Computers to Help with Conversations:
Affective Framework to Enhance Human Nonverbal Skills

by

Mohammed Ehsan Hoque

B.S., Pennsylvania State University (2004)

M.S., University of Memphis (2007)

Submitted to the Program in Media Arts and Sciences,

School of Architecture and Planning,

In partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

Signature redacted

Author _____

Program in Media Arts and Sciences
August 15, 2013

Signature redacted

Certified by _____

Rosalind W. Picard
Professor of Media Arts and Sciences
Program in Media Arts and Sciences, MIT
Thesis supervisor

Signature redacted

Accepted by _____

Pattie Maes
Associate Academic Head
Program in Media Arts and Sciences, MIT

Computers to Help with Conversations:

Affective Framework to Enhance Human Nonverbal Skills

by

Mohammed Ehsan Hoque

Submitted to the Program in Media Arts and Sciences
School of Architecture and Planning,
On August 9th, 2013, in partial fulfilment of the
Requirements for the degree of
Doctor of Philosophy

Abstract

Nonverbal behavior plays an integral part in a majority of social interaction scenarios. Being able to adjust nonverbal behavior and influence other's responses are considered valuable social skills.

A deficiency in nonverbal behavior can have detrimental consequences in personal as well as in professional life. Many people desire help, but due to limited resources, logistics, and social stigma, they are unable to get the training that they require. Therefore, there is a need for developing automated interventions to enhance human nonverbal behaviors that are standardized, objective, repeatable, low-cost, and can be deployed outside of the clinic.

In this thesis, I design and validate a computational framework designed to enhance human nonverbal behavior. As part of the framework, I developed My Automated Conversation coacH (MACH)—a novel system that provides ubiquitous access to social skills training. The system includes a virtual agent that reads facial expressions, speech, and prosody, and responds with verbal and nonverbal behaviors in real-time.

As part of explorations on nonverbal behavior sensing, I present results on understanding the underlying meaning behind smiles elicited under frustration, delight or politeness. I demonstrate that it is useful to model the dynamic properties of smiles that evolve through time and that while a smile may occur in positive and in negative situations, its underlying temporal structures may help to disambiguate the underlying state, in some cases, better than humans. I demonstrate how the new insights and developed technology from this thesis became part of a real-time system that is able to provide visual feedback to the participants on their nonverbal behavior. In particular, the system is able to provide summary feedback on smile tracks, pauses, speaking rate, fillers and intonation. It is also able to provide focused feedback on volume modulation and enunciation, head gestures, and smiles for the entire interaction. Users are able to practice as many times as they wish and compare their data across sessions.

I validate the MACH framework in the context of job interviews with 90 MIT undergraduate students. The findings indicate that MIT students using MACH are perceived as stronger candidates compared to the students in the control group. The results were reported based on the judgments of the independent MIT career counselors and Mechanical Turkers', who did not participate in the study, and were blind to the study conditions. Findings from this thesis could motivate further interaction possibilities of helping people with public speaking, social-communicative difficulties, language learning, dating and more.

Thesis Supervisor: Rosalind W. Picard
Title: Professor of Media Arts and Sciences, Program in Media Arts and Sciences, MIT

Computers to Help with Conversations:
Affective Framework to Enhance Human Nonverbal Skills

By
Mohammed Ehsan Hoque

Signature redacted

Reader:_____
Jeffrey Cohn
Professor of Psychology and Psychiatry
University of Pittsburgh
Adjunct Professor, Robotics Institute, Carnegie Mellon University

Reader:_____
Louis Philippe Morency
Research Assistant Professor, Institute for Creative Technologies
University of Southern California

Signature redacted

Reader:_____
Bilge Mutlu
Assistant Professor of Computer Science
University of Wisconsin - Madison

Acknowledgements

First, I would like to acknowledge my incredible advisor, Rosalind Picard, who has always been full of inspiration, support and cheer. The problem that I attempted to solve in this thesis was considered a very high-risk project, and we couldn't find any external funding agency to support it. Roz still felt that it was an important problem, provided all the resources, and encouraged me to take risks. Roz and I shared the passion of helping people with communication difficulties, and the ideas presented in this thesis are truly a collaborative effort. Besides being a great mentor, Roz cared for my well-being and always promoted a healthy work-life balance. When my mother was diagnosed with the final stage of cancer, she allowed me to take time off from my PhD for an entire academic year, and continued to send her best wishes while I was away. I hope to treat my students with the same level of kindness, compassion and respect that she has shown towards me.

I would like to thank my external committee members Jeffrey Cohn, Bilge Mutlu and Louis-Philippe Morency. I feel fortunate to have a very diverse committee who pushed me hard because they believed in me and wanted me to succeed as a scholar, scientist and teacher.

I would like to acknowledge my collaborator Jean-Claude Martin and Matthieu Courgeon from LIMSI-France for allowing me to build the MACH platform on their existing Multimodal Affective and Reactive Characters (MARC). They were very hospitable when I traveled to France for one week to work closely with them. I also appreciate Matthieu traveling all the way to Boston for one week to help me design the animations. Without their support, it probably would have taken me 6 more months to finish this thesis. In addition, I would like to acknowledge Amy Cuddy and Maarten Bos from Harvard Business School for helping me develop the coding scheme to label the interview data. Their enthusiasm towards my work was always a mood lifter. Special acknowledgement goes to Stefanie Shattuck-Hufnagel from MIT RLE for helping me setup the speech prosody reading group and tirelessly answering all my questions. Her excitement has always been a big source of inspiration.

Jon Gratch from ICT, USC, was the first to point out that relying only on the counselor's ratings might be risky and I should solicit other measures as well. I am so glad that I took his advice, as it turned out to be absolutely correct! James Cartreine from Harvard Medical School provided a lot of advice on the study design and pointed me to the literature to learn more about social phobia.

I was very glad to have the opportunity to interact with Sherry Turkle in capacity of my general exam. Despite having a position against the use of robots for conversations in

domains where emotion, feelings, our humanity are fully expressed, she always helped me to see the broader picture. I feel that my intellectual exchanges with Sherry made me more open and thoughtful.

"Intelligent Multimodal Interfaces" class by Randall Davis was undoubtedly the most exciting class that I have ever taken at MIT. Randy's ability to clearly and humorously articulate the ideas on multimodal interaction and interfaces eventually encouraged me write a thesis on it. Knowing that he couldn't make it to my defense due to a last minute conflict, Randy took the time to drop by 10 minutes before the defense to wish me luck. I was very touched by his generosity, and he has my highest respect as an academic.

