using Expectation Maximization (EM) and the real-time classification engine runs on the Zaurus PDA.

## 3.5 System Summary

This chapter describes the GroupMedia-Mithril hardware and software infrastructure which is used for data collection and real-time feedback in the following studies. The GroupMedia system is a wireless, distributed wearable system based on the Sharp Zaurus SL 5500/6000 and is capable of real-time data collection and inference for speech features, physiology, body motion and location and proximity information. The system is designed to support group interactions and features advanced networking and communications abilities. The ARM processor can easily run a real-time speech feature extraction engine. The system can also provide real-time feedback through the touch-screen interface or offline feedback to a central server.

Figure 3-3: BodyMedia Sensewear Armband, measures 2-axis accelerometry, skin conductivity, heat flux and skin temperature

# Chapter 4

# Conversational Interest

This chapter and the next two chapters describe various studies undertaken in order to better understand the role played by speech features in predicting interest and objective outcomes. In this study, we tried to measure speech features and predict self-reported interest in short (3 minute) conversations.

This study in measuring conversational interest is novel in several ways. From the perspective of psychological literature regarding human engagement and conversational scene analysis, we are building an automated interest measurement system, grounded in advanced machine perception techniques. Such a system can measure speech features that may be impalpable to a human observer, and even has the potential to replace the human observer entirely. Furthermore, we take a different approach from affective computing since we do not attempt to measure the underlying emotional states of the participants but rather the social signals that indicate interest in conversation.

As a starting point, it is important to distinguish between interest and engagement. According to Sipe, [53] Vocal engagement is a part of *being interested* and is observed as a physiological reaction where the person may be excited, aroused, may try to talk back, dramatacize or even take over the conversation. Aoki and Woodruff [2, 58] try to measure this form of conversational engagement using voice activity and prosodic cues. They measure engagement from conversational models of turn taking (a rudimentary relation to the

influence or engagement measure in our case). The engagement level for a voice channel can be used to change the communication parameters (e.g. if both parties seem engaged, the channel can switch from half-duplex to full-duplex).

Yu et. al. [59] use a more comprehensive model for measuring vocal engagement. They use a feature set similar to ours, including (functions of) fundamental frequency, energy, formants (harmonics of fundamental frequency) and coupled Hidden Markov Models (HMMs) to measure turn taking (similar to our influence parameters). While our labels are self-reported (after every 3-minute conversation), part of their data set consists of shorter spoken utterances that were hand-labeled for *engagement, valence and arousal*. They report approximately 60% accuracy on a 5-point scale for engagement in continuous speech.

Our definition of *interest* is for a longer duration conversation, and it captures a phenomenon quite different from these utterances of vocal engagement. We believe we have a more elaborate model that captures both the individual's speaking style (stress, activity) and the conversational dynamics (mirroring, engagement); vocal engagement for short utterances is part of the information we capture.

The study consisted of 20 participants, 10 males and 10 females, paired up with a same-sex conversation partner. (This was done to eliminate the possibility of attraction, which is explored in the next chapter. All participants were heterosexual). Each pair participated in 10 successive short (3 minute) conversations, hence each study lasted 30 minutes. The topics were randomly generated, from lists that both participants and experimenter came up with just before the start. After each conversation, they reported their interest rating on a scale of 1-10, 10 being highest. In 7 of the 10 sessions, the participants were complete strangers (the exceptions were for 2 male sessions and 1 female session). The data was checked for audio quality and 38 three minute sessions were removed (primarily because there were not enough speaking transitions to calculate the engagement measure using the influence model).

## 4.1 Results

The overall distribution of (sum of) ratings for each participant is shown in figure 4-1. As seen, people are consistently spread over all ratings.
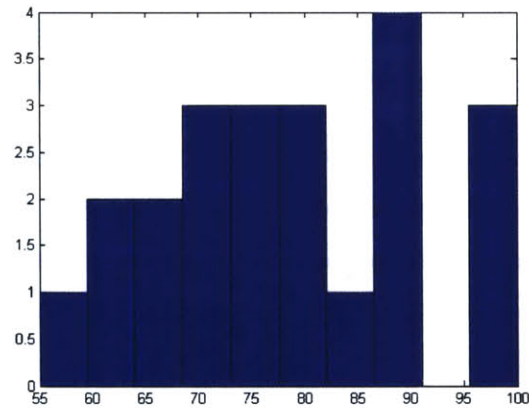


Figure 4-1: Distribution of (sum of) ratings for each person

### 4.1.1 Male Interest

The interest answers were correlated with the speech features from both men stress, activity, influence and mirroring. For the men, using all features we can predict 44% of the variance (r=0.66, p=0.02). Using the stress and activity measures alone (for both men), we can predict about 35% of the variance (r=0.59, p=0.01). The activity features predict 27% of the variance (r=0.52, p =0.007). The influence feature and back-and-forth features by themselves did not produce statistically significant correlations.

In addition to the actual ratings, we also calculated the mean rating for each male over the 10 sessions. Heuristically, interest in a conversation has many contributing factors: the person's interest in the topic itself, compatibility with the other person, his mood or state that day, etc. The mean rating for all 10 successive conversations, in a sense, represents factors other than the topic itself, and we call it the mood + compatibility factor. Measuring the mood, compatibility and various other hidden variables that might contribute would

require an array of psychological tests, many of which may bias the results or subjects. We believe that the mean, in a sense, represents that information, as it excludes the contribution made by the conversation topic. For men, the correlation of speech features with this mean rating was slightly higher than with their actual ratings (r=0.7144, p=0.003) implying that it is possible to predict their mood + compatibility for these conversations slightly more accurately than their overall interest in that conversation. Stress and activity from the person himself showed a high correlation with this factor (r=0.56, p=0.0001).

Another reason to calculate the mean was to verify if people really had different scale factors for their answers. Hence, instead of using the actual answers, we considered the deviation from their mean as the answer. For the men (unlike the women), the speech features from both participants showed significant correlation with the deviation from the mean (r=0.697, p=0.007). Again, stress and activity features were the most important (r=.63, p=0.003). The figure below shows the distribution of deviation-from-mean for the answers.
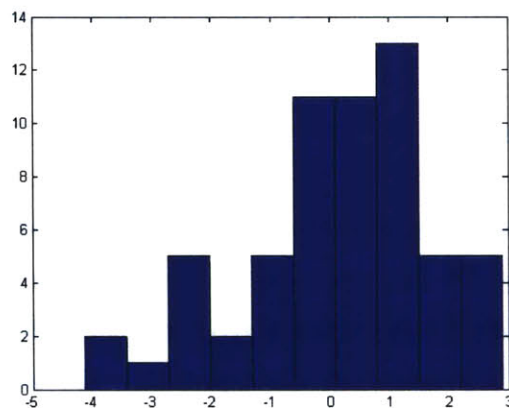


Figure 4-2: Distribution of deviation from mean ratings (or the 'mood + compatibility' component) for men

If we restrict ourselves to speech features from one male only, we find that his activity, stress, engagement and mirroring interactions, contribute to his mean rating (r=0.5634, p=0.01) but not significantly to his actual interest answer. On the other hand, speech features from the other person contribute to the first persons interest answer (r=0.57, p=0.01) but not towards the mean. In a sense, this might imply that a persons speaking style alone

is reflective of their mood + compatibility component that day, while the other persons speaking style can reflect the first persons overall interest in the conversation.
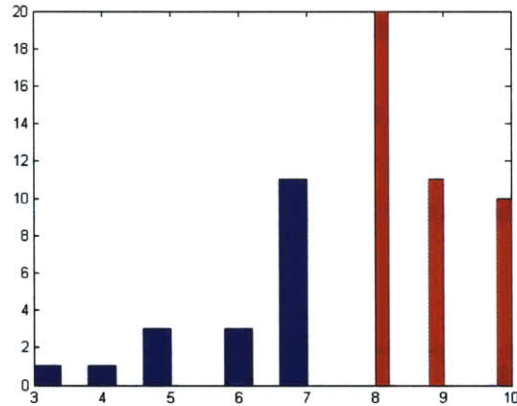


Figure 4-3: Dividing Male ratings into 2 classes

The distribution of male ratings can be split into a two class model we labeled them as high interest (rating >=8) and low interest (rating < 8), marked in red and blue in the figure. For this 2-class model, the speech features explain 47% (r=0.69, p=0.01). Stress and activity features for both people also show a high correlation (r=0.595, p=0.01).

