Toward a Social Signaling Framework: Activity and Emphasis in Speech

by

William T. Stoltzman

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author .....
Department of Electrical Engineering and Computer Science
August 11, 2006

Certified by .....
Alex P. Pentland
Toshiba Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Toward a Social Signaling Framework:

Activity and Emphasis in Speech

by

William T. Stoltzman

Submitted to the Department of Electrical Engineering and Computer Science
on August 11, 2006, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Language is not the only form of verbal communication. Loudness, pitch, speaking rate, and other non-linguistic speech features are crucial aspects of human spoken interaction. In this thesis, we separate these speech features into two categories—vocal Activity and vocal Emphasis—and propose a framework for classifying high-level social behavior according to those metrics.

We present experiments showing that non-linguistic speech analysis alone can account for appreciable portions of social phenomena. We report statistically significant results in measuring the persuasiveness of pitches, the effectiveness of customer service representatives, and the severity of depression. Effect sizes of these studies explain up to 60% of the sample variances and yield binary decision accuracies nearing 90%.

Thesis Supervisor: Alex P. Pentland
Title: Toshiba Professor of Media Arts and Sciences

Acknowledgments

First and foremost, I thank my ever-wise advisor, Sandy, who has inspired me more times than I can count. He was instrumental in getting my experiments off the ground and giving them the momentum they needed to succeed. Invaluable too were the Pentlandians, past and present, who showed me the ropes and were always available to chat or brainstorm. I owe a particular debt to Anmol Madan, whose work in the group helped me to formulate a complete framework.

In the design and execution of the experiments covered in this thesis, I received help from many individuals and institutions.

In the persuasion experiment, I thank Pat Bentley for her involvement in conceiving and organizing the study. Many thanks also to all the participants who donned a microphone and donated their time and voice to my research. After the analysis was complete, I enjoyed working with Ben Waber on a speech synthesis mini-project that I cite in this thesis. Finally, thanks to Gene Pettinelli and Victor Grau Serrat at CambridgeLight Partners for giving me the opportunity to expand my research.

In the call center experiment, I thank Vertex for allowing us to collaborate with them and collect real, usable data, and Dr. Marco Busi and Laura Dingwall for their organizational efforts. Special thanks are owed to Stefan Agamanolis who worked tirelessly to record many, many customer service calls.

In the depression experiment, I am indebted to the Medical Information Systems Unit at the Boston University Medical Campus/Boston Medical Center, which furnished all the data for the experiment. Thanks to Ramesh Farzanfar, Ph.D., Robert Friedman, M.D., and Edward Goldstein.

Last, but certainly not least, I thank my family and friends, without whose support and positive influence I'd be lost.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Social interaction has commonly been addressed within two different frameworks [27]. One framework comes from cognitive psychology and focuses on emotion. Ekman and Friesen [15] are the most well-known advocates of this approach, which is based roughly on the theory that people perceive others' emotions through stereotyped displays of facial expression, tone of voice, etc. The simplicity and perceptual grounding of this theory has recently given rise to considerable interest in the computational literature [28]. However, serious questions about this framework remain, including the question of what counts as affect? Does it include cognitive constructs such as interest or curiosity, or just the base dimensions of positive/negative, active/passive? Another difficulty is the complex connection between affect and behavior: adults are skilled at hiding emotions, and seemingly identical behaviors may have different emotional roots.

The second framework for understanding social interaction comes from linguistics, and treats social interaction from the viewpoint of dialog understanding. Kendon et al. [21] and Argyle [6] are among the best known pioneers in this area, and the potential to greatly increase the realism of humanoid computer agents has generated considerable interest from the human-computer interaction community [11]. In this framework, prosody and gesture are treated as annotations of the basic linguistic information, used (for instance) to guide attention and signal irony. At the level of dialog structure, there are linguistic strategies to indicate trust, credibility, etc., such

as small talk and choice of vocabulary. While this framework has proven useful for conscious language production, it has been difficult to apply it to dialog interpretation, perception, and for unconscious behaviors generally.

In this thesis, we expand upon a new conceptual framework introduced by Pentland, which focuses on social signaling of speaker attitude or intention through the amplitude, frequency, and timing of prosodic and gestural activities [27]. This framework is based on the literature of personality and social psychology, and is different from the linguistic framework in that it consists of *non-linguistic*, largely *unconscious* signals about the social situation, and different from the affect framework in that it communicates *social relation*, rather than speaker emotion.

It is different in another way as well: it happens over longer time frames than typical linguistic phenomena or emotional displays. It treats speech and gestures more like a texture than individual actions, and it appears to form a largely independent channel of communication. In the language of the affect framework, these signals are sometimes identified by the oxymoronic label 'cognitive affect,' whereas in the linguistic framework they might be related to dialog goals or intentions.

1.1 Approach

Our approach will be to analyze social interaction through speaking patterns. Our goal is to establish a framework for understanding these signals that will be both *automated*—able to run on a computer without human parsing, labeling, etc.—and *universal*—without the need to be trained for particular individuals. Toward this goal, we define two different social signals called Activity and Emphasis that offer high (and independent) explanatory power over speakers' intentions and attitudes. We then discuss the design and execution of a number of experiments and illustrate how a simple framework can unify the experimental outcomes.

1.2 Outline

The remainder of this thesis is organized as follows:

- In chapter 2, we provide background on social signaling, speech, and prosody. We introduce our speech analysis platform and explain our statistical methods.
- In chapter 3, we define Activity and Emphasis and propose a framework for using those as measures of various behavioral states.
- In chapter 4, we describe the setup and execution of an experiment to measure persuasiveness in speech.
- In chapter 5, we apply our speech processing techniques to gauge which factors contribute to success in the service and sales industry.
- In chapter 6, we show that depression can be readily correlated to speaking patterns, leading to the notion that our platform could be used as a clinical monitoring tool for mental health.
- In chapter 7, we round out our proposed framework by citing studies done in the fields of interest and attraction.
- We conclude and summarize our contributions in chapter 8.

Chapter 2

Background

2.1 Social Signaling and Thin-Slicing

Social signaling is what you perceive when observing a conversation in an unfamiliar language, and yet find that you can still 'see' someone taking charge of a conversation, establishing a friendly interaction, or expressing empathy for the other party [19]. While you cannot understand the words being spoken, you can still interpret and understand the prototypical (and often unconscious) behaviors that humans have evolved to display.

Research in social signaling gained momentum with the studies of Ambady and Rosenthal [3]. They are credited with developing the concept of 'thin-slicing,' which was later popularized in Malcolm Gladwell's bestselling book *Blink* [19]. Gladwell defines thin-slicing as "the ability of our unconscious to find patterns in situations and people based on very narrow slices of experience" (p. 23). These 'slices'—running the gamut from brief snippets of audio to facial expressions to a walk through someone's bedroom whom you haven't met—can predict various social phenomena with surprising accuracy.

For example, Ambady et al. published startling results where ordinary people were able to predict whether surgeons would be sued for malpractice, *just by listening to 20-second sound clips of doctor-patient conversations.* Even more surprising was that the raters were not basing their judgments on *what* was being said. The experiment

was performed a second time, where the content was filtered out of the audio signal, leaving only intonation, pitch, and rhythm.[1] In this seemingly restricted task, raters performed just as well as before [4]. That is the power of thin-slicing.

More recently, Pentland has shown that computer software is well suited to thin-slicing as well. Where humans are capable of around 70% binary decision accuracy, computers average about 80% accuracy in tasks such as predicting salary from a sound file of the first 5 minutes of a negotiation [26] or predicting who will exchange business cards at a meeting based on speech and motion patterns [18].

2.2 Speech and Prosody

The ability to speak is a uniquely human quality which pervades our society. Speech as a social signal is important to study because it is so common as a means of communication, from face-to-face interaction to public address to phone conversations.

Physiologically, speech originates with airflow at the glottis and a possible noise source from the vocal folds and is then filtered by the vocal tract (larynx, pharynx, oral cavity, nasal cavity, lips; see figure 2-1). The vocal folds may vibrate hundreds of times per second (corresponding to voice pitch), and in running speech, the vocal tract constantly changes shape to produce the different speech sounds. As a reference point, mean syllable durations in read script are in the range of 200–250 ms [31].

While language is a critical component of speech, it is indisputable that verbal communication does not end with syntactic and semantic content. Often, *how* something is said holds as much importance as *what* is actually said. To take a simple example, consider the phrase "I'm excited." Interpreted literally, its meaning is quite clear. This meaning can be reinforced by the speaker through energy and excitement—"I'm excited!!"—or further qualified through points of emphasis—"*I'm* excited." The speaker can even contradict the meaning of the words by speaking

[1] This may be done by low-pass filtering the sound signal with a frequency cutoff around 400 Hz, thereby removing the formants and noise bursts (generally between 500 and 5000 Hz) that make speech intelligible.