I would like to acknowledge Cynthia Breazeal and the cool roboticists from the personal robots group. Based on Roz's recommendation, I read Cynthia's PhD thesis and I was very inspired by her work. Cynthia's work has definitely impacted me in many ways and I appreciate her support throughout my PhD.

I was very lucky to be able to excite more than half of the staff members from MIT Career Services to help with my PhD work. In particular, Ellen Stahl, Heather Law, Christina Henry, Colin Smith and Lily Zhang volunteered their time to help with my studies and provided a lot of advice. Special thanks to Colin Smith and Natalie Lundsteen for spending numerous hours to carefully label the interview videos for me. Without their help and support, my research experiments would have been very challenging to run.

I would like to acknowledge all my colleagues from the Affective Computing that I had the opportunity to overlap, including Shaundra Daily, Hyungil Ahn, Hoda Eydgahi, Seth Raphael, Alea Teeters, Kyunghee Kim, Rob Morris, Ming-Zher Poh, Jackie Lee, Ellitott Hedman, Micah Eckhardt, Yuta Kuboyama, Mish Madsen, Dan Mcduff, Javier Hernandez, Akane Sano, Yadid Ayzenberg, Karthik Dinakar, Rana el Kaliouby, Matthew Goodwin, and Rich Fletcher. I have had the good fortune of collaborating with many of them on the capacity of iSet, Speech Enabled Speech games, and MIT Mood Meter. Our friendly group admins Daniel Bender, Lillian Lai and Kristina Bonikowski have played a very important role by helping me with many purchases and travel reimbursements.

I had to recruit 90 MIT undergraduate students to participate in my study. Asking all of them to come three times in the lab during a busy semester was a lot of work, and I appreciate their time, sincerity and enthusiasm in helping with my research. Sarah Wang from the Sloan Behavioral Research Lab had made an exception by allowing me to use the facility for more than 20 hours per week including the weekend. Thank you, Sarah!

Special acknowledgement goes to Mary Himinkool and Hartmut Neven for allowing us to use Google Feature Tracker for our research. Samsung kindly donated a 3D smart TV to our group, which I have used in my studies.

Over the years, I had the opportunity to work with many brilliant undergraduate students including Joseph Lane, Geza Kovacs, Benjamin Mattocks, Ryan Lacey, Anji Ren, Angela Zhang, Sigtryggur Kjartansson, Shannon Kao, Tamanna Islam, Abra Shen, Kristi Tausk, Nicole Pan, Benjamin Schreck, Eunice Lin, Michelle Fung, Sumit Gogia and Kevin Lin. Thank you all for contributing to the scientific discoveries presented in this thesis.

I had a great time being part of Bangladeshi Students Association of MIT and being involved in organizing many cultural events. It was great working and overlapping with Tanvir bhai, Tauhid, Shammi, Sarwar, Ketan, Maysun, Deeni, Nazia, Shadman, Urmi, Sujoy, Zubair, Saima and Sumaiya during my time at MIT. You all will be dearly missed!

We were lucky to have wonderful set of friends in Boston: Tinku bhai, Rumpa Apa, Nasim bhai, Zakia apa, Tareq bhai, Mafruha Apa, Shakib, Mishu, Adnan bhai, and Ashirul bhai. Interacting with all of you gave me the much needed work-life balance during my PhD. When we were blessed with our son, Wali, we didn't have any family support. As a first time parent and without the support, we were confused, sleep deprived, and miserable. I appreciate how you all have stepped in and treated us like a family. I remember Tinku bhai and Rumpa apu offering us to spend the Friday night with them so that they could take care of Wali, while we could get the 8 hours of uninterrupted sleep. The time spent with all of you will always remain very memorable to me!

I would like to acknowledge my dad (Enamul Hoque), brother (Eshteher), sister (Luna), brother-in-law (Babu bhai) and my wonderful niece Pareesa. We went through a difficult transition as a family as we lost our mother during my PhD. On the bright side, it brought us together in the most wonderful way. I also would like to acknowledge my father and mother in law for embracing me as a son as I went through a difficult transition of losing my mother during my PhD. I feel blessed to have your love and care in my life.

I am fortunate to have a wonderful wife, Nafisa Alam, who has been a very supportive partner during my PhD. She worked full time, took care of our son, and did all the household activities. Every evening, she would instant message me the dinner menu in Gchat so that I would come home early from the lab. She took the time to do those little things that always filled my heart with joy and love. My son, Wali, helped me to re-shuffle my priorities in life, and made me realize that his happiness means the world to me.

I would like to dedicate this thesis to the wonderful memories of my beloved mother who taught me the two most important lessons in life; 1) never quit and 2) never forget rule number 1. I miss her every day!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ASD	Autism Spectrum Disorder
AU	Action Unit
CBT	Cognitive Based Therapy
CP	Computer-aided Psychotherapy
FACS	Facial Action Coding System
FPS	Frames Per Second
HCI	Human-Computer Interaction
HCRF	Hidden Conditional Random Fields
HMM	Hidden Markov Models
MACH	My Automated Conversation coacH
ROC	Receiver Operator Characteristics
SD	Standard Deviation
SVM	Support Vector Machines
VE	Virtual Environments
VR	Virtual Reality
VRCBT	Virtual Reality Cognitive-based Therapy

Chapter 1: Introduction

As computers take on a ubiquitous role and intertwine themselves into our lives in many forms and shapes, they are still fundamentally limited in understanding and responding to our nonverbal cues. The limitation could be attributed to our own lack of understanding for the full range of human communication. As we take incremental steps to enhance the theoretical and computational understanding of nonverbal behaviors, what new interaction possibilities might it unveil with our computers? What might that interface look like?

Nonverbal behaviors are fundamentals human face-to-face communication (Albert Mehrabian, 2008) . Nonverbal behaviors are used by those around us to define our feelings and personality, and these impact our relationships and likelihood of being successful. Is it possible to use automated technology to enhance human nonverbal behaviors in contexts such as social interactions, job interviews, public speaking, language learning, training medical professionals, counseling, delivering customer service, or dating?

In this thesis, I design, develop, and evaluate a new possibility of interacting with computers to enhance human nonverbal skills. The system has been contextualized to help people improve nonverbal behaviors in job interviews. As part of the exploration of technology development, I present new studies on understanding expressions that provide explanations to the existing theory as well as contradict it. I demonstrate how the generated knowledge could be applied towards nonverbal sensing, and representation, and an interactive learning system to enhance human nonverbal behaviors, offering new insights and interaction possibilities.