Using the activity features (speaking time and voicing rate for both men) in a linear SVM classifier, it is possible to classify samples into these two classes (red and blue) with 74% accuracy.

### 4.1.2 Female Interest

The speech features are again highly correlated with the interest answers for women, and explain 45% of the variance (r=0.67, p=0.006). The stress and activity measures (for both women) predict just over 35% of the variance (r=0.6, p=0.004) of her interest answers. Unlike the male responses, the influence/engagement measures (r=0.46, p=0.05) and only her stress and activity measures (r=0.48, p=0.006) also played an important role.

Similar to the men, we also calculated the mean rating for each woman over the 10 sessions,

and labeled it the compatibility + mood factor. The speech features showed significant correlation with this mean (r=0.631, p=0.03). Stress and activity for both women (r=0.6, p=0.004), her stress and activity measures alone (r=0.53, p=0.001) and the engagement measures (r=0.36, p=0.06) also showed significant correlation.

In women, only the engagement parameters had a significant correlation (r=0.4, p=0.03) with the deviation of means for the answers. Speech features from just one woman are correlated with both her actual interest answer (r=0.58,p=0.005) and her mean rating (r=0.58,p=0.003).
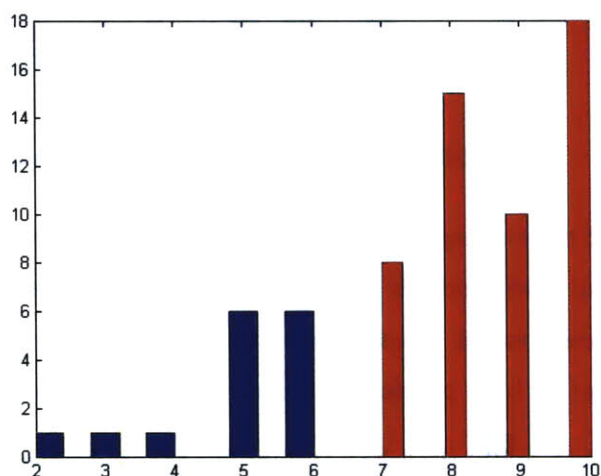


Figure 4-4: Dividing Female ratings into 2 classes - high and low interest

Similar to men, the distribution can be divided into classes of low interest (blue, answer < 7) and high interest (red, answer >=7). All features together explain 42.5% of the variance (p=0.01) and the stress and activity measures explain about 35% of the variance (p=0.005).

## 4.2 Discussion

The speech features explain about 45% of the variance in self-reported interest ratings on a scale of 1-10 for both men and women. Interestingly, the stress and activity measures

alone explain about 30-35% of the variance in ratings. This shows that our speech features capture at least some of the verbal body language and social signaling that people use when they are interested in a short conversation. In the next two chapters, we explore the role of these features in attraction and focus group situations.

Furthermore, these features can also be used to build a real-time classifier. The activity measures alone (speaking time and voicing rate) are used to build a high and low interest classifier that is 74% accurate (for men). Given the significant correlations, it is realistic to build real-time interest classifiers (on cell phones, wearables, maybe even VOIP networks) that are reasonably accurate ( 70% or more) using our features.

An interesting observation was that although we can explain 45% of the variance in the self-reported interest rating, we also explain 40% of the variance in mean rating, or mood + compatibility factor for both men and women. This leads us to ask the questionwhat are people really interested in? If self-reported interest ratings are really a function of the persons state, personality, and compatibility with the conversation partner, then they cannot be used as an objective measure of interest in products and services. We expected that the focus group study would shed more light on this result.

This effect was more pronounced for women, as there was very little correlation between our speech features and the deviation from mean rating. There were other differences between men and women, for example, the engagement feature played a role in predicting female interest. This suggests that the phenomenon of *being interested* in a conversation is different for men and women. Popularly speaking, "women are from Venus, men are from Mars"

# Chapter 5

# Attraction

Speed dating is relatively new way of meeting many potential matches during an evening. Participants interact for five minutes with their 'date', at the end of which they decide if they would like to provide contact information to him/her, and then they move onto the next person. A 'match' is found when both singles answer yes, and they are later provided with mutual contact information. Perhaps since speed dating provides instant-gratification in the dating game, it has recently become the focus of several studies that attempt to predict outcomes based on mate selection theory. Kurzban and Weeden [39] surveyed 10,000 speed-dates and concluded that outcomes are highly correlated with physically observable attributes like attractiveness and age.

Our motivation behind this study was to extend our definition of automated interest measuring and verify if it could be applied to social outcomes. Although it is somewhat understood as to what constitutes an *attractive voice* (e.g. Oguchi [45]), not much has been done to show if tone of voice or speaking style can be used as a predictor of attraction, especially in a speed-dating like face-to-face setting. Speed-dating is of particular experimental interest because there is a clear *'buying decision'* that has to be made at the end of each conversation.

In this experiment we analyzed 60 five-minute speed-dating sessions. Data was collected from several events held in 2004, and participants were singles from the MIT-Harvard

community in the ages of 21-45. In addition to the 'romantically attracted' question, participants were also asked two other yes/no questions: would they like to stay in touch just as friends, and would they like to stay in touch for a business relationship. If they replied 'yes' to any of these 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. Since our data was collected from real-life sessions, it was also more immune to the effect of self-entertainment. (i. e. we classify participants' natural intonation, not special 'interested' or 'bored' voices generated by actors).
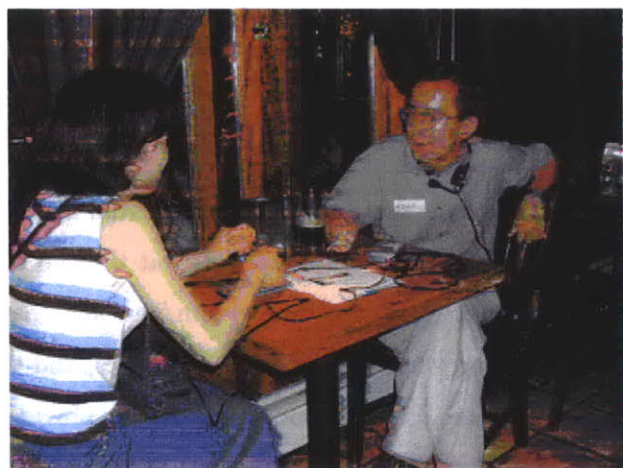


Figure 5-1: Speed dating session in progress. Audio is recorded using Zaurii PDAs, participants are seen wearing noise-cancelling microphones

## 5.1 Results

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 could account 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 engagement measure was the most important individual feature for predicting the

'friendship' and 'business' responses. The mirroring measure was also significantly corre-
lated with female 'friendship' and 'business' ratings, but not with male ratings. The stress
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.
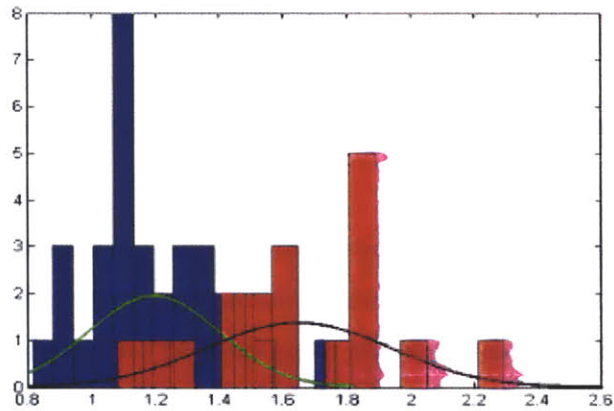


Figure 5-2: Frequency distribution of female 'attracted' responses (red=yes) versus predictor
value. The cross-validated binary linear decision rule has 72% accuracy

Figure 5.2 shows a two-class linear classifier for the female 'attraction' responses, based
on the social signaling measures; this classifier has a cross-validated accuracy of 71% for
predicting the 'attracted' response. Feature selection was based on the regression results.
The two fitted Gaussians are simply to aid visualization of the distributions' separability.