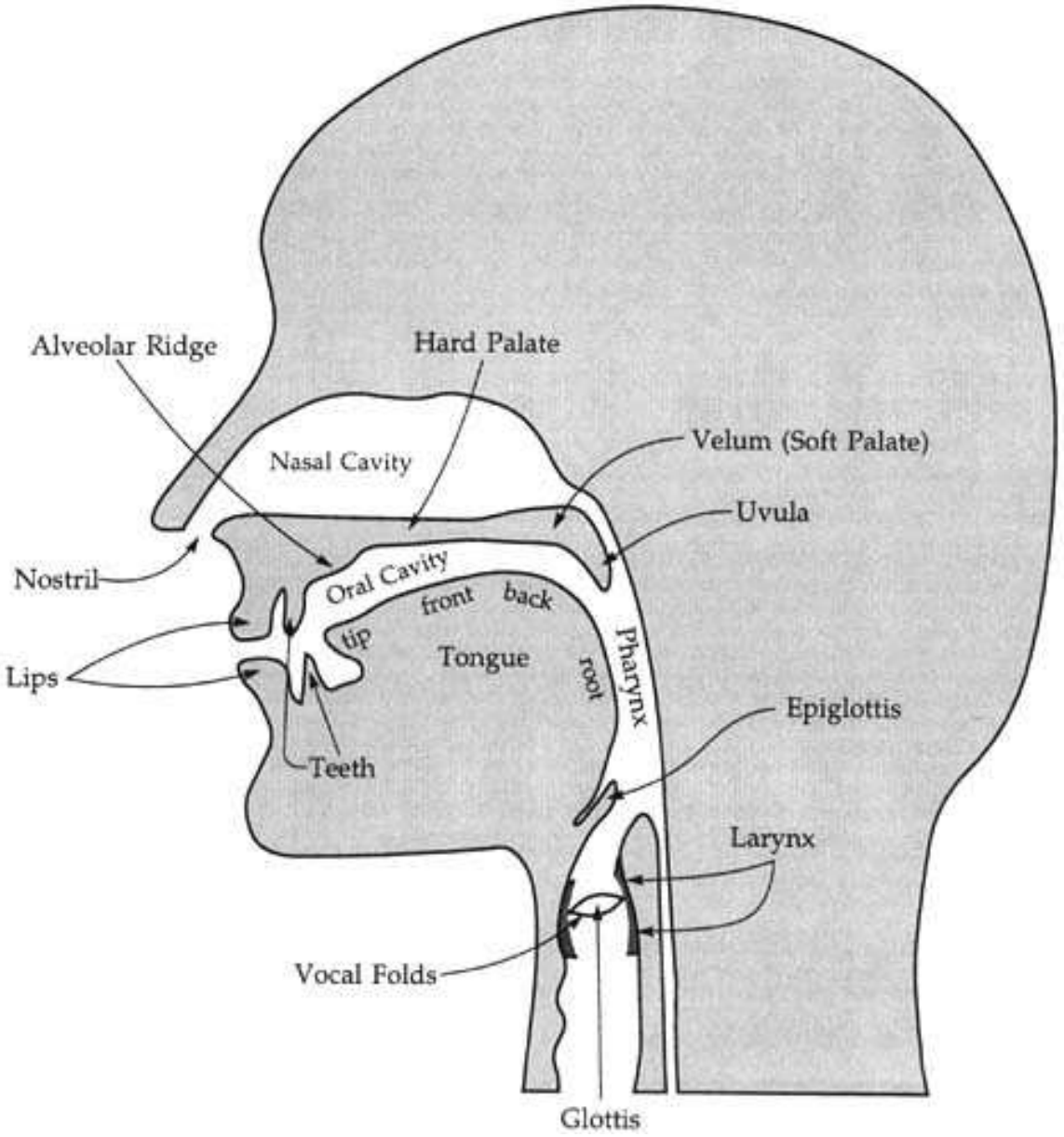

Figure 2-1: The human vocal tract.

sarcastically.

The non-linguistic cues that a speaker uses to guide listeners and signal intent are collectively called *prosody*. Prosody includes such factors as voice pitch, pacing, and loudness and may occur consciously or unconsciously. The sophisticated verbal and vocal channels are so well integrated in humans that they seldom provide inconsistent information to listeners [25]. Frick reports studies showing that, in prosodic communication of emotion, there is little evidence for either personal idiosyncrasies or cultural differences [17], suggesting that prosodic signals are *evolved patterns*, rather than *learned conventions*.

2.3 Speech Analysis Platform

Our speech analysis platform is built around measuring the prosodic features of speech. One key advantage of this approach is that this sort of analysis is fast and efficient, making it computationally feasible on resource-limited platforms, such as cell phones and other embedded devices.

All the speech features we employ are based on *voiced* speech, which are segments of speech whose spectra show strong harmonic structure. This occurs when the vocal folds are vibrating periodically and the vocal tract is unobstructed; essentially, these are vowels. *Unvoiced* speech (consonants) can appear on either side of voiced segments to form syllables.

2.3.1 Speech Features

We begin by extracting a basic set of speech features from audio sampled at 8000 Hz. The processing is done with a 256 sample window (32 ms) and 128 sample step size (16 ms). We consider the following features:

- $\mathbf{f_0}$ – The fundamental frequency. Essentially, the pitch of the voice. In adults, f_0 is generally between 90 and 300 Hz, with females typically higher in the range than males [31].

- **Spectral entropy** – Measure of the randomness of the segment in the frequency domain.
- **Spectral autocorrelation** – Autocorrelation of the Fourier Transform of the window. A voiced segment will exhibit strong peaks in this signal due to its periodicity.
- **Energy** – The volume (loudness) of a segment.
- **d/dt Energy** – The time-derivative of volume.

Figure 2-2 shows a spectrogram of 9 seconds of female speech with f_0 and energy overlaid.

Next, we apply speech analysis techniques described in Basu [8] to determine which segments are voiced, and how those segments can be grouped together to constitute a phrase, or a 'speaking' segment. We take this approach because it is robust to low sampling rates, far-field microphones, and ambient noise, all of which can plague real-world situations.

Using the raw features from above, we employ a two-level hidden Markov model (HMM) to identify voiced segments (where the vocal folds are vibrating, as in a vowel sound) and group them into speaking regions. See figure 2-3. After performing this analysis, we calculate the following features over a desired time period, often five minutes:

- **Length of a voiced segment** – The duration of a sonorant (vowel) sound. Essentially the duration of each syllable of speech.
- **Length of a speaking segment** – The duration of a phrase, as decided by the voiced/speaking HMM.
- **Fraction of time speaking** – Percentage of the total pitch time taken up by phrases.
- **Voicing rate** (also referred to as **speaking rate** or **speech rate**) – The number of voiced segments (essentially, syllables) per unit time. The voicing rate is only

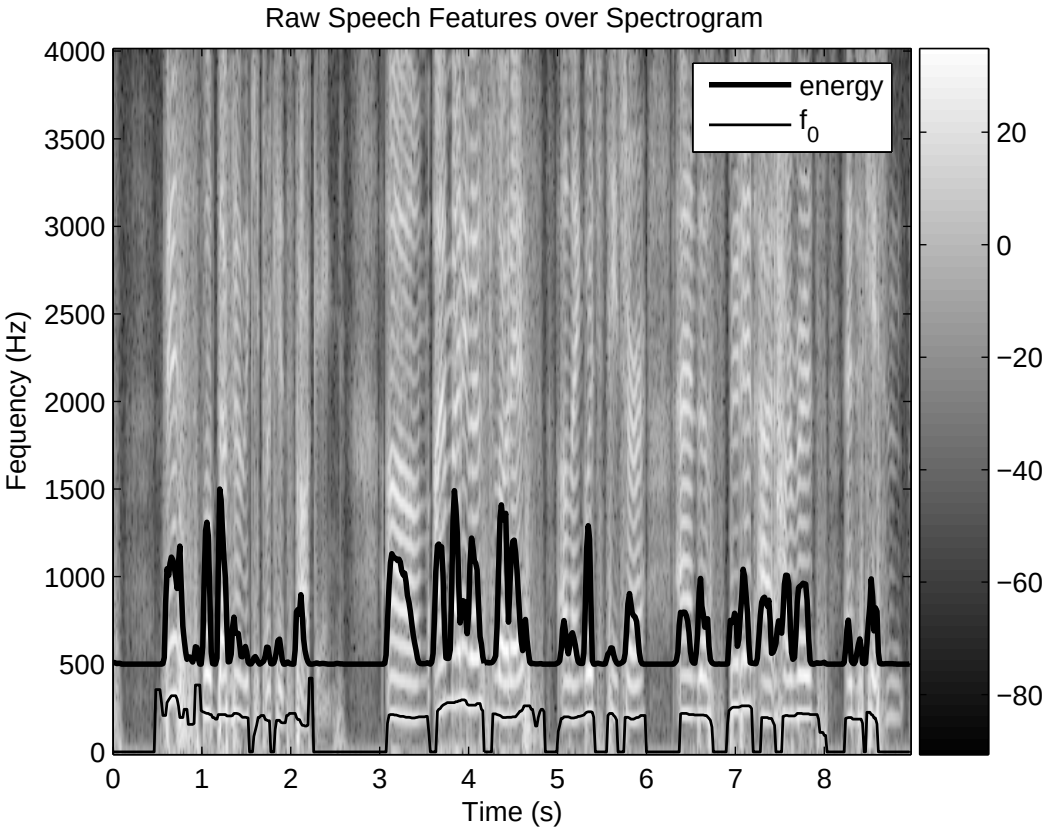

Figure 2-2: Some raw speech features plotted over a spectrogram of a female speaker. f_0 is shown tracking the fundamental frequency, and energy is superimposed (with no relation to frequency).