1.1 SCENARIO

Let us consider Bob, a young teen diagnosed with social difficulties, including difficulty engaging with people in face-to-face scenarios. Bob finds the social rules of taking turns, making eye contact, and smiling to be polite, confusing. His difficulties in socially interacting with others have only met with taunting, bullying and rejection by his peers. As a result, Bob now has retreated into an online world. It is a fun experience for Bob to practice social language and conversation, and make new friends through online interaction, but he still craves real face-to-face interactions. He wishes to practice his social interactions, but fears social stigma. Is it possible for Bob and others to practice and improve social interactions in their own environment so that they could control the pace of the interaction, practice as many times as they wish and still be in the driver's seat of the interaction?

1.2 NONVERBAL BEHAVIOR

Face-to-face interaction is like dancing, where people continuously adjust their behaviors based on the actions of their interlocutor (L. P. Morency, 2010). During conversations, humans rely on the behaviors of their conversational partners to decide on the next set of actions. Often they try to encode *'extra information'* with their behaviors to strengthen (e.g., affirmation), supplement (e.g., pointing) and contradict (e.g., sarcasm) what they say. According to Mehrabian et al. (A Mehrabian, 1968), Birdwhistel et al. (Judee K Burgoon & Buller, 1994), and Burgoon et al. (Birdwhistell, 1970), the extra information, known as nonverbal behavior, is more significant than the spoken words.

According to Duncan (Duncan Jr. Starkey, 1969), nonverbal behaviors include, 1) body motion or kinesic behavior (gesture and other body movements, including facial expressions, eye movement, and posture); 2) paralanguage (voice qualities, speech non-fluencies, and such non-language sounds as laughing, yawning and grunting); 3) proxemics (use of social and personal space); 4) olfaction; 5) skin sensitivity to touch and temperature; and 6) use of artifacts, such as dress and cosmetics.

In this thesis, I focus more on paralanguage (e.g., intonation) and facial expressions, relate them with social skills, and motivate the scientific findings and technology development in relation to enhancing human nonverbal skills.

1.2.1 Importance of Nonverbal Behavior in Building Social Skills

Nonverbal behavior plays an integral part in a majority of social interaction scenarios. Being able to adjust nonverbal behavior and influence others' responses are considered valuable social skills (Ferris, Witt, & Hochwarter, 2001). For example, greater interpersonal skills are shown to be a predictor of better performance in school (Halberstadt & Hall, 1980; Nowicki Jr. & Duke, 1994). In the context of doctor-patient interactions, doctors with a higher level of nonverbal skills are perceived as more competent and caring (Blanch, Hall, Roter, & Frankel, 2009). For businesses and customer-facing individuals, nonverbal skills play a role in signaling trustworthiness to boost their business (Wood, 2006).

Studies have found that nonverbal behavior bivariately relates to contextual performance in teamwork (Morgeson, Reider, & Campion, 2005). Teamwork requires coordination skills to share the workload with each other, where nonverbal behavior plays an important role (Campion, Medsker, & Higgs, 1993). Of individuals with high general mental ability, those with higher social skills were associated with higher salaries (Ferris et al., 2001).

Nonverbal behaviors are also indicative of affinity in social relationships. For example, in 1992, Gottman and Buehlman conducted a study in which they interviewed couples with children and analyzed their nonverbal behaviors. They were able to develop a discriminant

function using an *a posteriori* modeling technique that could predict who divorced with 94% accuracy (Buehlman, Gottman, & Katz, 1992). In a follow-up study in 1998, Gottman predicted which newlywed couples would remain married four to six years later (Gottman, 2003). Gottman's latest study in 2000 found that his model had 87.4% accuracy in predicting divorce in the first 5 years of marriage (Carrère, Buehlman, Gottman, Coan, & Ruckstuhl, 2000).

Research has shown that the job interview is an effective construct for evaluating social skills (Allen I Huffcutt, Conway, Roth, & Stone, 2001; Posthuma, Morgeson, & Campion, 2002). Through the employment interview, employers can gauge how potential employees will interact with other employees. Studies found that the most important constructs in employment interviews are "personality traits and social skills" (Allen I Huffcutt et al., 2001). Given the importance of social skills in most customer-facing jobs, big companies like Walmart, Nike, Starbucks, Dunkin' Donuts, and eBay use automated web-based technologies like HireVue ("HireVue," n.d.) and JobOn ("JobOn," n.d.) that require candidates to record answers to interview questions to be assessed by the company later. Using these recordings, employers eliminate unfavorable candidates, often using simple behavioral rules. For example, Holiday Inn was reported to eliminate candidates who smiled less than a given threshold during interviews (LaFrance, 2011). Such practices highlight the changing nature of technology use for assessment in professional placement and underline the growing need for technologies that help people improve their communication skills in such contexts.

1.2.2 Importance of Automated Nonverbal Social Skills Training

Feldman et al. (Feldman, Philippot, & Custrini, 1991a) hypothesize people who are socially deficient may suffer from a lack of nonverbal decoding and encoding skills. One of the health conditions includes Asperger's Syndrome, which is an Autism Spectrum Disorder (ASD) and is characterized by difficulties in social and nonverbal interactions (Attwood & Lorna, 1998). One of the health conditions includes Asperger's Syndrome, which is an Autism Spectrum Disorder (ASD) and is characterized by difficulties in social and nonverbal interactions (Attwood & Lorna, 1998). While people with Asperger's Syndrome have normal or above-average intelligence, they have less well developed social skills (Krieger, Kinébanian, Prodinger, & Heigl, 2012). Along with implications in personal life, social difficulties make it difficult for these individuals to seek employment and retain their job. Howlin (Howlin, 2000) summarizes the relevant literature and concludes that 45-95% of individuals with Asperger's Syndrome have a high probability of being unemployed. According to Muller et al. (Muller, Schuler, Burton, & Yates, 2003), people with Asperger's

Syndrome have trouble adapting to their environment, job routines, and interacting with employers and co-workers.

Another common and debilitating health condition is social phobia— the fear of being judged by others and of being embarrassed. Social phobia afflicts 13% of the population, propagates to major difficulties in professional, academic, and social life, and could result in drug and alcohol addiction in an attempt to self-medicate (Craske, 1999)(Ruscio et al., 2008).