Figure 5.3 illustrates a two-class linear classifier for the 'business' responses, based on the
social signaling measures; this classifier has a cross-validated accuracy of 74% for predicting
the 'attracted' response. By considering the overlapping region as a third class, we can
increase the cross-validation accuracy to 83% for the yes and no response regions. The two
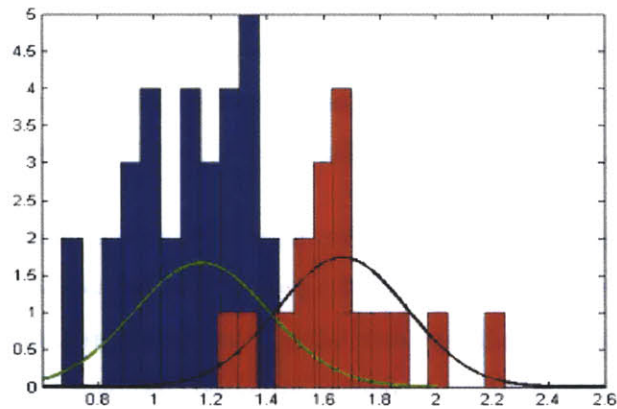
Figure 5-3: Frequency distribution of female 'business' responses (red=yes) vs. predictor value. The cross-validated three-class linear decision rule produces 83 % accuracy.

fitted Gaussians are simply to aid visualization of the distributions' separability.

It was also observed that the accuracy of the predictions increased when the classifier was trained for a particular person. We believe this is important because people have different speaking styles, which potentially could be captured by our model. We had 8 conversations involving subject J and trained our model only on her data set. For the romantic interest question for subject J, for example, the cross-validation accuracy of our predictions increased to 87.5%.

We also used SVM classifiers to test the separability of the 'yes' and 'no' responses, with linear, polynomial and radial kernels. As seen, the performance of the SVM classifier is only slightly better than a simple linear classifier (with appropriate feature selection). The RBF kernel is more prone to overfitting.

| Question | SVM Accuracy (Linear Kernel) | SVM Accuracy (RBF Kernel) |
|---|---|---|
| Are you interested in this person? (asked of females) | 0.71 | 0.70 |
| Would you like to be friends with this person? (asked of females) | 0.76 | 0.79 |
| Are you interested in maintaining a professional relationship with this person? (asked of females) | 0.73 | 0.71 |
| Are you interested in this person? (asked of males) | 0.64 | 0.62 |
| Would you like to be friends with this person? (asked of males) | 0.79 | 0.82 |
| Are you interested in maintaining a professional relationship with this person? (asked of males) | 0.71 | 0.69 |

## 5.2   Real Time Feedback

If our description of social signaling is precise, and the effects are measurable, then a wearable 'social signaling meter' is the next step. We wanted to use such a device to understand how speed daters would react if made more aware of their social signals.

We have built a real-time speech feature calculation engine (SpeedDating Pro) that runs on a Sharp Zaurus Linux PDA. SpeedDater Pro is adapted from the real-time speech feature processing implementation for the Zaurus. The Zaurus is a wireless, ARM-based handheld with a microphone input and a touch screen interface. The current version of the speech feature code is written in C++, and incorporates only the stress and activity features (the mirroring measure has been implemented but required time syncing between Zaurii. The engagement measure has not been implemented). It also calculates the probability of the

57

Figure 5-4: (left) Display of probability of the other person saying YES (right) Display of measured social signals

other person saying 'yes' based on the coefficients from the model training. Zaurii are time synced using ntpdate (a linux utility) and use UDP broadcasts to communicate with each other. The display can show chances of a second date or the measured social signals, and is updated after every 90 seconds of conversation.

Although we have not yet conducted a rigorous experiment using the real-time system, it has received mixed reactions from users. Many users (usually early adopters of technology) were amazed by the idea that computers and cellphones could predict how well they (humans) were doing in the speed dating game. Others (especially older women) found it that it changed certain aspects of the interaction.

## 5.3  Discussion

The social signaling seems to communicate and be involved in mediating social variables such as status, interest, determination, or cooperation, and arise from the interaction of two or more people rather than being a property of a single speaker. Semantics and affect are important in determining what signaling an individual will engage in, but they seem to

be fundamentally different types of phenomena. The social signaling measured here seems to be a sort of 'vocal body language' that operates relatively independently of linguistic or affective communication channels, and is strongly predictive of the behavioral outcome of dyadic interactions.

The correlations indicate that features that did not play an important role in conversational interest (mirroring, engagement measures) were more predictive in the case of opposite-sex attraction. This suggests that the behavioral phenomenon of attraction reflected in speech is different from that of interest reflected in conversation.

Often male participants ask us (as researchers), as to what would be a good strategy to employ in speed dating. Based on these results, it seems that it is important to let the woman talk (activity measure), try not to dominate the conversation (engagement measure), and show some verbal empathy by using words like ah-uh, hmm.. etc. (engagement measure).

# Chapter 6

# Focus Groups

According to ESOMAR, annual spending in qualitative market research in the US alone is about $2B every year [23]. Qualitative market research is used to develop products/services and target customers in consumer goods, movies, television content and commercials, advertisements, travel, tourism, healthcare, and other industries. The most common form of qualitative consumer research is focus groups; other methods include one-on-one interviews, ethnographic studies, telephone surveys, online focus groups, etc. Consumer research analysis drives mission-critical decisions such as the focus of advertising campaigns and product development, but a significant portion is misallocated as a result of unreliable research. However, nearly 4 out of 5 new consumer products fail or seriously fall short of expectations [60, 5] , largely because the developers did not fully understand customer preferences. Many experts in the industry agree that focus groups areoften inaccurate - yet, they are forced to rely on them for the lack of a better alternative. There is a significant need for these companies to better measure their customers true behavior and preferences. In this chapter, we try to apply our speech feature and interest metrics to focus groups in an attempt to increase their accuracy.

Kitzinger [38] defines a focus group as a form of group interview that capitalizes on communication between research participants in order to generate data. Instead of the researcher asking each person questions, people are encouraged to talk to each one another: ask

questions, exchange anecdotes, comment on experiences and perspectives. The method is particularly useful to examine not only what people think, but also how they think and why they think that way. The idea behind focus groups is that group processes can help people explore and clarify their views in ways that would be less accessible in one-on-one interviews.

A typical focus group session consists of 4-8 participants and a moderator trained to raise questions and guide the conversation. The group usually lasts for about 1-2 hours, and participants may be laymen or experts in a certain field. Almost always, monetary compensation for the participants is involved.

## 6.1 Issues with (current) focus groups

- Lack of Consumer insight – People often do not share true consumer insight; instead they may state what is expected of them, especially in situations where monetary compensation is involved. According to Patnaik, [51], traditional focus groups are like *a customer terrarium*, as people are displaced from their natural surroundings and behavior. This often makes it impossible for researchers to know what customers are feeling. According to Johnston, [34], as people are exposed to more (market) research, they begin to understand what is expected of them. Often they parrot back marketing or advertising messages that offer little consumer insight.

- Group Effects – Focus groups are known to be highly susceptible to group biases, e.g. group polarization and takings more *risky* (extreme) stances. The polarization effect is a known group psychology effect – when people are placed into a group and these people have to deal with some situation, the group as a whole typically has some overriding attitude toward the situation (Gray [28]) Geis et. al. [29] show that in focus groups for lovers of AIDs patients, there were many more angry comments about they medical community than one-on-one interviews, as the synergy of the group reinforced vented feeling of rage, allowing the group to take a more extreme stand.