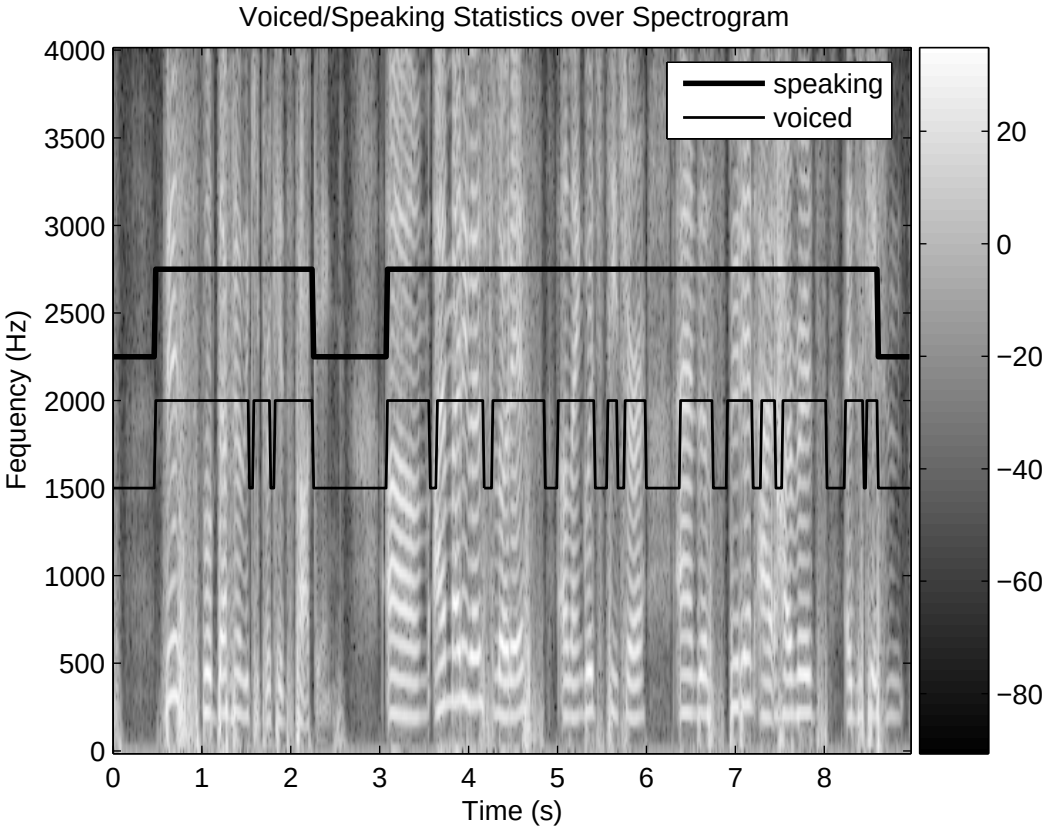

Figure 2-3: Voiced and speaking segments plotted over a spectrogram of a female speaker. Voiced segments exhibit strong harmonics and can be visually identified by areas having distinct bands. High portions of the plots represent voiced or speaking segments. (There is no relation of these two plots to frequency.) A two-level HMM is used to determine the voiced and speaking segments.

calculated over speaking segments, i.e. pausing between phrases does not affect this number.

- **Entropy of length of speaking segments** – Measurement of the randomness in the lengths of phrases.
- **Entropy of length of pauses** – Measurement of the randomness in the lengths of pauses between phrases.

Where appropriate, we calculate the following feature over the entire signal:

- **Duration** – Duration of the session in seconds.

Note that from this collection of speech features alone, it is essentially impossible to recover the actual words spoken, thereby mitigating most privacy or intellectual property concerns.

For further information about obtaining and using our speech analysis platform, see appendix A.

2.4 Statistical Analysis

In this framework, amplitude, frequency, and timing of prosodic activities are shown to correlate to speaker attitude and intention. Many of the relationships we observe are linear, and so we choose to describe them with linear models obtained through single- or multiple-variable regressions. For trends that are not so readily modeled, we may use simple quadratic classifiers that are best matched to the second moment statistics (i.e., the mean and covariance) of the feature vectors.

For regressions, we report the coefficients for each of our explanatory variables. To avoid clutter, we omit the constant term of the regression line; this can be easily recovered. See DeGroot and Schervish [14] for reference.

For quadratic classifiers, we report the mean and covariance matrices that completely specify the classifier. See Therrien [33] for details.

We assess the strength of our models by reporting standard statistics:

- r – The *correlation coefficient*, ranging from -1 to 1, measures the strength of a linear relationship. Values close to 1 indicate strong positive relationship, values close to -1 indicate strong negative relationship, and values near 0 indicate weak (linear) relationship.

- r^2 – This value, ranging from 0 to 1, is used in multiple-variable regressions to give the *proportion of variance explained* by the fitted regression. The closer r^2 is to 1, the smaller the sample variation around the regression line is compared to the variation around the sample mean.

- p – The p-value is the probability that we could observe a particular r or r^2 if there were no true underlying relationship. Smaller p-values indicate more *statistically significant* findings.

- Binary decision accuracy – This number measures the strength of a model attempting to separate sample data into two mutually exclusive, collectively exhaustive classes (e.g. success *vs.* failure). Reported as a percentage, it tells what *proportion of the observations were correctly classified.* The incorrect classifications are then divided into *false positives* (failure classified as success) and *false negatives* (success classified as failure).

We are now left with the question of how to assess "goodness" of the measured values of these statistics. Unfortunately, this is a somewhat arbitrary procedure, very much dependent upon context. For instance, the value $r = .9$ may be unacceptably small in a well-controlled physical experiment, but that same r could be quite large in a broad anthropological study. Cohen provides some guidelines for correlation in the behavioral sciences (which encompass the research presented in this thesis): a "medium effect size" is merited when $r > .30$ ($r^2 > .09$), and a "large effect size" is taken to be $r > .50$ ($r^2 > .25$) [12]. In terms of statistical significance, common cutoffs for accepting a relationship are $p < .05$ and $p < .01$ [14].

Chapter 3

Framework

3.1 Activity and Emphasis

Recently, Pentland constructed measures for four types of social signaling: Activity, Stress (later renamed to Emphasis), Engagement, and Mirroring [27]. These measures were extrapolated from a broad reading of the voice analysis and social science literature, and have been generally established as valid [10] [22] [26]. The first two measures—Activity and Emphasis—are the basis for our social signaling framework. We forgo Engagement and Mirroring in this framework, because both these measures deal specifically with dyadic interaction, while we choose to focus on singular signaling. Previous studies have shown the contribution of Activity and Emphasis to substantially outweigh that of the other two features in predicting certain social phenomena [22]. Our findings about Activity and Emphasis should hold true whether a speaker is addressing one, many, or unspecified numbers of people.

In this thesis, we use the following definitions:

- **Vocal Activity** is a combination of the percentage of speaking time, length of phrases, and speed of speech production.
- **Vocal Emphasis** is a combination of the variations in loudness, pitch, and spectral entropy.

We hypothesize that Activity is manifest when a speaker is outwardly projecting

and is in a state of social *offering*. Emphasis will occur primarily when a speaker is signaling openness—to comment or new information, for example—and is in a state of social *invitation*. Both Activity and Emphasis can vary on a spectrum from low to high, and, more importantly, can vary independently.

3.2 Relating Speech Prosody to Social Signals

We propose that high or low amounts of Activity and Emphasis correspond to four representative behavioral states: Persuasion, Service, Depression, and Interest; see table 3.1. Persuasiveness, as when delivering a pitch, is primarily a result of high Activity—a persuader must offer information and project enthusiasm. On the other hand, those in the service sector, particularly customer service, find success by inviting customer input, a hallmark of high Emphasis, while at the same exhibiting low Activity to avoid overwhelming the customer. Depressed individuals, not surprisingly, exhibit neither Activity nor Emphasis—neither offering nor inviting interaction. And, finally, people who find themselves interested in or attracted to a subject show high levels of both Activity and Emphasis, indicating both a desire to make conversation and a willingness to listen.

		Emphasis	
		low	high
Activity	low	Depression	Service
	high	Persuasion	Attraction

Table 3.1: Quadrants of the Activity/Emphasis framework.

We devote the next four chapters of this thesis to show strong evidence supporting our proposed framework with experiments designed to examine each of the Activity/Emphasis quadrants.

Chapter 4

Persuasion — High Activity, Low Emphasis

In this chapter, we describe an experiment to quantify persuasiveness by looking only at speaking patterns. We develop an automated speech analysis program called The ElevatorRater to analyze "elevator pitches" and, more generally, any brief (less than 5 minutes) presentation meant to persuade. Elevator pitches are short, spoken overviews of an idea or product, intended to elicit the interest and support of the listener. Everyone has an elevator pitch of his or her own—think about how you might introduce yourself and your interests to a new acquaintance—but these mini-speeches are particularly prevalent in the business and entrepreneurial world, where an individual may need to propose an idea in a very constrained amount of time (ostensibly, during an elevator ride with an executive).

It is tempting to think that good content will ensure positive reception of a pitch. After all, the content conveys the message. But it turns out that other factors—ones we might not even be consciously aware of—play a surprisingly large role in how listeners perceive a speaker. In fact over 35% of the sample variance for persuasion in our study was accounted for by just looking at a selection of non-linguistic speech features.

We show that perceived persuasiveness generally rises in speakers exhibiting high amounts of Activity but low amounts of Emphasis, i.e. delivering much information

in a given amount of time but with well-regulated volume dynamics.

We show that gender—of both the speaker and listener—plays a significant role. In terms of delivering a pitch, men and women exhibit different speech feature profiles. In terms of receiving a pitch, we find that the perception of the listener is on average influenced by his or her gender and that of the speaker.

Finally, while we might like to believe that elevator pitches are judged solely on content, we report experiments showing that ratings of persuasiveness are deeply confused with speaking style.

4.1 Background

While humans have no difficulty identifying persuasive speech, little previous work has been done to qualify the traits and patterns responsible. Indeed, building a comprehensive model is daunting, as it is reasonable to believe that one's ability to persuade is based on a large number of factors, such as speaking style, voice quality, word choice, preparation, physical appearance, dress, etc. Much of this information, however, is redundant, and recent studies show that a number of prosodic speech features correlate highly with charisma and persuasion [29].

4.2 Elevator Pitch Experiment

4.2.1 Data Collection

We collected the data during several identical sessions where students (mostly MBAs from MIT's Sloan School of Management) volunteered to gather in small groups to practice public speaking. In each session, participants were asked to give a short, prepared pitch on a topic of their choice. With the speakers' written consent, we recorded their pitches using a headset microphone. Pitches ranged in length from 30 seconds to 5 minutes and all were recorded at 8 kHz. A large number of pitches were centered around funding requests for new technology (in the true spirit of an elevator pitch), but topics varied widely and included themes as diverse as thesis

proposal outlines, the similarities between sharks and humans, and the future of wearable computing. Each session consisted of roughly 10 participants, and in total, we recorded and processed 42 pitches (20 male and 22 female).