Unfortunately, it is difficult for people to gain access to help due to unavailability of resources, experts, and logistics. Hilton et al. (Hilton, Scuffham, Sheridan, Cleary, & Whiteford, 2008) states that only 22% of U.S. workers with high levels of psychological distress receive intervention. A typical therapy or intervention is conducted weekly or biweekly, and it is difficult for people to take time off from work, arrange child care, and justify a possibly long traveling time (Cartreine, Ahern, & Locke, 2010). Therefore, there is a need for developing automated interventions to enhance human nonverbal behavior that are standardized, objective, repeatable, low-cost, and can be deployed outside of the clinic.

1.3 CAN COMPUTERS RECOGNIZE NONVERBAL BEHAVIORS?

In order to develop automated systems to help people enhance nonverbal social skills, computers need to recognize the nonverbal behaviors. But nonverbal behaviors are known to be subtle, multidimensional, noisy, overlapping, and often contradictory, making it an extremely difficult problem for computers to reliably recognize them (M. Hoque, Mcduff, Morency, & Picard, 2011). Consider the computational challenges involved in understanding and recognizing nonverbal behaviors. Let us examine the game of chess. The first player can open with any of the 20 actions, and the second player can do the same. Therefore, after the first two moves, there are 20 x 20 = 400 branches to specify. Based on those numbers, according to Skeath and Dixit (Dixit & Skeath, 2004), the number of possible moves in chess is on the order of 10^{120}. A computer making a billion calculations a second would take approximately $3x10^{103}$ years to consider all of these moves. Due to the complexity involved in considering all the moves in chess, it has become a common platform for computer theorists to design new optimization algorithms and complexity analysis. Let us contrast the game of chess with nonverbal behaviors. Using the 43 muscles of our face, it is possible for humans to produce approximately 10,000 unique facial expressions in any given time (human timing is in milliseconds) (Ekman & Friesen, 1978). In other words, for two people interacting, the number of possible facial paths after the opening moves would equal 10,000 x 10,000 = 100,000,000. There are also other nonverbal modalities such as prosody, gestures, and body movements. This creates a grand challenge in the field of computing, requiring breakthroughs in multiple areas. Despite the challenges, the affective computing community has come a long way in developing computational frameworks that can model the interplay, redundancy, and

dependency among the affective modalities. This thesis takes an exploratory step in that direction.

1.3.1 Facial Expressions

Paul Ekman and his colleagues first introduced the notion of six basic emotions: surprise, fear, happiness, sadness, anger, and disgust (Ekman & Friesen, 1978). They argue that the basic emotions are the basis for all human emotions. Given the theory and availability of data, most of the previous exploratory studies have attempted to classify so-called "basic emotions" from images and videos ((Keltner & Ekman, 2000), (Juslin & Scherer, 2005) as reported in (Gunes & Pantic, 2010)).

Basic emotions are widely believed to be universally expressed, and their dynamics are typically much stronger than in spontaneous day-to-day facial expressions, which make them a natural place to start training expression recognition systems. Also, given that the majority of the available affective datasets contain basic emotions, it is desirable to work on them towards developing a common benchmark. However, it is also important to push the boundary of understanding the nuances of facial expressions with naturalistic data congruent with relevant tasks. This phenomenon has been echoed by many individuals diagnosed with autism who find it difficult to apply universal emotions to more naturalistic expressions. For example,

> *...She said that she could understand "simple, strong, universal" emotions but was stumped by more complex emotions and the games people play. "Much of the time," she said, "I feel like an anthropologist on Mars." .. She has instead to "compute" others' intentions and states of mind, to try to make algorithmic, explicit, what for the rest of us is second nature.*
>
> *-Interview with Temple Grandin, by Oliver Sacks* (Sacks, 1995)

1.3.2 Prosody

Studies suggest that how we say things is as important as what we say (Albert Mehrabian, 2008) (Judee K Burgoon & Buller, 1994) . The phenomenon of adding extra information by varying the rhythm and melody of spoken language is called *prosody* (Hirst & Di Cristo, 1998)(Cutler, Dahan, & Van Donselaar, 1997). The features of prosody include pitch (F0) (slope, contour type, discontinuity), speaking rate (phoneme duration, lengthening, pause behavior), loudness (energy, perception), voice quality (used for emotion). Our ability to respond to prosody is innate; even newborns can extract and respond to prosody from spoken words (Sambeth, Ruohio, Alku, Fellman, & Huotilainen, 2008).

Prosody is pervasive. It can convey meaning at many different levels. For example, Bolinger (Bolinger, 1989) states that using prosody we could change our focus (Turn right |

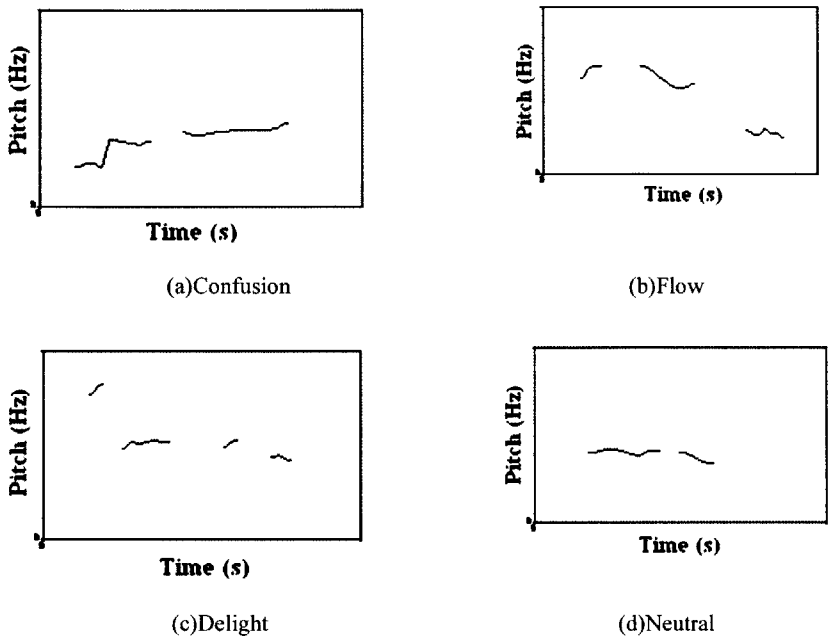


Figure 1-1. Pictorial depiction of the word "OK" uttered with different intonations to express different emotions (a) Confusion, (b) Flow, (c) Delight, (d) Neutral

Turn RIGHT), syntax and phrasing (Turn right. | Turn. Right.), pragmatic functions (Turn right. | Turn right?), and affect and emotion (annoyance, panic). Even the two letter word "OK" can have several meanings based on its prosodic contour (M. E. Hoque, Yeasin, & Louwerse, 2006), as shown in Figure 1-1.