- A common issue with focus groups is when certain individuals dominate the conversation, and minority insight vital to the discussion is suppressed. In Kitzinger's study of focus group discussions for elderly residential care patients, [38] she mentions how individual opinions are silenced by strong opinions of other people.

- Sometimes focus groups also produce inconsistent results because of variations in the moderator's style.

## 6.2 Hypothesis and Experimental Design

Digital Innovation (MAS.967) is a class taught to MBA and other graduate students every spring at the Media Lab. Each student in the class is expected to write an individual paper at the end of the semester, which determines part of their final grade. The paper has to be around one of the eight areas of innovation introduced in lectures.

Our focus group experiment used class project ideas as the subject of discussion. The study was held about 3 weeks before final papers were due, while students were still in the process of deciding their paper topics. The focus group discussions was moderated by MBA students who had past industry experience working with siilar focus groups.

The class paper topic presents an interesting discussion opportunity. Each topic was broad enough that many papers could be written about it. Peer discussions were likely to influence people to change their topic because another one might seem more interesting or easier to write about.

When students select a final topic, there is a clear *buying decision* involved  they have to write their paper on the same topic later. If a student wanted to change his/her topic later, he/she would require special permission from the TA. In contrast with traditional focus groups, this process closes the loop, as participants are obligated to pursue what they say. This allows us to compare participants speaking style with *what they really did*, and not simply *what they say*. Of course, it also gives them more incentive to state the truth during

the study. Our hypothesis was that speech features culd provide two types of insight in this study.

- The first was to measure speech features and correlate them with ill-effects observed in traditional focus groups, e.g. a particular speaker dominating the conversation and suppressing minority viewpoints, or polarization within the group (described in more detail in the previous section). This information could be used to automatically identify focus group discussions that were likely to be inaccurate.

- The second kind was being able to predict if people would change their final papers later based on their measured speaking style, e.g. someone who was not confident during discussion might request a topic change later (even though they picked one at the end of the discussion).



Figure 6-1: A focus group session in progress

Before start of discussion, each student was asked to select their (probable) final paper topic. Students then engaged in a 30 minute discussion about various topics. The moderator asked the following questions (each 5-7 minutes) to keep the discussion going. The sequence of events was as follows,

1. Students choose one topic out of eight and fill out feedback form

2. "What talks from the class did you find interesting?"

3. "What is you paper going to be about?"

4. "What factors do you think are important in choosing a topic?"

5. Students fill out feedback form with final chosen topic. This is assumed to be their final topic for class grade purposes, and any changes after this require the TA's permission

6. "So what made you choose that particular topic?"

## 6.3 Results

This experiment was unique because we scaled the speech feature analysis from a dyadic interaction to a group interaction (4 people). Our influence and empathy measures are well defined only for dyadic interaction, so the analysis was done by using the activity measure (speaking time, voicing rate) and stress measure (std. deviation of pitch, spectral entropy, std. deviation of energy), for every individual.

| Participant No. | Topic before discussion | Topic post discussion | Submitted Paper Topic | Submitted paper different? |
|---|---|---|---|---|
| 101 | 3 | 3 | 3 | No |
| 102 | 3 | 3 | 5,3 | Yes |
| 103 | 7,3 | 3 | 3 | No |
| 104 | 3 | 3 | Did not submit | Yes |
| 201 | 8 | 8 | 8 | No |
| 202 | 8 | 8 | 8 | No |
| 203 | 8 | 7 | 7 | No |
| 204 | 5 | 5 | 5 | |
| 301 | 8 | 2 | unrelated | Yes |
| 302 | 8 | 8 | 8 | No |
| 303 | 3 | 3 | 3,7 | Yes |
| 304 | 8 | 8 | 8 | No |

As seen, after discussion with their peers, one person in each session changed his or her topic (103, 203 and 301). Although three sample points are in no way statistically significant, speech features may provide some insight about the behavior of the groups.

The graphs below show the interactions of the second focus group in more detail (subjects 201-204, question 1). Fig 6-2 shows the raw audio and the results of speaking/not-speaking detection after question one ("What talks from the class did you find interesting"). The stress and activity measures are calculated for each person from the speaking segments.

Figures 6-3 uses speech features to visualize each groups behavior for the first question (roughly first 5 minutes of conversation). Each circle represents one person in the focus group study. The radius of the circles is proportional to the fraction speaking time (activity measure), the thickness of the circle is proportional to the mean voicing rate (activity measure), and the thickness of the arrows denotes the transition probability (for the first question, the transitions between speakers were non-recurring and hence transitions had equal probability)

In each figure the grayed out circle represents a participant who changed his or her topic at the end of the discussion, i.e. participants 103, 203 and 301 respectively. It is very interesting to note that in roughly the first five minutes of conversation, each of these participants had the highest speaking time and the lowest voicing rate (or very close to the lowest).

Subjects 103, 203 and 301 who changed their topic post-discussion spoke the longest and the slowest (activity measure) during the first five minutes of conversation. When we listened to the raw audio for the first question to understand the cause, we found that the person was speaking in a slightly unsure manner, slowly and using words like 'umm', 'maybe' and so on. This might explain the indecision in opinion that was captured by the activity measures (fraction speaking time and voicng rate). Although this study is too small to be statistically significant, it shows that the speech features could potentially play a predictive role.