After each speaker presented a pitch—but before any group feedback—the listeners filled out an anonymous survey asking three questions (with boldfaced and italicized words as shown):

Q1 How **persuasive** is the **speaker**, apart from the details of the pitch?

Q2 How **convincing** is the **content** of the pitch, apart from the way in which the speaker delivered it (e.g. if you had *read* it)?

Q3 How **effective** is the **presentation style**, apart from details of the pitch and the way in which the speaker delivered it? (where "presentation style" addresses number of "ums," sentence structure, pacing, flow of information, etc.)

Each question could be scored any integer value from 1 ("hardly") to 10 ("very"). Participants were also asked their gender and the gender of the speaker. The surveys were collected after each pitch, encouraging participants to rate all speakers independently, rather than comparatively within a session.

The survey responses were summarized into an average score for each question for each speaker. The resulting distributions look roughly normal, with (mean, std. dev.) of (6.7, 1.4), (6.6, 1.3), and (6.5, 1.4) for questions 1–3, respectively.

4.2.2 Relationship of Persuasion, Content, and Style

The three survey questions (Q1–Q3) were intentionally designed in order to isolate three different aspects of a pitch: persuasion, content, and style. Somewhat surprisingly, we find that these factors are highly intertwined. Taking the pair-wise correlations between speakers' average scores on each of the three questions shows clear relationships:

$$\text{corr}(Q1, Q2) = .83$$

$$\text{corr}(Q1, Q3) = .94$$

$$\text{corr}(Q2, Q3) = .77$$

Here, $N = 42$, and all three p-values are less than 10^{-8}.

These relatively high correlations offer two possible implications: (1) style, content, and persuasion are intrinsically related, or, (2) people are not very good at distinguishing among these characteristics. The first theory does not make much sense. For example, a truly charismatic speaker does not lose his or her flair when peddling weak content (think of the stereotypically slick used-car salesperson). By the same token, fascinating content can still be delivered by a boring, awkward speaker.

Thus we support the second theory, which is very telling in terms of human judgment. We suspect that what listeners think they are perceiving as good or bad content is actually very heavily influenced by the context (i.e. the delivery). Put another way, your ability to persuade a crowd may have as much to do with your presentation style as your message.

4.2.3 Statistical Analysis

For each of the speech features, we computed the mean and standard deviation (where applicable), resulting in a set of summary statistics for the entire pitch. Since the upper-bound for the length of pitches recorded was about 5 minutes, this approach is justified by Pentland et al., who find that 5-minute audio chunking provides ample information for the prediction of social phenomena [26].

Visually inspecting plots of the survey outcomes against the speech statistics showed that, where present, relationships were generally linear. We modeled the effect of speaking patterns on persuasion using a multivariate linear regression.

4.3 Results

4.3.1 Persuasion as Explained by Speech Features

The single feature we find to be most correlated with high marks in persuasiveness was voicing rate ($r = .46, p = .003$)—persuasive speakers talk faster than others. This finding agrees with a study on charisma by Rosenberg and Hirschberg, who find that a faster speaking rate (in terms of syllables per second) corresponds to a higher charisma rating [29].

The average length of a voiced segment shows a fairly strong negative association with persuasion ($r = -.36, p = .02$), but does not provide much new information when already considering voicing rate. In fact, with a correlation of $r = -.93$ between the two measures, they might be considered stand-ins for each other. This makes intuitive sense, as a high voicing rate would typically imply short, succinct syllables.

The mean-scaled standard deviation of the energy also shows a statistically significant negative trend, decreasing with persuasion ($r = -.34, p = .03$). This indicates that persuasive speakers employ well-regulated volume dynamics.

The remainder of the individual speech features prove uninteresting in the context of persuasion, save, perhaps, the standard deviation of the spectral entropy ($r = .29, p = .07$).

We build The ElevatorRater upon a combination of these explanatory variables. Considering just voicing rate and mean-scaled standard deviation of the energy, we achieve an r^2 of .36 ($p \ll .01$). A scatter plot of this model is shown in figure 4-1. Augmenting the model to include the standard deviation of the spectral entropy as well gives $r^2 = .38$ ($p \ll .01$). Figure 4-2 shows a scatter plot of the augmented model. Coefficients for both regressions are shown in table 4.1.

4.3.2 Gender Differences

Several interesting trends emerge from the data when segmenting according to gender of the speaker and gender of the listener. Table 4.2 shows the four sub-groups

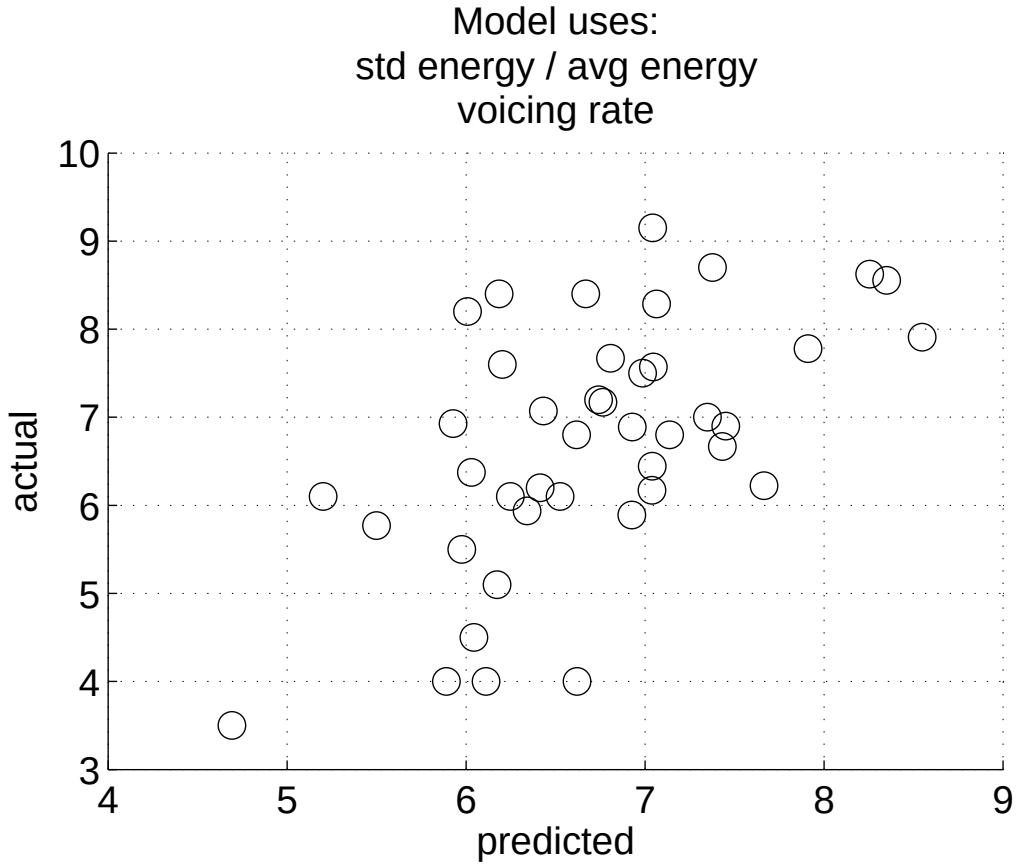

Figure 4-1: Actual vs. predicted values of persuasion by The ElevatorRater. Model incorporates just two speech features and explains 36% of the sample variance ($N = 42$).

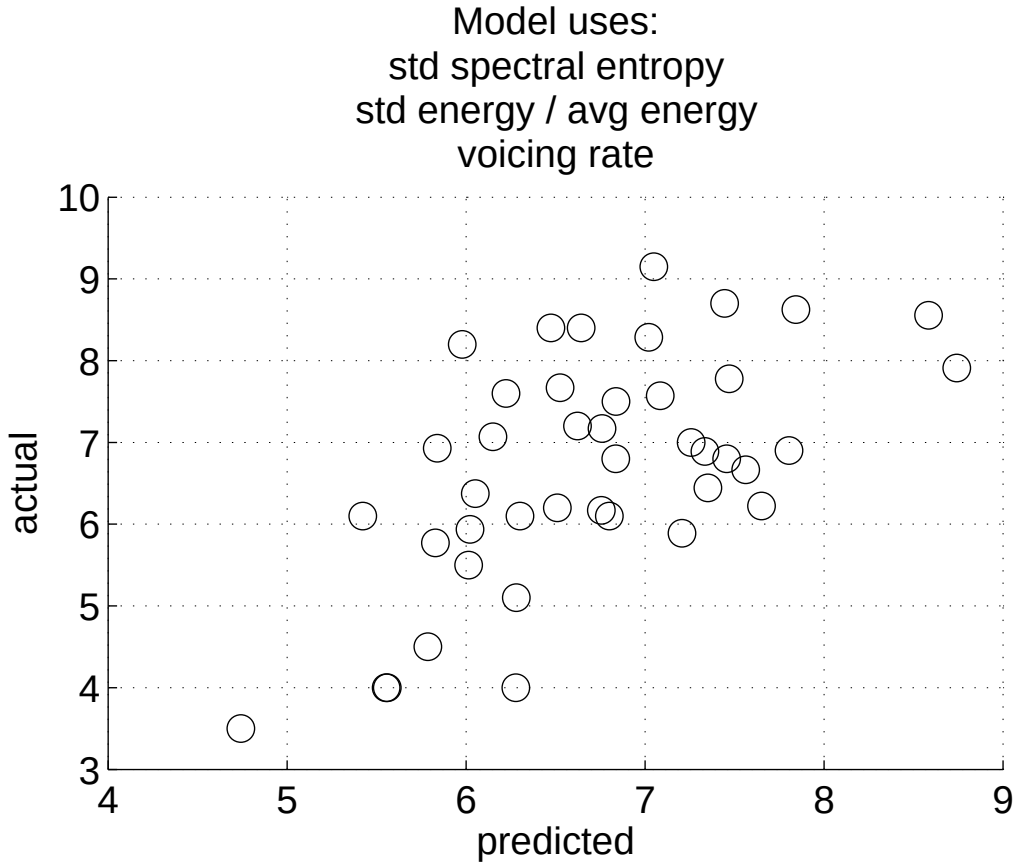

Figure 4-2: Actual vs. predicted values of persuasion by The ElevatorRater. Model incorporates three speech features and explains 38% of the sample variance ($N = 42$).