1.3.3 Speech

Speech is a very important modality in the context of Human-Computer Interaction. Spoken language is meant to be heard, not read. When speech is automatically recognized by an Automated Speech Recognizer (ASR), it is represented with text, and thus loses its dynamicity, structure (punctuation, capitalization, formatting), and, most importantly, the tone. In this thesis, I present a framework that is able to capture the relevant nuances from spoken language in real-time.

1.4 AFFECTIVE FRAMEWORK AND NONVERBAL SKILLS

Consciously or unconsciously, humans exhibit numerous nonverbal behaviors using vocal nuances, facial expressions and body postures to communicate their intentions and feelings. During face-to-face interactions, those behaviors take precedence over the spoken words.

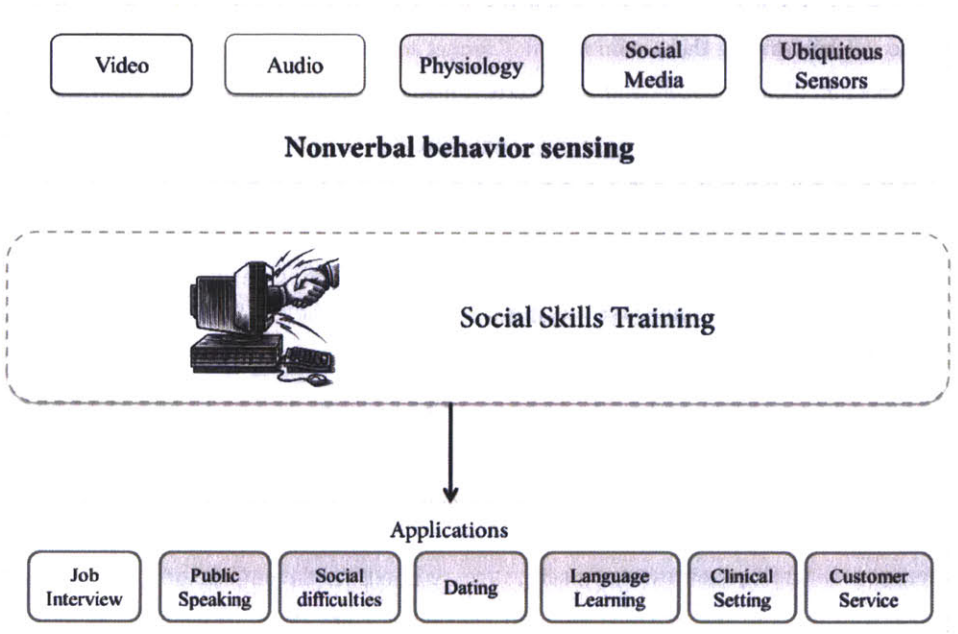


Figure 1-2. Outline of thesis aims and future possibilities. Future possibilities are displayed in shades.

An automated system that can capture and interpret the full range of nonverbal cues, and allow humans to practice and enhance their nonverbal skills, remains an open problem. To develop such a framework is an ambitious undertaking.

In this thesis, I present new studies on understanding expressions that extend explanations to the existing theory as well as contradict it. I also demonstrate how the knowledge generated could be applied towards automated classification of affective states offering new insights and interaction possibilities to enhance human nonverbal behaviors (Figure 1-2).

1.5 AIMS AND CHALLENGES

The sensing of nonverbal behavior needs to go beyond the basic emotions, and interpret and incorporate more difficult subtle and naturalistic behavior. What are the elements of spontaneous datasets that make them so much more difficult? How are they different? Is there a surprising element present? Can we go beyond the naïve interpretation of "a customer is smiling, therefore, the customer must be satisfied" and automatically infer the underlying meaning behind expressions? This thesis takes on these questions through explorations of new studies, dataset, and algorithm development.

For a computer, nonverbal behaviors are nothing but a multidimensional array of numbers. Representing those numbers in a format to the users, with little or no experience, and expecting them to understand and interpret them is still an open problem. In this thesis, I present iterations with real users towards designing the final interface for users to understand and reflect on their nonverbal behaviors.

Human interpretation of nonverbal behaviors is subjective. The behaviors are dependent on the application and individual interpretation. Computation, on the other hand, is objective. Modeling the subjective judgment of nonverbal behavior for a given context using objective data remains an open problem. In this thesis, I focus on developing technology and designing an intervention that allows users to combine human subjectivity with automatically computed objective nonverbal data. For example, first, I design an intervention allowing users to interact with a professional career counselor, in the context of a job interview, to obtain feedback on their nonverbal behavior. Second, the users were invited to interact with an automated system on their own as much as they wished, for the purpose of reflecting on the system's feedback to enhance their nonverbal behavior. Third, the users would come back and interact with the same counselor in the job interview scenario. This approach allows users to engage with an objective automated system, using subjective expert opinion from the counselor as a guideline. This method also makes possible a comparative behavior measure of the user as they interact with the human expert before and after using the system, a study we conduct in this thesis.

The overall aim for this thesis is twofold: first, to expand our theoretical knowledge of nonverbal behavioral understanding and sensing; second, to push the boundary of real-time nonverbal behavior recognition by creating an interactive system to help people enhance their nonverbal skills.

In this thesis, the nonverbal sensing is focused on audio and video only, with future plans to incorporate other sensors. Towards developing a proof-of-concept, an interaction scenario of job interviews has been defined. The goal is to determine whether it is possible for humans to enhance their nonverbal skills through interaction and affective feedback through an automated system. Future interaction possibilities include public speaking, social difficulties, dating, language learning, medical clinical settings, customer service, and more.

1.6 THESIS CONTRIBUTIONS

The contributions of this thesis intersect fields of human-computer interaction (HCI), computer vision, machine learning, multimodal interfaces, and experimental psychology. The contributions are theoretical, methodological, and practical.

- I design a new interaction scenario with computers that could provide affective feedback to enhance human nonverbal skills. To further the concept, I developed a system called "MACH: My Automated Conversation coacH" consisting of a 3D character that is able to "see," "hear," and "respond," in real-time using a webcam and a built-in microphone. MACH is a real-time framework that can recognize spoken utterances (along with their prosodic contours) and facial expressions (smiles and head gestures). In addition, MACH uses the sensed data to drive the behaviors and actions of a conversational virtual agent and provide naturalistic affective feedback within a given context.

- I present computational evidence of humans demonstrating facial signatures of delight while experiencing frustrating stimuli, contradicting the popular belief that "true smiles" mean you are happy. An example is shown in Figure 1-3.