66
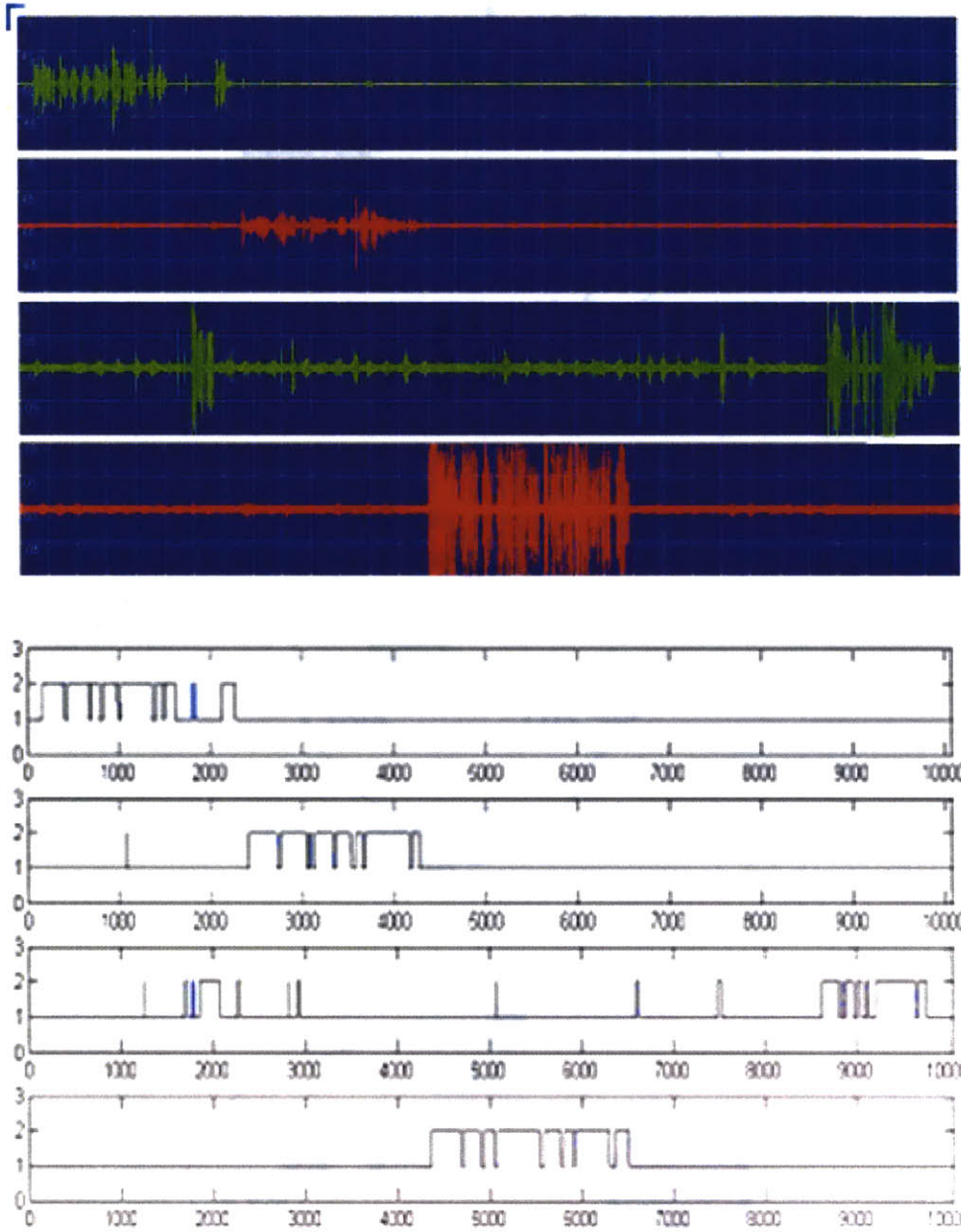
Figure 6-2: (top) Raw audio for subjects 201-204. (bottom) Results of speaking/not-speaking detection . The stress and activity measures are calculated for the speaking sections
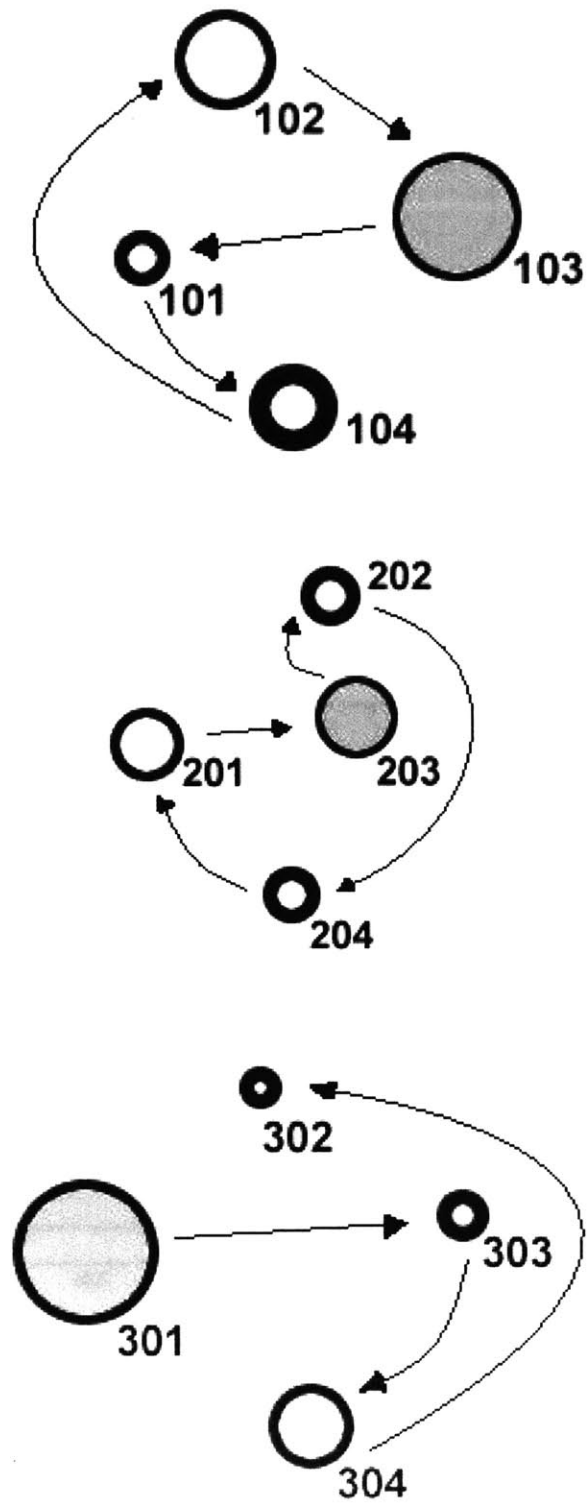
Figure 6-3: Interactions for all three focus groups for question 1. The circle radius is proportional to fraction speaking time and the thickness is proportional to voicing rate. The grayed out circle represents subjects 103, 203 and 301 who changed thier topic after discussion.

## 6.4 Discussion

Our initial analysis of interactions within focus groups shows that speech features may offer insight about the working of the groups. Preliminary analysis shows that they may help identify bad effects observed during focus group discussions, e.g. domination by a particular subject or identify subjects who are easily influenced. Focus groups are a ubiquitous market research tool in development and marketing of consumer products and services. The ability to filter out flawed or biased focus groups would be invaluable to these companies that spend $1.4 billion every year on focus group studies.

Speech features may also help predict final outcomes, in this case, the submitted paper. In the conversation interest study, the results suggested that self-reported interest was, to some extent, a function of the person's mood and compatibility that day. It is a promising direction for future research to see how that applies in the context of focus groups.is a promising direction for our future work.

# Chapter 7

# Interest Networks

The preceding chapters suggest that it is possible to measure and predict interest and engagement with a reasonable degree of accuracy. In this chapter, we describe a prototype application where a *wearable interest meter* is combined with location, proximity and social network context information. This wearable system uses machine perception to quantify a users interest level and social context and propagate relevant information to others in the his or her social network.

Although we have not evaluated the usability and data collected from this system, in this chapter we illustrate our ideas and possible applications.

## 7.1  Current System

The Interest Networks system for an individual is capable of measuring the following features:

### 7.1.1  Bluetooth Proximity

The system periodically performs a bluetooth scan for devices in its vicinity (currently set at 4 minutes). The choice of an appropriate period between bluetooth scans is a balance

Figure 7-1: Bluetooth based Interest Networks system based on Zaurus SL 6000 or Motorola A760. The system is capable of real-time speech feature processing and bluetooth and wifi scanning.

between detecting short interactions and conserving battery life. Scan results (BTIDs) are compared with a buddy list stored on the Zaurus. Bluetooth scanning is enabled on the Zaurus using the BlueZ stack implementation, and a scan takes about 10 seconds.

There is a simple graphical user interface to add new people (devices) to the Buddy List. This can also be automated, for example, the Zaurus can prompt for identifiers for frequently scanned devices.

## 7.1.2   Speech features

The system uses speaking/not-speaking detection, and then calculates the activity and stress measures for the speaking sections. The activity and stress measures can be calculated for an individual without having to synchronize the Zaurus with other people. From the results in chapter 4, it is reasonable to start with the stress and activity measures (they explain about 35% of the variance in conversational interest)

It is also possible to calculate the empathy and engagement measures for two or more people. This requires time-syncing between the Zaurii for different people (within 1 sec accuracy for empathy and 120ms accuracy for influence calculations). Time syncing is done through a middleware implementation (similar to enchantment), which exchanged wireless UDP packets.

### 7.1.3 Kismet location (warwalking)

The system uses wireless access point IDs to generate approximate location (within 100-300 feet resolution). This allows it to classify known / frequently visited locations like home, workplace, etc. Location inference can be done far more accurately using GPS (outdoors only), assisted GPS (indoors and outdoors), cell-tower/wifi triangulation, IR beacons, etc. However, using access point IDs does not require any additional infrastructure either in the environment or on the Zaurus.

The Zaurus runs kismet (a wireless network sniffer for Linux) in warwalk mode, scanning the area for various wireless networks and access point IDs. Kismet is invoked for about 20 seconds every 4 minutes. The Zaurus is connected to the wireless network (if available) between scans. Similar to bluetooth scans, when the Zaurus frequently scans particular wifi access points, the user can be prompted to provide labels.