Feature	Model 1	Model 2
Voicing rate	100.6	98.0
Std. dev. of energy	-1.5	-1.3
Std. dev. of spectral entropy	—	4.8

Table 4.1: Coefficients for the two- and three-variable regressions.

considered. The number of samples for each sub-group correspond to the number of individual surveys filled out.

speaker gender	listener gender	number of samples
F	F	49
	M	82
M	F	52
	M	79

Table 4.2: Four sub-groups and the number of surveys filled out in each.

Figure 4-3 shows the sample mean (vertical line) for responses to each of the three questions with a 95% confidence interval for the population mean (gray box). The mean responses for each of the four sub-groups are overlaid—circles for female speakers; squares for male speakers; filled shapes represent same-gender speaker/listener; outlined shapes represent mixed-gender.

Notice that several sub-groups show means that are significantly ($p < .05$) different from the population mean (indicated by points lying outside the gray boxes in figure 4-3). Males tended to give particularly high scores to female speakers—their average rating for females is above the population mean in all three questions. On the other hand, male listeners tended to score other males (for Q1 and Q2—persuasion and content) below the population mean. For Q2 (content), females rated other females above the population mean.

Two other patterns in our data set are worth commenting upon. First, females were rated higher on average than males on all three questions (circles higher than

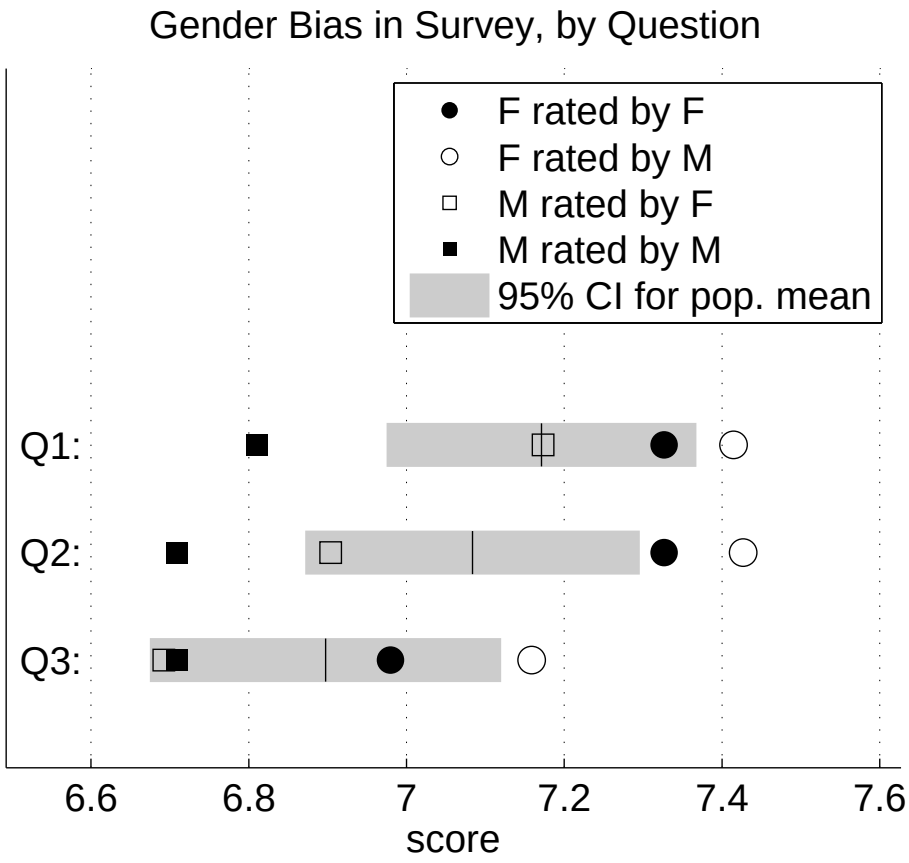

Figure 4-3: Gender sub-group means shown against 95% confidence interval of true population mean for that question.

squares in figure 4-3), with $p < .02$ for all three questions. Second, cross-gender scores seem to be higher than same-gender scores (open shapes higher than filled shapes in the figure), though the only observation that reached statistical significance of $p < .05$ was that male speakers were rated more highly by females than by males in terms of persuasion (Q1).

It turns out that males and females also exhibit some differences in speaking patterns when delivering pitches. For males, the mean-scaled standard deviation of f_0 increases with persuasion ($r = .47, p = .04$), though there is no trend for females. Similarly, male persuasion shows a positive correlation with entropy in the length of speaking segments ($r = .39, p = .09$).

4.4 Discussion and Conclusion

In this chapter, we showed that persuasion correlates with speaking patterns. We found that a simple linear model can explain 38% of the sample variance ($N = 42$) associated with perceived persuasion in a pitch. The most important factors are the voicing rate and the mean-scaled standard deviation of the energy, which together account for 36% of the sample variance. Voicing rate was positively correlated with persuasion, while standard deviation of energy was negatively correlated. So, simply put, listeners perceived greater persuasion in those speakers who spoke quickly and maintained even volume dynamics. We could hypothesize that these two features mean a speaker is delivering large amounts of information per unit time while still displaying calm collection.

We implemented our findings in software that we called The ElevatorRater. The software assesses pitches along the axes of the three significant speech features and can be used to improve one's own public speaking skills. The ElevatorRater works independently of content, so privacy concerns are mitigated and natural language processing difficulties are avoided.

The two most important features in persuasion—voicing rate and standard deviation of energy—fit our proposed model of Activity and Emphasis. Those speakers exhibiting high amounts of Activity (in the form of faster speech production) and lower amounts of Emphasis (in the form of low volume variance) were considered to be the most persuasive.

The ElevatorRater is one way to help identify the subtle characteristics of speech that contribute to persuasiveness. From a coaching perspective, speakers who practice delivering a pitch using The ElevatorRater learn how to focus on *how* they are saying things, and consequently will improve their pitch's reception without necessarily changing *what* they are saying.

Where gender is concerned, we show that there are significant differences in both a speaker's perception and his or her speaking patterns. In terms of perception, several gender-based subsets of the sample population showed significant differences from the

sample mean. One might consider controlling for effects such as physical attraction in future studies of persuasion. In terms of speaking patterns, males exhibit at least two speech trends that females do not. Enhanced versions of The ElevatorRater could exploit these asymmetries to form better assessments, given (or inferring) the gender of the speaker.

4.5 Future Directions

We found informal validation of our model in a preliminary effort to simulate persuasion in voice[1]. Taking pitches that had been human-rated as mediocre, we modified the sound signal to take on properties of high or low persuasion. To increase the perceived persuasiveness of a pitch, we time-compressed the signal, thereby increasing the voicing rate (making sure to use a technique that would leave f_0 and other spectral elements intact), and modulated the amplitude of the signal to make the volume more uniform around the mean. To decrease the perceived persuasiveness, we applied the inverse transforms to the base pitch.

In a limited study, we found that listeners comparing the modified pitches to the originals did indeed perceive the higher or lower levels of persuasiveness that we sought to simulate. This success opens the door for a host of speech synthesis and speech modification applications and needs to be investigated more rigorously in the future.

In an effort to further expand our initial study, we have partnered with a local venture capital firm to initiate a broader study of persuasion, focused exclusively on entrepreneurs. This program will build upon The ElevatorRater platform to assess pitches in request of funding. The setup will be more standardized (all speakers will answer the same set of questions) and will be more representative of "real world" elevator pitches, with significant stakes involved. We expect to complete this study in the 2006-2007 time frame.

[1]This work was done in conjunction with Ben Waber.

Chapter 5

Service —
Low Activity, High Emphasis

In this chapter we describe our setup and execution of a customer service call center experiment, and report results linking speaking style of both the customer and the service agent to the success of the call. We build a model that is greater than 85% accurate at separating successful from unsuccessful calls. Our analysis shows that low Activity and high Emphasis on the part of the agent are common traits among the successful calls. We hypothesize that this combination avoids overwhelming or stifling the caller, while at the same time signaling openness to comment.

Customer satisfaction is paramount in the service industry, and our analysis techniques could prove very useful in the booming call center market. We propose several related applications that could follow from this research, including agent self-training, real-time feedback, style matching, and manager review.

5.1 Background

Corporate call centers are a crucial means by which a company can enhance its accessibility to customers. Call centers serve to fill the following basic information needs [5]:

1. answering customer questions

2. acting on customer requests

3. resolving customer issues

4. rectifying customer complaints

Companies recognize that such customer access adds value to the sales transaction and, in some markets, may be one of the few factors differentiating them from their competition [16]. Customers now expect (and demand) telephone access to companies [13], and the world market for call centers is estimated at hundreds of billions of dollars [9].