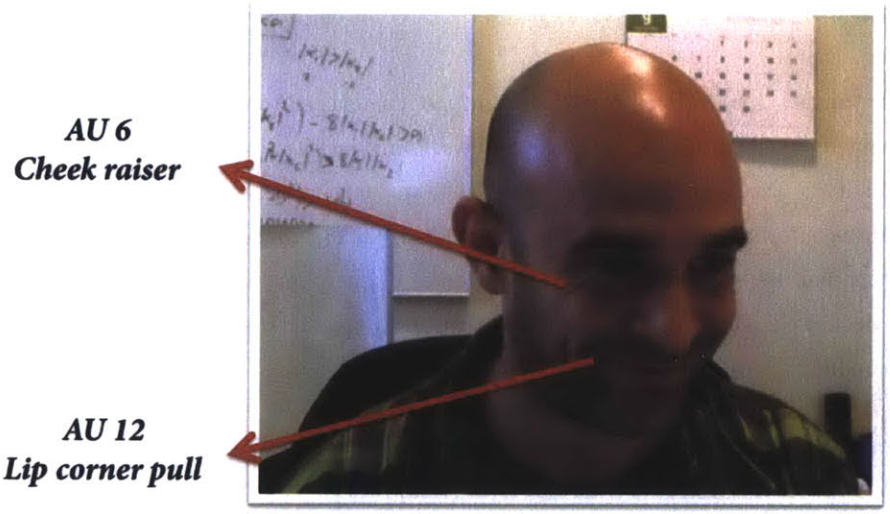


Figure 1-3. A participant demonstrating an expression usually associated with "delight" (AU 6 and AU 12) while experiencing a frustrating stimulus.

- I present new findings on morphological and dynamic properties of polite and amused smiles in the context of natural face-to-face interactions, confirming and furthering the existing literature of smiles. I analyze naturally occurring instances of amused and polite smiles in the context of banking, noting also if they were shared, which I defined as the rise of one starting before the decay of another. My analysis confirms previous findings showing longer durations of amused smiles, while also suggesting new findings about the symmetry of smile dynamics. I report more symmetry in the velocities of the rise and decay of the amused smiles, and less symmetry in the polite smiles. I also find the fastest

decay velocity for polite but shared smiles. These findings are steps towards creating a taxonomy of smiles and other relevant expressions.

- I have used and validated a machine learning algorithm that can model the temporal patterns of smiles and can classify different smiling instances as accurately as humans, or better. The study reveals that the happy smiles build up gradually, while frustrated smiles appear quickly but fade fast. This work opens up possibilities of providing better ways to help people disambiguate similar expressions with different meanings.

- I develop and evaluate a novel data visualization that automatically gathers and represents multidimensional nonverbal cues in an educational and intuitive interface. The interface allows users to understand, reflect on, and interpret their behaviors. The interface was designed using a user-centric iterative approach to automatically capture the nonverbal nuances of interaction.

- I validate the MACH framework in the context of job interviews with 90 MIT undergraduate students. I designed a weeklong intervention to measure the effect of MACH. The findings indicate that MIT students using MACH being perceived as stronger candidates compared to the students in the control group. The results were reported based on the judgments of the independent MIT career counselors and Mechanical Turkers, who did not participate in the study, and were blind to the study conditions.

1.7 OTHER RELEVANT CONTRIBUTIONS

As a Ph.D. student, I have worked on other projects that indirectly influenced my PhD dissertation. Here, I provide short descriptions of my contributions to those projects.

1.7.1 Speech Games for Children on the Autism Spectrum

I developed rapidly customizable speech-enabled games to help people with speaking deficiencies (e.g., speaking rate, loudness). The platform allowed participants to control the characters of the games by modifying certain properties of their speech. The games were made open source and are freely available for people who may not otherwise have the ability to spend $2000-4000 per license (as sold by Visi-Pitch™) for games with fewer features and no customization capability. As part of a team, I ran a 6-week-long intervention with eight participants to demonstrate that a subgroup of individuals on the autism spectrum can benefit from computerized therapy sessions in addition to or in place of conventional therapy — a

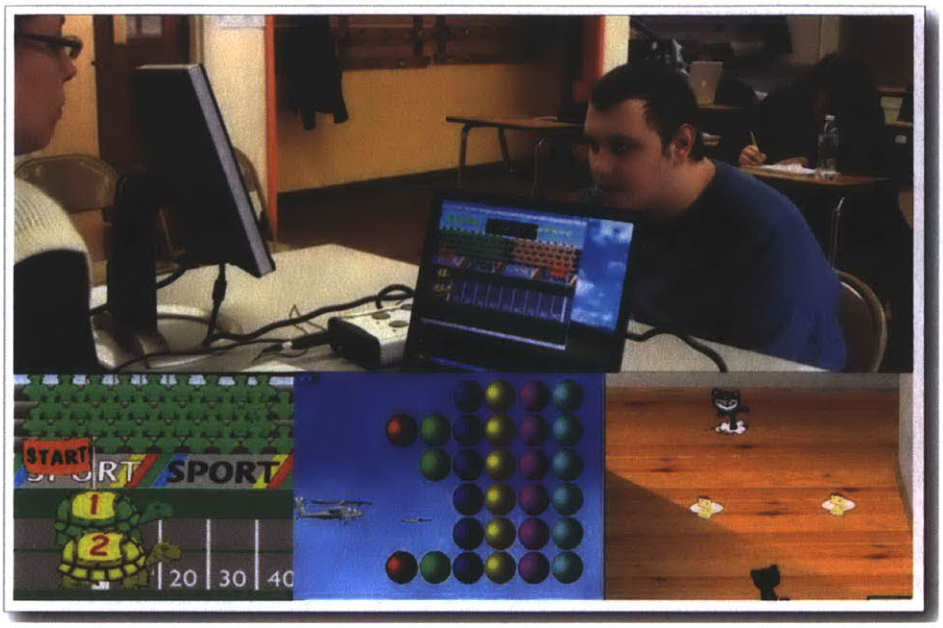


Figure 1-4. A session between a speech therapist and a participant with speech difficulties. The laptop's screen is being projected into the other monitor, so that the participant sees what the therapist sees, while sitting face-to-face. The bottom half of the screen shows the screen shots of speech-enabled interactive games.

useful option for individuals who do not have regular access to speech-language therapists. This work has resulted in a publication (M. E. Hoque, Lane, Kaliouby, Goodwin, & Picard, 2009).