## 7.2 Proposed Applications

### 7.2.1 Buddy Zing and Collaborative Filtering

Using the location, proximity and speech features, the system can build a real-time context for the user. In other words.. 'where are you, whom are you with, and is something interesting going on? With appropriate privacy restrictions, the system can share information between family, friends, and workgroups. For example, 'where are all your friends on

Saturday night, and who seems to be having the most fun? We call this system Buddy Zing.

In the context of workgroups, if the system detects several people from the same team in proximity and engaged in an *interesting discussion*, it can notify pre-approved distant users that they might want to 'patch in to the conversation. Upon receiving such a notification a distance-separated team member can subscribe to the audio channel or later look at the appropriate annotations.

By building implicit profiles of people's interest ratings, it might also be possible for the system to make recommendations. As an example... your best friend always has a great time at this coffee shop. Some of these applications may require more accurate location information..

## 7.3   Selective Recording and Annotation

An accurate wearable interest meter could be very useful in recording and annotating pre-approved interesting events and discussions. Startlecam [31] was a wearable project that used skin conductivity responses to start a video recorder. More recently, the diary application [22] uses location, proximity and the common-sense database to implicitly build a 'life-log.

The Zaurus features an optional video camera, and even without video, it is possible to record audio or start a Linux text-to-speech converter (available for the Zaurus from IBM). The latest version of the Zaurus, CL 3000, has a 4Gb hard drive, which provides ample storage space content-rich audio or video snapshots.

## 7.4   Qualitative Market Research

The advertising and market research companies spend about $16 B every year to better understand how customers think and why they think the way they do. Bernard et. al.

[11] show in their study how traditional forms of retrospective data reported by people in surveys, interviews, etc. ('how often do you do such and such, what did you think of it) can be highly inaccurate.

An automated, wearable interest meter could be used to better understand what part of an experience or activity really matters (also known in advertising as *relevancy testing*), by measuring consumer behavior and interest reactions in amusement parks, cruises, casinos, malls etc. With a wearable, unobtrusive device, this information could be collected for a much larger sample size (with consent, and in exchange of some form of compensation). This information could be used to make better products or services and create more focused advertising.

# Chapter 8

# Conclusions and Future Work

## 8.1 Summary

In this thesis we attempt to build a *human interest detector* using speech features, phys-
iological features, body movement, location and proximity. We explain these features in
detail and justify as to why they should be considered. The speech features, consisting
of activity, stress, empathy and engagement measures are used in three large experimental
evaluations measuring interest and engagement in conversation, attraction in speed-dating,
and understanding interactions in focus groups.

In the conversational interest experiment, the speech features predict about 45% of the
variance in self-reported interest ratings for 20 male and female participants. Stress and ac-
tivity measures play the most important role, and a simple activity-based classifier predicts
low or high interest with 74% accuracy (for men). This study provides evidence that our
chosen features do capture some element of the verbal body language and social signaling
that is indicative of human interest, within a few minutes.

In the speed-dating study, we use the speech features measured from five minutes of con-
versation to predict attraction between people. The features play an important role in
predicting the outcome of these dyadic interactions. They predict 40% of the variance in

outcomes for attraction, friendship and business relationships. Speech features are used in an SVM classifier that is 75%-80% in predicting outcomes based on speaking style. We use these results to build a prototype real-time dating meter on a PDA.

We then evaluate the insight that speech features might provide when measuring consumer interest in focus group study. Although our experiment and analysis are exploratory, the speech features help to identify a pattern of behavior where subjects changed their opinions after discussion.

Finally, we describe a prototype wearable interest meter and various application scenarios. We portray a world where cell phones can automatically measure interest and engagement, and share this information between families and workgroups. Such a system could automatically patch in pre-approved team-members into the conversation, or selectively record and annotate discussions. We have also received commercial interest in using it as a wearable tool for gather qualitative consumer research data.

We believe that this is just the beginning, and the results from this thesis show that it is possible to build a *human interest meter*.

## 8.2   Future Work

In the course of this work, we have opened up many exciting doors that would be worth exploring.

As a start, it is important to better understand the role of speech in predicting interest. It is important to conduct larger-scale studies and see if the measured effects change. It may also be possible to increase the accuracy of our predictions by combing speech content with non-linguistic features.

Although a generic interest meter is very attractive, a system like ours is likely to be more accurate if trained for a particular person. This could be implemented on a cell phone platform and the system could use continuous supervised or unsupervised learning from the

users behavior. Our attempts to build an individual classifier for the attraction experiment generated promising results.

From a consumer research application, focus groups pose a problem with large industry potential. Our analysis is only preliminary, but reveals that a machine perception based system may help to better understand the working of a group, and identify flawed or biased sessions. Our talks with consumer research and product development companies have already generated some interest.

There are other applications for this technology. For example, the adoption of Voice Over IP (VOIP) communications is growing exponentially. There are many concerns with limited bandwidth availability between several simultaneous calls.

A possibility is to use an active interest meter on the server-side to distribute limited bandwidth to calls where speakers are most active and highly engaged.

It is important to study and analyze the role played by the other features we have described - physiology, body movements, location and proximity patterns, in conjunction with speaking style. However, the space is vast and understanding the interconnections between all those features and interest could be the subject of at least several dissertations.

Finally, we personally find the wearable interest meter application and the idea of distributed interest networks extremely exciting. Once these features are better understood, such a device may change the way we interact and handle information.

There is so much more to explore, and so little time.

# Appendix A

# Supporting Material for Conversational Interest Experiment

## A.1 Anonymized Raw Data

Please visit our website groupmedia.media.mit.edu or email anmol@media.mit.edu

## A.2 Regression Tables

## A.3  Feedback Form

**Men (r value, p value)**
* indicates that the results are not statistically significant (p value > 0.05)

All features,
0.6630, 0.0265 - my actual rating
0.7144, 0.0037 - my mean rating, denotes personality + compatibility + mood that day
0.6975, 0.0076 - my deviation from mean
0.6889, 0.0106 - my 2 class interest rating
0.5886, 0.1758 - other person's actual rating
0.6755, 0.0174 - other person's mean rating, denotes personality + compatibility + mood
0.6698, 0.0211 - other person's deviation from mean
* 0.6148, 0.1009 - other person's 2 class interest rating

My features only
* - my actual rating
0.5634, 0.0158 - my mean rating, denotes personality + compatibility + mood that day
0.4982, 0.0851 - my deviation from mean
* - my 2 class interest rating
0.5042, 0.0746 - other person's actual rating
0.5207, 0.0507 - other person's mean rating, denotes personality + compatibility + mood
* - other person's deviation from mean
0.6016, 0.0044 - other person's 2 class interest rating

All his features only
0.5663, 0.0144 - my actual rating
* - my mean rating, denotes personality + compatibility + mood that day
0.5182, 0.0539 - my deviation from mean
0.6331, 0.0012 - my 2 class interest rating
* - other person's actual rating
* - other person's mean rating, denotes personality + compatibility + mood
* - other person's deviation from mean
* - other person's 2 class interest rating

Stress and activity
0.5889, 0.0128 - my actual rating
0.6606, 0.0001 - my mean rating, denotes personality + compatibility + mood that day
0.6284, 0.0031 - my deviation from mean
0.5951, 0.0105 - my 2 class interest rating
0.5330, 0.0612 - other person's actual rating
0.5998, 0.009 - other person's mean rating, denotes personality + compatibility + mood
0.5938, 0.0110 - other person's deviation from mean
0.6020, 0.0083 - other person's 2 class interest rating

My stress and activity only
* - my actual rating
0.5596, 0.00008 - my mean rating, denotes personality + compatibility + mood that day

* 0.3950, 0.0939 - my deviation from mean
* - my 2 class interest rating
0.4766, 0.0166 - other person's actual rating
0.4797, 0.01 - other person's mean rating, denotes personality + compatibility + mood
0.4412, 0.03 - other person's deviation from mean
0.5912, 0.00002 - other person's 2 class interest rating