Call centers offer a compelling environment for our research for several reasons. First, all interaction is necessarily vocal, thus isolating speech from other social signaling channels (e.g. gestures, facial expressions). Second, customer satisfaction is notoriously low in such settings [9], leaving ample room for improvement. Finally, the literature acknowledges that little work has been done in suggesting what variables are related to caller satisfaction [9] [16]; the only metrics generally studied are those pertaining to call center operations (e.g. average speed of answer, queue time, abandonment rate, type of music played while on hold, etc.), whereas we are primarily interested in how to improve sales by enhancing agent-caller interactions.

5.2 Vertex Call Center

To test our framework in the context of sales calls, we collaborated with Vertex, one of the United Kingdom's largest providers of business process outsourcing [1]. In particular, we worked with their customer service branch in a call center serving Tesco plc. Based in the UK, Tesco is one of the world's leading international retailers, with sales of £37.1 billion (approx. $70 billion) in 2005 [2].

5.2.1 Experimental Setup

Raw call center data (i.e. recorded agent-customer interactions) are notorious for being fiercely guarded intellectual property in this highly competitive industry. (No

doubt, this difficulty has contributed to the dearth of research in this area.) Our collaboration was made possible in part by the fact that our approach works independently of linguistic content, thereby reducing privacy concerns.

We were able to collect speech features from customer service agents handling calls about a Tesco home phone product. All manner of calls relating to this product were processed, including sales, cancellations, questions, billing problems, complaints, etc. All calls were 30 seconds or longer.

Over a period of two days, we gathered information from 70 such calls, handled by 8 different agents (2 male and 6 female). Immediately after each call, the agent was asked to declare the call as 'successful' or 'unsuccessful.' Of our 70 samples, 39 were rated as successful, with the balance rated unsuccessful.

5.3 Results

Considering just the agents' speech features, we ran a stepwise forward linear regression to find that four features hold explanatory power in identifying successful sales calls: standard deviation of the spectral entropy, average length of a speaking segment, voicing rate, and call duration ($r^2 = .50$, $p \ll .01$). The regression coefficients, summarized in table 5.1, show that the length of a speaking segment is negatively correlated with success, while the remaining three features are positively correlated.

Looking instead at just the callers' speech features, we also find significant features that correlate with a successful call (bearing in mind that 'success' here is still defined from the agent's viewpoint): standard deviation of energy, average length of a speaking segment, and call duration ($r^2 = .36$, $p \ll .01$). All three features are positively associated with success; see table 5.2 for individual coefficients.

Finally, we took both the agents' and callers' speech features together and, not surprisingly, came up with a still more powerful model. As in the individual cases, we ran a stepwise forward linear regression to determine the best set of features. The results were fully reflective of the individual models: on the agent side, we found standard deviation of the spectral entropy, average length of a speaking segment, and

voicing rate; on the caller side, we found standard deviation of energy; on both sides, we found the call duration ($r^2 = .58$, $p \ll .01$). This data is summarized in table 5.3.

The call duration feature was obviously the same for both the agent and the caller, so the only individually significant feature that did not carry over to the combined model was the callers' speaking segment lengths. A closer inspection, however, suggests that the callers' long speaking segments are probably complementary to the agents' short speaking segments, such that only one needed to be included in the model.

Labeling the training data as 1 for success or 0 for failure, our model distributes the calls as shown in figure 5-1, where the light gray histogram represents successes and the dark gray histogram represents failures. Fitting Gaussians to the resultant distributions shows the optimal decision boundary at 0.5, which yields classification accuracy of 87% (with 9% false positives and 4% false negatives).

5.4 Discussion and Conclusion

We find that success in a sales call is highly explainable by simple, non-linguistic speech features. Analyzing audio streams from both the agent and the caller, yields binary decision accuracy of over 85% in classifying a call as successful or not. In cases where one or the other audio stream is not available (for technical reasons or privacy concerns), we show that individual models are viable and quite potent as well.

Our findings offer a reasonable interpretation about good customer service. First, the fact that longer call duration correlates with success shows that more interaction between agent and caller is fruitful. Looking next at the caller, increased volume dynamics—a type of vocal Emphasis—can signal interest; see chapter 7. Finally, on the agent side, high voicing rate and short speaking segments combine to provide high transfer of information while allowing ample opportunities for the customer to speak. High standard deviation of spectral entropy in the agent's voice comes from variation in pitch and pitch accents (e.g., the voice patterns at the end of a sentence when asking a question). We hypothesize that this sort of variation indicates indicates

Feature	Sign	Coefficient
Standard deviation of spectral entropy	+	3.47
Average length of a speaking segment	−	-0.17
Voicing rate	+	25.97
Call duration	+	0.0014

Table 5.1: Agent speech features associated with successful sales calls. Also shown is the sign of the correlation and the coefficient from linear regression.

Feature	Sign	Coefficient
Standard deviation of energy	+	0.66
Average length of a speaking segment	+	0.20
Call duration	+	0.0011

Table 5.2: Caller speech features associated with successful sales calls. Also shown is the sign of the correlation and the coefficient from linear regression.

Side	Feature	Sign	Coefficient
Agent	Standard deviation of spectral entropy	+	3.35
	Average length of a speaking segment	−	-0.21
	Voicing rate	+	18.64
Caller	Standard deviation of energy	+	0.42
Both	Call duration	+	0.0010

Table 5.3: Agent and Caller speech features associated with successful sales calls. Also shown is the sign of the correlation and the coefficient from linear regression.

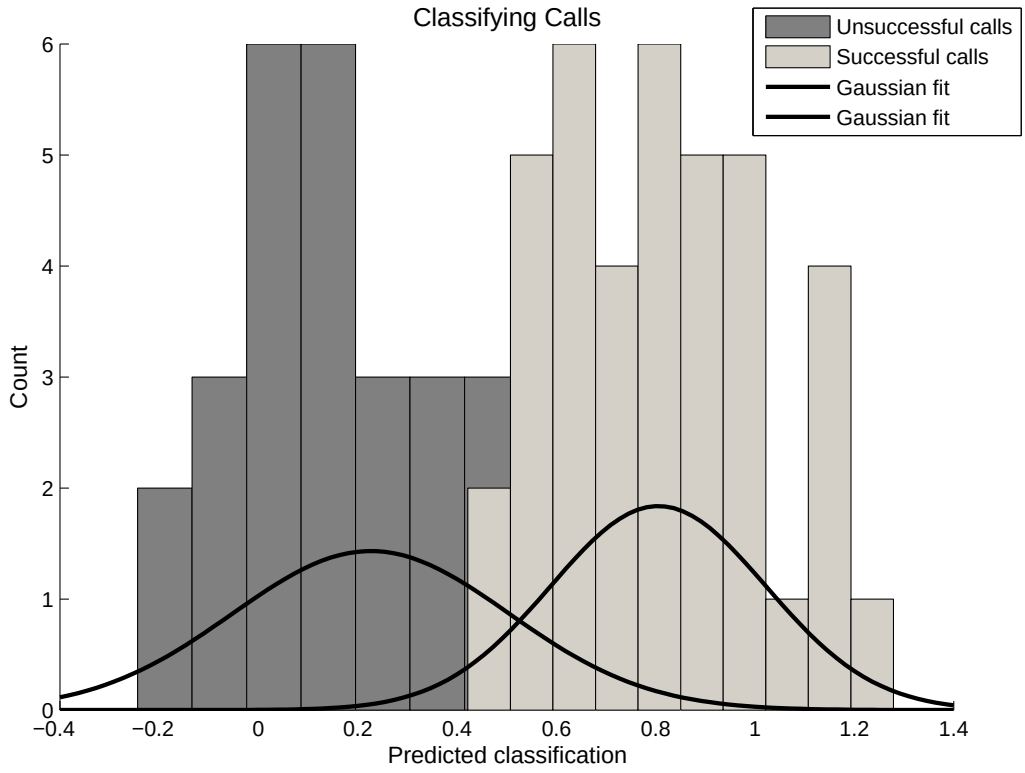

Figure 5-1: Classification of sales calls as successful or not based on agent and caller speech features. A binary decision boundary at 0.5 yields accuracy of 87% over the sample data.

openness to comment or new information. Simply put, successful calls tended to involve agents who were unobtrusive, receptive, and sparked interest from the caller.

In terms of our proposed framework, we focus here on the agent, who seeks to maximize the number of successful calls. The strategy that emerges from our study is one of low Activity (short speaking segments) and high Emphasis (large standard deviation in spectral entropy). The fact that we found high voicing rate to also be a factor in success may seem to contradict the notion of 'low Activity,' but we feel that in this case, the high voicing rate is simply a means to achieve short speaking segments.

5.5 Future Directions

While we focused mainly on how to identify successful sales calls, the environment is ripe for other applications. Since humans, as social beings, act with a fair degree of predictability, a call center could easily amass a database to associate agent/caller speaking patterns with general trends (e.g. success, failure, interest, annoyance). We see many possibilities in this direction, including:

- **Self-training and real-time feedback.** Agents would have a tool visible to them during customer service interactions providing them with real-time feedback on how their speaking patterns compare to known trends—e.g. "customer is interested but agent is being overbearing and may lose sale." The agent would then be able to adapt his or her strategy to improve the chances of reaching a positive outcome.

- **Style matching.** One chronic dilemma of call centers is that of matching a caller with the agent who will best serve the caller's needs. Current practice may assign an agent randomly or, at best, geographically. But it is indisputable that different customers will be best served with different agent styles. For example, a fast-talking, detail-oriented representative may strike a rapport with one customer while inducing nervousness and confusion in another. Some cur-

sory speech analysis could assist in matching a caller with an agent who has previously worked well with speakers of that type.

- **Manager review.** These tools could offer call center managers new ways to assess why some calls were successful and others were not. Alternatively, a manager could effectively oversee large groups of agents by obtaining real-time data summarizing their speaking patterns. The manager could identify problems early and intervene before they caused loss of sale. Trying to accomplish this without such tools—say, by listening to a large number of agents at once—would be cumbersome if not impossible.