1.7.2 MIT Mood Meter

I designed the computer vision system of "MIT Mood Meter" that counted smiles of people in four different public spaces of MIT using cameras. By sensing data in real-time, we presented a map of MIT as a heat map of smiles, refreshed every 2 seconds on the World Wide Web. With data collected for more than 10 weeks, we were able to objectively answer questions such as whether students from one department smile more than another, and whether people smile less during midterms. This was one of the pioneering examples of testing computer vision algorithms through a real-world deployment and collecting data to draw new insights. This work has resulted in a publication (Hernandez, Hoque, Drevo, & Picard, 2012).

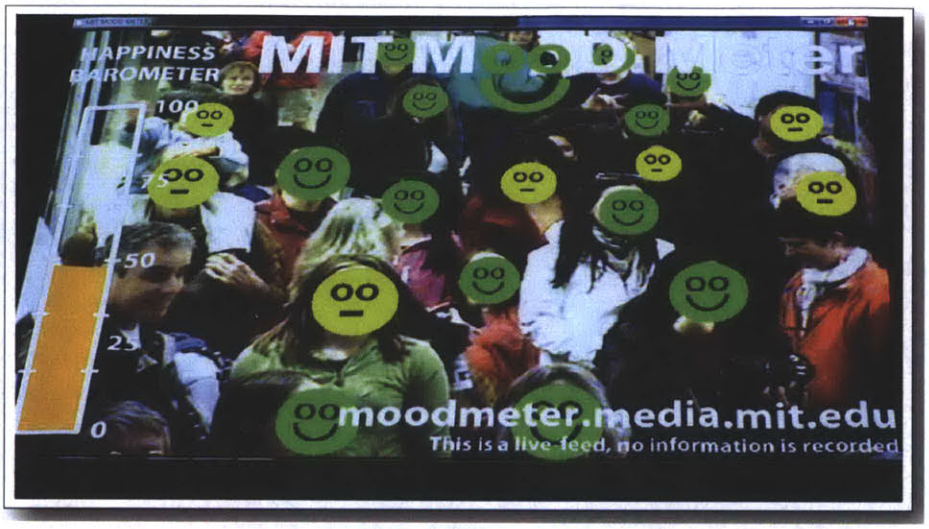


Figure 1-5. The interface of the MIT Mood Meter which was deployed at 4 different locations of MIT for 10 weeks.

1.8 THESIS OUTLINE

The outline for the remainder of the thesis is as follows.

Chapter 2: Background

The research described in this dissertation draws inspiration from several areas of research, including nonverbal behaviors, social/technical interventions, nonverbal behavior sensing, nonverbal data visualization, and conversational virtual agents. This chapter presents theoretical background and describes the relevant work, highlighting the existing lack of technology to provide automated nonverbal sensing with feedback to humans in an agent-human interaction and the application of such technology to training scenarios.

Chapter 3: Job Interview – An Interaction Scenario

This chapter uses a job interview as a possible scenario of helping people with their nonverbal social skills. Among most of the social interactions, job interviews follow a predictable turn-taking structure which is easier to model. It describes a job interview study with 28 interview sessions, involving employment-seeking MIT students and career counselors, and its outcome. The analysis of the study informs the contextual inquiry of nonverbal behavior modeling of the interviewer as well as the structure of the feedback provided to the interviewee.

Chapter 4: Technology Design, Development, and Explorations

This chapter explores technologies that can automatically capture nonverbal behaviors and use the sensed data to drive the behaviors of the conversational virtual agent. As part of the exploration, it describes a series of studies investigating the morphological and signal-level differences of smiles expressing frustration, delight, or just politeness. This chapter describes a machine learning algorithm that can model the temporal patterns of the smiles and predict its underlying meaning as good as humans or better. It also provides details on the relevant prosodic properties of speech, their characteristics, and the challenges of automatically computing them.

Chapter 5: Training Interfaces for Social Skills

This chapter describes the process that I followed to represent nonverbal behavior as part of training interfaces. I present prototypes and user evaluations of four successive interfaces informed by our intuitions, previous work, and user studies. This chapter contains the "lessons learned" on each iteration, and describes how I continuously incorporated the user feedback to inform the design of the next prototype.

Chapter 6: My Automated Conversation coacH (MACH)

In this chapter, I bring together theoretical knowledge and technology developed from the exploration and user-centric training user interfaces to form an automated affective framework, My Automated Conversation coacH (MACH). MACH consists of a conversational virtual agent that is designed to enhance the nonverbal skills of users in the context of job interviews. The chapter ends with a description of the technical modules contributing to the behavior synthesis aspects of MACH.

Chapter 7: Evaluation of MACH in the Job Interview Scenario

This chapter describes the experimental set-up that was designed to validate the effectiveness of MACH. There were 90 MIT students who interacted with MACH in the context of a job interview. The chapter reports the summary of the results from four different measures along with a discussion.

Chapter 8: Discussion, Conclusions and Future Directions

This chapter summarizes the main contributions of this thesis and highlights some remaining challenges and limitations.

Chapter 2: Background

This thesis draws upon many areas, including nonverbal behaviors, behavioral interventions, automated nonverbal sensing, visualization of nonverbal data, and conversational virtual agents. This chapter provides theoretical background along with relevant technical details in all those areas. This chapter highlights the fact that while nonverbal behavior is extremely useful for interlocutors in many face-to-face interaction scenarios, the rules for nonverbal behavior could conflict depending on the context.

A literature review provides evidence that behavioral interventions to improve nonverbal social skills confront many challenges, including ensuring that the skill generalizes beyond the given task. In addition, this chapter describes the limitations of the existing affective multimodal systems in terms of sensing that allows users to understand, interpret, and enhance their nonverbal behaviors in a given interaction scenario; and it highlights the opportunity to contribute to this research space.

2.1 SOCIAL NONVERBAL SKILLS

2.1.1 What Are Nonverbal Behaviors?

A majority of the information during face-to-face interactions gets communicated nonverbally (Birdwhistell, 1970; Judee K Burgoon & Buller, 1994; A Mehrabian, 1968). My working definition in this thesis for "nonverbal behaviors" includes all the communicative modalities expressed from a person that are not verbal. However, given the large body of literature in this research space, there exist many definitions of nonverbal behaviors. Below I provide a summary.