Activity only
0.5210, 0.0079 - my actual rating
* 0.41, 0.12 - my mean rating, denotes personality + compatibility + mood that day
0.4838, 0.0232 - my deviation from mean
0.5421, 0.041 - my 2 class interest rating
0.5009, 0.0145 - other person's actual rating
* - other person's mean rating, denotes personality + compatibility + mood that day
0.4864, 0.0217 - other person's deviation from mean
0.5478, 0.0033 - other person's 2 class interest rating

Stress only
* - my actual rating
0.3786, 0.0702 - my mean rating, denotes personality + compatibility + mood that day
* - my deviation from mean
0.3950, 0.0499 - my 2 class interest rating
* - other person's actual rating
* - other person's mean rating, denotes personality + compatibility + mood that day
* - other person's deviation from mean
* - other person's 2 class interest rating

Engagement, no significant correlations

Empathy, no significant correlations

**Women**

All features,
0.6733, 0.0064 - my actual rating
0.6309, 0.0301 - my mean rating, denotes personality + compatibility + mood that day
* 0.5533, 0.2 - my deviation from mean
0.6560, 0.0126 - my 2 class interest rating
0.6801, 0.0048 - other person's actual rating
0.6437, 0.0197 - other person's mean rating, denotes personality + compatibility + mood
* 0.5597, 0.1829 - other person's deviation from mean
0.7023, 0.0017 - other person's 2 class interest rating


My features only
0.5744, 0.0047 - my actual rating
0.5851, 0.0031 - my mean rating, denotes personality + compatibility + mood that day
* - my deviation from mean
0.5553, 0.0093 - my 2 class interest rating
0.5361, 0.0172 - other person's actual rating
0.4779, 0.0807 - other person's mean rating, denotes personality + compatibility + mood
* 0.4166, 0.2542 - other person's deviation from mean
0.6017, 0.0016 - other person's 2 class interest rating

Other person's features only
0.5744, 0.0047 - my actual rating
0.5851, 0.0031 - my mean rating, denotes personality + compatibility + mood that day
* 0.4249, 0.2234 - my deviation from mean
0.5553, 0.0093 - my 2 class interest rating
0.5361, 0.0172 - other person's actual rating
0.4779, 0.0807 - other person's mean rating, denotes personality + compatibility + mood
* 0.4166, 0.2542 - other person's deviation from mean
0.6017, 0.0016 - other person's 2 class interest rating


Stress and activity
0.5960, 0.004 - my actual rating
0.5696, 0.0105 - my mean rating, denotes personality + compatibility + mood that day
*   - her deviation from mean
0.5909, 0.0049 - my 2 class interest rating
0.6211, 0.0014 - other person's actual rating
0.5691, 0.0107 - other person's mean rating, denotes personality + compatibility + mood
0.4923, 0.0905 - other person's deviation from mean
0.6219, 0.0014 - other person's 2 class interest rating

My stress and activity only

0.4795, 0.0067 - my actual rating
0.5273, 0.0012 - my mean rating, denotes personality + compatibility + mood that day
\*   - her deviation from mean
0.4672, 0.0098 - my 2 class interest rating
0.4212, 0.0348 - other person's actual rating
\* - other person's mean rating, denotes personality + compatibility + mood
\* 0.38, 0.09 - other person's deviation from mean
0.4982, 0.0036 - other person's 2 class interest rating

Engagement
0.4615, 0.005 - my actual rating
0.3623, 0.0685 - my mean rating, denotes personality + compatibility + mood that day
0.3945, 0.0330 - my deviation from mean
0.4341, 0.0116 - my 2 class interest rating
0.4759, 0.0031 - other person's actual rating
0.3815, 0.0449 - other person's mean rating, denotes personality + compatibility + mood
0.3258, 0.1383 -  other person's deviation from mean
0.4742, 0.0033 - other person's 2 class interest rating

Empathy, no significant correlations

Please list 10 generic topics you find interesting to talk about. These could be from current events, arts and science, politics or your work.

| Conversation # | Topic |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

**Please rate each 3-minute conversation a scale of 1 to 10.**

How interesting did you find talking about this particular topic? (1 not interesting, 10 extremely interesting)

| Conversation # | Rating |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

How much did you agree with the other person's opinion? (1 disagreed completely, 10 agreed completely)

| Conversation # | Rating |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |

# Appendix B

# Supporting Material for Attraction Experiment

## B.1  Anonymized Raw Data

Please visit our website groupmedia.media.mit.edu or email anmol@media.mit.edu

## B.2  Regression Tables

| Question | r value | p value |
|---|---|---|
| Asked of females: | | |
| Are you interested in this person? | 0.673 | 0.05 |
| | 0.663 | 0.01 |
| Are you interested in being friends? | 0.6484 | 0.04 |
| | 0.6340 | 0.01 |
| Are you interested in a business/ professional relationship? | 0.7066 | 0.03 |
| | 0.7066 | 0.01 |
| | 0.6946 | 0.001 |
| Asked of males: | | |
| Are you interested in this person? | 0.6071 | 0.05 |
| | 0.5886 | 0.01 |
| Are you interested in being friends? | 0.6284 | 0.04 |
| | 0.6155 | 0.01 |
| Are you interested in a business/professional relationship? | 0.6035 | 0.05 |
| | 0.5699 | 0.01 |

Features:
4 - Female fraction of speaking time
5 - Female std deviation of energy
6 - Female std deviation of format frequency (stress)
7 - Female std deviation. of  spectral entropy (stress)
8 - Female voiced rate
9 - Male fraction of speaking time
10 - Male std deviation of energy
11 - Male std deviation of format frequency
12 - Male std deviation. of  spectral entropy
13 - Male voiced rate
14, 15 – Female influence parameters (on female, on male)
16,17 – Male influence parameters (on female, on male)
18,19 - Female and male back-and-forth counts

# Appendix C

# Supporting Material for Focus Groups Experiment

## C.1  Anonymized Raw Data

Please visit our website groupmedia.media.mit.edu or email anmol@media.mit.edu

## C.2  Feedback Form

# Please Choose one of these as your class paper topic :

(1) Richard Hackman's - scheme for organizing groups
(2) Carl Marci -- 'concordance', focus groups;
(3) Eagle & Pentland -- peoples' position and proximity, social networks
(4) Caneel, Madan, & Pentland -- tone of voice, social intelligence
(5) Huberman -- mining email. corporate knowledge management
(6) Hughes -- recording and indexing meetings and interactions
(7) Fernandez -- social capital
(8) Ambady -- thin slices of behavior

**Now please choose as your \*\*final\*\* class paper topic. In order to change your selection after this, you will have to email Joost Bonsen, class TA.**

(1) Richard Hackman's - scheme for organizing groups
(2) Carl Marci -- 'concordance', focus groups;
(3) Eagle & Pentland -- peoples' position and proximity, social networks
(4) Caneel, Madan, & Pentland -- tone of voice, social intelligence
(5) Huberman -- mining email. corporate knowledge management
(6) Hughes -- recording and indexing meetings and interactions
(7) Fernandez -- social capital
(8) Ambady --  thin slices of behavior

# Bibliography

[1] Ambady N., Rosenthal R *Thin Slices of Expressive Behaviour as Predictors of Inter-personal Consequences : A Meta Analysis* PhD Thesis Harvard University (1992)

[2] Aoki P., Woodruff A., *"User Interfaces" and the Social Negotiation of Availability*

[3] Briol, Petty *Overt Head Movements and Persuation: A self-validation and analysis*, Journal of Personality and Social Psycology 76 (1999)

[4] Burgoon J.,Saine T.,*The Unspoken Dialogue : An Introduction to Non-verbal Communication* Houghton (1978)

[5] Business Insight (formerly Reuters), consumer goods report 2004 - The failure rate is highest at about 90

[6] Breazeal C., *Designing Social Robots* MIT Cambridge Press (2001)

[7] Basu, B., (2002) Conversational Scene Analysis, doctoral thesis, Dept. of Electrical Engineering and Computer Science, MIT. 2002. Advisor: A. Pentland

[8] Batliner A., Fisher K., Huber R., Spilker J.,*Desperately seeking emotions: Actors wizards and human beings*, in Proc. of ISCA workshop on speech and emotion, ISCA (2000)