Additionally, a future study might strive to take into account the customers' perspective for measures of success and satisfaction; the agents' perception of these values may differ from those of the caller.

Chapter 6

Depression — Low Activity, Low Emphasis

In this chapter, we describe a collaboration with the Boston Medical Center in a preliminary study to design a telecommunication system for monitoring the mental health of depressed individuals.

While several studies of depression have been done in clinical or long-term settings, we show that even short, one-sided audio clips offer a window into the illness. Taking a thin-slicing approach we are able to determine which patients would get better, get worse, or not change significantly. Among those patients who did show change in depression severity, we find simple measures—voice pitch variation and speaking rate—correlate intuitively with their depression severity.

Our findings give clear support to the notion that low levels of both Activity and Emphasis can signal depression.

6.1 Background

Depression is a very real problem in the world. According to the Global Burden of Disease study—initiated by the World Bank in 1992 and carried out by the World Health Organization (WHO)—unipolar major depression will rank second in magnitude of global disease burden in 2020 in established market economies, up from rank 4

in 1990 [24]. In the United States alone, the estimated monetary costs for depression exceeded $44 billion in 1990 [23].

Fortunately, depression is a treatable illness [23]. But despite the gravity of the condition, very little work has been done to employ objective physiological measures in the diagnosis of depression, monitoring of treatment response, or predicting of early signs of relapse [32]. Current clinical practice for assessment of depression severity still centers on the Hamilton Depression Rating Scale, an instrument developed in the late 1950s.

The Hamilton Depression Rating Scale (HAM-D) consists of 21 questions (though countless variations exist) measuring a variety of factors, including insomnia, psychomotor retardation, anxiety, loss of weight, etc. [20]. Higher HAM-D ratings correspond to more severe depression. The HAM-D is taken to be the "gold standard" among depression researchers due to its extremely high total score reliability and its clinically-proven discrimination validity [7]. But despite holding a position as *de facto* depression assessment standard for 50 years, the scale has recently come under criticism for being a potential source of subjectivity and having poor inter-rater and retest reliability [32] [7]. Further, the HAM-D requires skilled clinicians to administer the survey and interpret the results, which can often be a costly undertaking [34].

In an effort toward augmenting or replacing the HAM-D with physiological measures, there is evidence to support a relationship between speech prosody (and other physiological cues) and the severity of depression. Sung et al. are able to accurately track depression state using non-invasive, continuous monitoring of speech and movement patterns [32]. Sobin and Sackeim cite studies showing that depressed patients may show slowed responses, monotonic phrases, and poor articulation. Furthermore, they suggest that such speech patterns will return to normal values as patients improve [30].

6.2 TLC-Depression

In an effort to help adult patients with unipolar depression improve adherence to their antidepressant medication regimens, investigators at the Medical Information Systems Unit, Boston University Medical Campus/Boston Medical Center launched a study named TLC-Depression: **T**elephone-**L**inked **C**are for Adherence to Treatment Regimen in **Depression**. The study would assess the effectiveness of a computer-based telecommunications system to:

1. monitor patients' adherence to their treatment regimens, focusing on antidepressant medication-taking and follow-up office visits, as well as monitoring their psychological and general health status over time.

2. provide patients with education and counseling aimed at improving their adherence to their medication regimen and follow-up office visits.

3. generate reports for mental health caregivers from information collected from their patients.

Participants in the study were asked to call the TLC-Depression system once a week on a prearranged day and time of their preference. Interaction with the system lasted less than 10 minutes and consisted of reminders to take medication, schedule appointments, etc. At the end of the phone session, the patients were allowed to leave a one-minute voice message for their care practitioner on any topic desired. It is these messages that we used for our analysis.

The study spanned from January 2004 to May 2005.

6.2.1 Subjects

The study population for TLC-Depression consisted of 120 adult patients from the Boston Medical Center Department of Psychiatry. All patients had been diagnosed with Major Depressive Disorder and/or Dysthymic Disorder and were taking at least one prescribed antidepressant medication. (Patients diagnosed with Bipolar Disorder,

schizo-affective disorders, or significant personality disorders were ineligible for the program.) All patients spoke and understood conversational English.

During the TLC-Depression program, patients made five checkup visits to the clinic at one-month intervals. At the first (T_0) and last (T_4) visits, clinicians administered the Hamilton Depression Rating Scale questionnaire. We used the HAM-D results of these two visits to represent initial and final depression severities.

From the initial 120 patients who participated in TLC-Depression, 81 stayed in the program through the final (T_4) checkup. Reasons for patients' withdrawals include transportation problems, phone disconnection, moving away from Boston, stopping depression medication, physical injury, and feeling of inadequate compensation for study participation. Of those patients who did complete the program, 56 did not leave a message and 25 left at least one message. Among those leaving at least one message, 23 left multiple messages.

	Number of messages		
	0	>0	>1
Total number of patients	56	25	23
Average T_0 HAM-D score	24.5	23.3	23.5
Average aggregate improvement (reduction in HAM-D points)	1.5	5.7	5.9
Percent who improved, according to HAM-D score	57%	80%	83%

Table 6.1: Summary of patients with both initial (T_0) and final (T_4) Hamilton Depression Rating Scale (HAM-D) surveys.

6.3 Results

6.3.1 Survey Data

The raw data shows striking differences in improvement between the group that left at least one message and the group that left no messages. See table 6.1. On average, the HAM-D scores for those who left messages was 5.7 points lower (i.e. less severe) at T_4 than at T_0, compared with a drop of just 1.5 points over the same period for those who left no messages. Further, 80% of patients in the message-leaving group showed some improvement over the study period, whereas only 57% showed any improvement in the no-message group. Both of these differences are statistically significant with $p \ll .01$, but it is unclear whether the relationships are causal.

6.3.2 Call Analysis

Classifying the Callers

We first sought to answer two questions:

1. Can we identify those patients who would get worse over the study period?
2. Can we distinguish patients who would change substantially (more than 15% increase or decrease in HAM-D rating) from those who did not?

The first question would be valuable in allowing clinicians to flag those patients potentially needing special attention. The second question, while not too interesting by itself was important for a further analysis; see below.

In order to answer these questions, we take a 'thin-slicing' approach. In the same way that Ambady et al. were able to predict malpractice suits from snippets of doctor-patient conversations, we are able to accurately answer the above two questions based on a single call from a TLC-Depression patient.

We took speech features from the first call that a patient made to TLC-Depression and applied a two-variable quadratic classifier. To separate those patients who would get worse from those who would get better or stay the same, we found the two variables

with the most explanatory power were: (1) average length of a voiced segment, and (2) entropy in the length of pauses. The Gaussian parameters are summarized in table 6.2. We achieved 88% accuracy in binary segmentation (with 12% incorrectly classified as having improved).

Next, we considered identification of those who showed appreciable change (greater than 15% improvement or regression from T_0 to T_4). We achieve 84% binary decision accuracy in this task by applying a quadratic classifier over the following two features: (1) fraction of time speaking, and (2) entropy in the length of pauses (with 12% misclassified as having improved significantly and 4% misclassified as showing little change). The Gaussian parameters are summarized in table 6.3.

It is interesting to note that the pause-length entropy showed discriminatory power in both questions.

Monitoring Depression Severity

While making predictions about a patient's future wellbeing from an initial call is valuable information—especially as an early warning system—we would also like to be able to monitor progress throughout the program as the patient makes regular calls. To this end, we find that there exists a simple relationship between speech features and mental health improvement.

Patients who *did not* exhibit much change in HAM-D rating over the course of the study (those in the less-than-15%-change group discussed above) did not exhibit meaningful patterns in speech from one call to the next. This is, of course, not surprising, given their relatively constant depression severity.

Allowing ourselves now to restrict analysis to those who *did* exhibit substantial HAM-D rating change during the study, we are able to explain more than half the variance of the percent improvement (positive or negative) in HAM-D rating from T_0 to T_4. We accomplish this by considering a linear combination of (1) the standard deviation of f_0, and (2) the voicing rate ($r^2 = .504$, $p \ll .01$). The coefficients for the two features, respectively, are 2.51 and 14.91, indicating that both values increased with more positive HAM-D percent improvements.

	Features	Means	Covariance
Patients who got worse	$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$	$\mathbf{m} = \begin{bmatrix} 0.2443 \\ 1.3897 \end{bmatrix}$	$\mathbf{K} = \begin{bmatrix} 0.0008 & 0.0025 \\ 0.0025 & 0.0516 \end{bmatrix}$
Patients who got better or stayed the same	$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$	$\mathbf{m} = \begin{bmatrix} 0.2699 \\ 1.5185 \end{bmatrix}$	$\mathbf{K} = \begin{bmatrix} 0.0024 & -0.0101 \\ -0.0101 & 0.1539 \end{bmatrix}$

Table 6.2: Gaussian parameters for quadratic decision rule to classify those depression patients who got worse. The two variables involved are $[y_1, y_2]^T$ = [average length of a voiced segment, entropy in the length of pauses]^T.

	Features	Means	Covariance
Patients who changed $< 15\%$	$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$	$\mathbf{m} = \begin{bmatrix} 0.6751 \\ 1.4430 \end{bmatrix}$	$\mathbf{K} = \begin{bmatrix} 0.0266 & 0.0155 \\ 0.0155 & 0.1137 \end{bmatrix}$
Patients who changed $> 15\%$	$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$	$\mathbf{m} = \begin{bmatrix} 0.6920 \\ 1.5162 \end{bmatrix}$	$\mathbf{K} = \begin{bmatrix} 0.0088 & 0.0214 \\ 0.0214 & 0.1483 \end{bmatrix}$

Table 6.3: Gaussian parameters for quadratic decision rule to classify those depression patients who did not change appreciably (less than 15% change in HAM-D rating from T_0 to T_4). The two variables involved are $[y_1, y_2]^T$ = [fraction of time speaking, entropy in the length of pauses]^T.

6.4 Discussion and Conclusion

Our findings leave us optimistic about the potential of non-linguistic speech features acting as a proxy for the HAM-D in some cases. We showed that there exists a strong relationship between positive improvement over the course of the TLC-Depression study and both higher vocal pitch variance and more rapidly produced speech. These findings are pleasingly intuitive and agree with the literature on the subject [32] [30].

Within our proposed social signaling framework, our results clearly support the hypothesis that low Activity and low Emphasis can signal depression. Here, Activity is represented by speaking rate, and Emphasis is represented by voice pitch variation. Low measurements in these two dimensions were a strong indicator of negative improvement over the course of the TLC-Depression study.

We were surprised to not find any association between speech features and *absolute* HAM-D ratings (our models worked with *percent change*), especially given strong correlations reported by Sung et al. [32] for a host of physiological measurements. We may perhaps attribute this to the fact that Sung was able to conduct continuous data collection over the course of days and weeks; our analysis of 30-second phone calls offers a considerably more restrictive view. It is, of course, the ability to work under such restrictions that makes our technique widely applicable with low barriers to implementation, and so we accept some limitations of the model.

6.5 Future Directions

Based on the positive findings presented here, we are in discussions with Boston Medical Center investigators to propose a follow-up study that will integrate speech feature analysis into the TLC-Depression program. This will offer further validation of our models and eventually—if the association between speech features and mental health is reliable—become an objective measure suitable for patient diagnosis. We expect to apply for the grant in late 2006.

In this iteration of our analysis, it may make sense to begin taking into account

external factors, such as type and dosage of medication, age, gender, etc.

Chapter 7

Attraction — High Activity, High Emphasis

In this chapter, we present experimental work from "Thin Slices of Interest" by Madan [22] to measure attraction as a function of speaking patterns. Madan takes the same approach as we do and uses the same speech processing toolkit.

We show that Madan's findings round out our social signaling framework by offering evidence that high vocal Activity and high vocal Emphasis are strongly correlated with interest and attraction.

7.1 Experimental Setup

Madan conducted his study at a real-world "speed dating" event. Speed dating is a relatively new way for singles to meet many potential matches in a single evening. Participants interact with their randomly chosen 'dates' (other participants) in five-minute sessions. At the end of a session, each individual indicates to a 3^{rd} party whether he or she would like to provide contact information to the other. A 'match' is found when both parties answer 'yes,' and they are later provided with mutual contact information.

Madan collected 60 five-minute speed dating sessions from individuals ages 21–45. The audio for each couple was recorded into separate streams using unobtrusive

directional microphones.

7.2 Results

While Madan found little correlation between male speaking patterns and attraction (i.e. 'yes' responses), he found that female speaking patterns significantly explained both female ($r = .48$, $p = .03$) *and* male ($r = .50$, $p = .02$) attraction. He concludes that female social signaling is more important in determining a couple's attraction response than male signaling. It is unclear whether males simply signal less (or in ways that we do not measure) or whether they are masking their behavior in this particular context.

For female attraction, the most important factor was high Activity, though high Emphasis also played a role. Activity features were fraction of time speaking and voicing rate. The Emphasis feature was standard deviation of f_0. Coefficients are shown in table 7.1. Together, the Activity and Emphasis features produced a classifier with a cross-validated decision accuracy of 71% in predicting attraction.

Feature	Coefficient
Fraction of time speaking	3.038
Voicing rate	10.62
Standard deviation of f_0	1.748

Table 7.1: Coefficients for the linear regression of female speaking patterns against attraction. Note that all coefficients are positive.

7.3 Discussion and Conclusion

We find Madan's experimental results to be consistent with our proposed framework, for at least females. Madan shows that Activity—in the form of high voicing rate and large percentage of time speaking—in combination with Emphasis—in the form

of increased voice pitch variation—serve to signal female attraction in a speed dating situation.

We have hypothesized that Activity can indicate desire to *provide* dialog and information and that Emphasis may serve to *invite* comment and conversation. As such, it is reasonable to believe that females, when attracted to a conversation partner, will exhibit ample amounts of both of these behaviors within a five-minute session.

Chapter 8

Conclusion

In this thesis, we proposed a framework in which varying levels of vocal Activity and vocal Emphasis corresponded to the perception of four different representative behavioral states: persuasiveness, effectiveness in service, depression, and attraction. We conducted experiments focusing on each of these four behaviors and found strong correlations supporting our hypothesis.

Needless to say, these metrics are merely a guidance—we would not expect any such framework to be perfect 100% of the time because people do not behave with perfect regularity 100% of the time. We are confident, however, that this groundwork offers insights towards identifying social phenomena from simple, thinly-sliced observations. More sophisticated techniques in speech and addition of other social signals (e.g. body motion, facial expressions, etc.) could be applied in further research.

Our studies were intentionally grounded in real-world situations so that future work may extend upon our results to provide solutions to practical problems. In some settings, notably telecommunication, our work is especially relevant because voice is the only means of interaction. There have already been efforts to deploy similar technologies on mobile platforms as personal coaches or interest meters, and the work in this thesis further expands those possibilities.

Appendix A

Speech Feature Extraction Code

Our speech analysis platform is a suite of MATLAB functions designed and built and at the MIT Media Lab. Updated code, usage instructions, and examples are freely available at `http://groupmedia.media.mit.edu/`.

Bibliography

[1] www.vertex.co.uk.

[2] www.tescocorporate.co.uk.

[3] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psycholological Bulletin*, 111:256–274, 1992.

[4] Nalini Ambady, Debi LaPlante, Thai Nguyen, Robert Rosenthal, Nigel Chaumeton, and Wendy Levinson. Surgeons' tone of voice: A clue to malpractice history. *Surgery*, 132, 2002.

[5] Jon Anton. The past, present and future of customer access centers. *International Journal of Service Industry Management*, 11(2):120–130, 2000.

[6] M. Argyle. *Bodily communication*. Methuen, 1987.

[7] R. Michael Bagby, Andrew Ryder, Deborah Schuller, and Margarita Marshall. The hamilton depression rating scale: Has the gold standard become a lead weight? *Am J Psychiatry*, 161:2163–2177, 2004.

[8] Sumit Basu. *Conversational Scene Analysis*. PhD thesis, Massachusetts Institute of Technology, September 2002.

[9] Lynne Bennington, James Cummane, and Paul Conn. Customer satisfaction and call centers: an australian study. *International Journal of Service Industry Management*, 11(2):162–173, 2000.

[10] Ron Caneel. Social signaling in decision making. Master's thesis, Massachusetts Institute of Technology, 2005.

[11] J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1–2):89–132, 2003.

[12] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, revised edition, 1977.

[13] Deborah Cowles and Lawrence A. Crosby. Consumer acceptance of interactive media in service marketing encounters. *The Services Industries Journal*, 10(3):521–540, 1990.

[14] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Addison Wesley, third edition, 2002.

[15] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, 1977.

[16] Richard A. Feinberg, Ik-Suk Kim, Leigh Hokama, Ko de Ruyter, and Cherie Keen. Operational determinants of caller satisfaction in the call center. *International Journal of Service Industry Management*, 11(2):131–141, 2000.

[17] Robert W. Frick. Communicating emotion. *Psychological Bulletin*, 97(3):412–429, 1985.

[18] Jonathan Gips and Alex (Sandy) Pentland. Mapping human networks. To appear in IEEE International Conference of Pervasive Computing and Communications, March 2006.

[19] Malcolm Gladwell. *Blink*. Little, Brown and Company, 2005.

[20] Max Hamilton. A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, 1960.

[21] A. Kendon, R. M. Harris, and M. R. Key. *Organization of behavior in face to face interaction.* The Hague: Mouton, 1975.

[22] Anmol P. Madan. Thin slices of interest. Master's thesis, Massachusetts Institute of Technology, 2005.

[23] Cynthia D. Mulrow. Treatment of depression: Newer pharmacotherapies. Evidence report no. 7, Agency for Health Care Policy Research, February 1996. AHRQ Publication No. 99-E014.

[24] CJ Murray and AD Lopez. Evidence-based health policy: Lessons from the global burden of disease study. *Science*, 274:740–743, 1996.

[25] Clifford Nass and Kwan Min Lee. Does computer-synthesized speech manifest personality? *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001.

[26] A. Pentland, J. Curhan, R. Khilnani, M. Martin, N. Eagle, R. Caneel, and A. Madan. A negotiation advisor. ACM Symposium on User Interface Software and Technology, October 2004.

[27] Alex (Sandy) Pentland. Social dynamics: Signals and behavior. International Conference on Developmental Learning, October 2004.

[28] R. Picard. *Affective Computing.* MIT Press, 1997.

[29] A. Rosenberg and J. Hirschberg. Acoustic/prosodic and lexical correlates of charismatic speech. *Proceedings of Interspeech 2005, Lisbon*, 2005.

[30] Christina Sobin and Harold A. Sackeim. Psychomotor symptoms of depression. *Am J Pyschiatry*, 154:4–17, 1997.

[31] Kenneth N. Stevens. *Acoustic Phonetics.* The MIT Press, 1998.

[32] Michael Sung, Carl Marci, and Alex (Sandy) Pentland. Objective physiological and behavioral measures for tracking depression. 2005. Excerpt from Sung's PhD Thesis at MIT.

[33] Charles W. Therrien. *Decision Estimation and Classification.* John Wiley & Sons, 1989.

[34] Janet B. W. Williams. Standardizing the hamilton depression rating scale: past, present, and future. *Eur Arch Psychiatry Clin Neurosci*, 251:II/6–II/12, 2001.