[9] BodyMedia devices, www.sensewear.com

[10] Bluetooth Acclerometer, CCG Georgia Tech http://www.cc.gatech.edu/ccg/resources/btacc

[11] Bernard, H.R., P. Killworth, D. Kronenfeld, and L. Sailer *The Problem of Informant Accuracy: The Validity of Retrospective Data* Annual Review of Anthropology 13:495-517 (1984)

[12] Chartrand T., Bargh J.,*The Chameleon Effect: The perception behaviour link and social interaction* Journal of Personality and Social Psycology, 76 (1999)

[13] Choudhary T. *Sensing and Modelling Human Networks*, PhD Thesis, Dept. of MAS, MIT (2003)

[14] Caneel R., Madan A., *Real Time Speech Feature Calculation Software for the Zaurus*, distributed as open source under the GPL licence off our website http://www.media.mit.edu/wearables

[15] cowie R., Douglas-Cowie E., Tsapatsoulis N.,Votsis G., *Emotion Recognition in human-computer interaction*, IEEE Signal Processing Magazine, vol 18(1) pp32-80 (2001)

[16] Contexual Computing Group, Georgia Tech http://www.cc.gatech.edu/ccg/

[17] Csikszentmihalyi M., *Flow: The Psychology of Optimal Experience*, Perennial; Rep edition (1991)

[18] Dallaert F., Polzin T., *Recognizing Emotions in Speech* in Proc. 4th ICSLP, pp 1970-1973 IEEE (1996)

[19] Dunbar, R. *Grooming, Gossip, and the Evolution of Language* Harvard Univ Press (1998)

[20] DeVaul R., Sung M., Gips J., *MiThril 2003: Architecure and Applications* in Proc of ISWC 2003 White Plains NY (2003)

[21] Eagle N., Pentland A., *Reality Mining: Sensing Complex Social Systems*, appears in Journal of Personal and Ubiquitous Computing (2005)

[22] Eagle N., Singh P., Morgan B., *Diary application* available at http://xnet.media.mit.edu/life/soothsayer/Soothsayer.swf

[23] ESOMAR Industry Report 2005 (www.esomar.com). ESOMAR is the world's leading market research association and is recognized as the premier international industry advocate (2005)

[24] Wearable Computing Group, ETH Zurich http://www.wearable.ethz.ch/

[25] Fuller G.D, *GSR History and Physiology* Biofeedback Methods and Procedures in Clinical Practice

[26] Fernandez, R., *A Computational Model for the Automatic Recognition of Affect in Speech* Ph.D. Thesis, MIT Media Arts and Science, February 2004. Advisor: R. Picard (2004)

[27] Gips J, Dong W, Pentland A. *No More Business Cards - Automatic Measurement of Interest at Conferences* MIT Media Lab Technote 586, available at hd.media.mit.edu

[28] Gray, P. *Engaging students' intellects: The immersion approach to critical thinking in psychological instruction* Teaching of Psychology, 20, 68-74 (1993)

[29] Geis S.,Fuller R., Rush J., *Lovers of AIDS patients : psychosocial stresses and counselling needs*

[30] Gladwell M., *Blink: The power of thinking without thinking*, Little Brown (2005)

[31] Healey J., Picard R. *StartleCam: A Cybernetic Wearable Camera*, proceedings of ICWC (1998)

[32] Henley N.M. *Body Politics: Power, Sex and Non-verbal communication: A Perceptual Study* Springer-Verlag (1977)

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

[34] Johnston P., VP Marcus Thomas LLC, quoted in *American Demographics: The New Science of Focus Groups* March issue (2003)

[35] Kapoor A.,Picard R,*A Real Time Head Nod and Shake Detector*, appears in IEEE PUI Orlando (2001)

[36] Kimbara I., Parrill F.,*Does Mirroring in Speech and Gesture Lead to Perception of Greater Rapport?* University of Chicago (http://home.uchicago.edu/ feyparr/research/papers/mirror.pdf)

[37] Knapp M.L. *Non-Verbal communication and Human Interaction* New York, Holt (1978)

[38] Kitzinger J., *Qualitative Research: Introducing focus groups*. British Medical Journal available at www.bmj.com

[39] Kurzban R. *HurryDate : Mate preferences in Action*, to appear in Evolution and Human Behaviour (2005)

[40] Leathers D.G. *Successful Non-Verbal behavior: Principles and Applications, 3rd ed* MA (1997)

[41] Larsen R.J., Diener E. *Promises and Problems with the circumplex model of emotion* In M. Clark (ed), Review of Personality and Social Psycology 13 pp. 25-59 (1992)

[42] Vivometrics Lifeshirt, www.vivometrics.com

[43] Marci C. *Physiological Evidence for the Interpersonal Role of Laughter During Psychotherapy* Journal of Nervous and Mental diseases, Oct (2004)

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

[45] Oguchi T., Kikuchi H., *Voice and Interpersonal Attraction* Japanese Psychological Research Volume 39 (1) pp.56 (1997)

[46] O'Shaughnessy, *Speech Communications, Human and machine* 2nd ed. MIT Press (2002)

[47] Oudeyer, P-Y. *The production and recognition of emotions in speech: features and algorithms* International Journal of Human Computer Interaction, 59(1-2):157–183 (2003)

[48] Pentland A., Curhan J., Khilnani R., Martin M., Eagle N., Caneel R., Madan A.,*A Negotiation Advisor* MIT Media Lab Technote 577, appears in UIST Santa Fe (2004)

[49] Pentland, A. (2004) Social Dynamics: Signals and Behavior, ICDL, San Diego, CA Oct 20-23, IEEE Press

[50] Picard R., *Affective Computing*, MIT Press (1996)

[51] Patnaik D., managing associate at Jump Associates, quoted in *American Demographics: The New Science of Focus Groups* March issue (2003)

[52] Reynolds C., Strauss M., *Bluetooth GSR sensor*, Affective computing group, MIT Media Lab http://arsenal.media.mit.edu/notebook/

[53] Sipe L., *Talking back and taking over: Young children's expressive engagement during storybook read-alouds*, The Reading Teacher, Vol.55 -5 (2002)

[54] Schroder M., Cowie R., Douglas-Cowie E.,Westerdijik M.,, *Accoustic correlates of emotion dimensions in view of speech synthesis*, in Poc. 7th EUROSPEECH, ISCA (2001)

[55] Terry M.,Mynatt E.,Ryall K., Darren L.,*Social Net: Using Patterns of Physical Proximity to Infer Shared Interests*, CHI MI (2002)

[56] E. Vyzas (1999), *Recognition of Emotion and Cognitive States Using Physiological Data* Mechanical Engineer's Degree Thesis, MIT, June 1999: Advisor Dr. R. Picard

[57] Valbonesi L., Ansari R., McNeill D., et. al. *Multimodal Signal Analysis of Prosody and Hand Motion: Temporal Correlation of Speech and Gestures* EUSIPCO 2002 Conference, Tolouse, France, (2002)

[58] Woodruff A., Aoki P., *How Push-to-Talk makes Talk Les Pushy* Proc. GROUP '03 ACM Florida pp. 170-179 (2003)

[59] Yu C.,aoki P.,Woodruff A., *Detecting User Engagement in Everyday Conversation,* draft in publication.

[60] Zaltman G., *How Customers Think: Essential Insights into the Mind of the Market,* HBS Press (2003)

[61] G. Zhou, J.H.L. Hansen, J.F. Kaiser, *Linear and Non-linear Speech feature analysis for Stress classification* ICSLP Volume 3 (1998)

[62] G. Zhou, J.H.L. Hansen, J.F. Kaiser, *Methods for Stressed Speech Classification: Nonlinear TEO and Linear Speech Based Features* IEEE ICASSP-99: Inter. Conf. on Acoustics, Speech, and Signal Processing, vol. 4, pp. 2087-2090, Phoenix, Arizona. (1999)