According to Argyle (Argyle, 1969), nonverbal behavior includes bodily contact, posture, physical appearance, facial and gestural movement, direction of gaze and paralinguistic variables of emotional tone, timing, and accent. Duncan (Duncan Jr. Starkey, 1969) divides nonverbal communication into five components: 1) body motion or kinesic behavior (gesture and other body movements, including facial expressions, eye-movement, and posture); 2) paralanguage (voice qualities, speech non-fluencies, and such non-language sounds as laughing, yawning and grunting); 3) proxemics (use of social and personal space); 4) olfaction; 5) skin sensitivity to touch and temperature; and 6) use of artifacts, such as dress and cosmetics. Knapp (Knapp, 1972) includes body motion, or kinesic behavior, facial expression, physical characteristics, eye behavior, touching behavior, paralanguage, artifacts, and environmental factors as nonverbal behaviors. Harrison et al. (Harrison, n.d.) divide the

nonverbal behavior domain into four codes: performance codes based on bodily actions; artifactual codes (the use of clothing, jewelry, etc.); mediational codes involving manipulation of the media; and contextual codes such as employment of nonverbal signs in time and space. Harper et al. (Harper, Wiens, & Matarazzo, 1978) narrow their consideration of nonverbal phenomena to those that are the most relevant to the structure and occurrence of interpersonal communication, and the moment-to-moment regulation of the interaction. In summary, most of the definitions generally revolve around body area and body activities. Several of them mention artifacts. Most of the definitions focus on the use of paralanguage and manipulation of facial expressions under nonverbal behavior. In this thesis, I focus more on paralanguage (e.g., the intonation) and facial expressions, and motivate the rest of the thesis based on those two modalities.

2.1.2 The Codebook of Nonverbal Behaviors

> *We respond to gestures with an extreme alertness and, one might almost say, in accordance with an elaborate and secret code that is written nowhere, known by none, and understood by all. – Edward Sapir (1927)*

As Sapir observed, nonverbal behaviors are ubiquitous but often not understood. The behaviors are known to be subtle, fleeting, contextual, and often contradictory. Consider speaking rate as an example. There are no strict rules on how fast one should speak to be effective. A rapid speaking rate could be interpreted as someone having a lot of confidence, or a nervous speaker rushing to conclude. Similarly, people who speak slowly may be regarded as having high status or being knowledgeable. There are stereotypes with voice pitch as well. One may associate a deep-pitched voice with a big male and a high-pitched voice with a small female. Pausing may be a sign of embarrassment or oppression, but it can also be a sign of respect or building up of suspense.

Vocal cues are an example of nonverbal behavior that can express information about various features of the speaker such as physical characteristics, emotional state, and personality (Kramer, 1963). Vocal cues can be persuasive for children as young as 12 months old (Vaish & Striano, 2004). Varying pitch and volume can help clarify messages, pose questions, or establish the ideal tone. Vocal cues can also express negative traits. Vocalized pauses or fillers such as "ah" or "um" are distractions that may express nervousness or lack of sincerity of the speaker. The meaning of nonverbal behaviors could be strengthened or contradicted when they are combined with multimodality. For example, leaning backward with an absence of smiles is considered more disengaged than just absence of smiles.

Nonverbal behavior can also act as relational communication. One's nonverbal cues can show dominance or intimacy regarding a relationship—how one feels about the relationship with the person with whom one is communicating. Dominance is characterized by "exerting control over conversational distance," "less smiling," and "greater postural relaxation" (J. K. Burgoon, 1985). In contrast, non-dominance has been characterized by more smiling, laughing, and shrugging of shoulders (J. K. Burgoon, 1985). Mignault & Chauduri mentioned that dominance or non-dominance could be conveyed with subtle changes in head movement (Mignault & Chaudhuri, 2003). For example, the head being directed upward can express dominance, while upward head tilts can convey non-dominance (Mignault & Chaudhuri, 2003). Carney et al., associated dominance with larger gestures (Carney, Hall, & LeBeau, 2005). Intimacy can also be conveyed through nonverbal behavior such as close physical proximity, open body posture, and eye contact(Andersen & Sull, 1985).

Researchers have also reported gender effects with nonverbal behaviors. For example, Mast and Hall found that women exhibit higher status with a downward head tilt, while males exhibit higher status with "formal dress and a forward lean" (Mast & Hall, 2004).

Nonverbal behavior is expanding to new areas such as mediated communication as more people begin to interact online as opposed to face-to-face. There is a transition from text-based computer-mediated communication to computer-mediated interaction given the emergence of Internet-based 3D virtual environments and avatars to represent users. This transition challenges the prevailing theories and definitions of nonverbal behavior for the analysis of computer-mediated interaction (Antonijevic, 2008). For many, attractiveness has become a function of online representation and expressions, and less about physical attributes. Many believe that online nonverbal behavior will have an impact on attraction in the future (Levine, 2000).

Nonverbal behaviors contribute to developing rapport in face-to-face interactions (Tickle-Degnen & Rosenthal, 1990). Rapport is the basis of understanding and communicating with coworkers and business partners. People can use nonverbal behaviors such as a smile, an open stance, and eye contact to establish rapport. Another technique for nonverbal communication that helps build rapport is mirroring. By paralleling a conversation partner's breathing, speaking rate, tone of voice, body language, mirroring can enhance communication and rapport.

2.1.3 Why Are Nonverbal Behaviors Important?

Understanding nonverbal behavior and being able to respond to it are considered very important aspects of communication (J. A. Hall & Bernieri, 2001a). It has been hypothesized by Feldman et al. that nonverbal communication is a form of social skill that is prevalent among people who are viewed as socially competent (Feldman, Philippot, & Custrini, 1991b).

People who are sensitive to nonverbal behaviors are perceived as more empathetic and tolerant. They appear to have high levels of engagement, emotional competence, skills, and to be more likely to have succeeded as supervisors, negotiators, or sales-persons [62].

Greater interpersonal skills are also shown to be a predictor of better performance in school (Halberstadt & Hall, 1980; Nowicki Jr. & Duke, 1994), (Kupersmidt, Coie, & Dodge, 1990; Ladd, 1990; Parke et al., 1997; Patrick, 1997; Ray & Elliott, 2006; Welsh, Parke, Widaman, & O'Neil, 2001; Wentzel, 1991a, 1991b)). A relevant study conducted on Hispanic teenagers in the United States demonstrated that their ability to understand nonverbal elements—due to their lack of fluency in English—is boosting their academic self-confidence, and as a result their school performance as well ((Acoach & Webb, 2004), cited by (Ivan & Duduciuc, 2011)).

In the context of doctor-patient interactions, doctors who are sensitive to nonverbal skills are perceived as more competent and caring. For example, Hall et al. demonstrated that future doctors who had received higher scores on standardized tests were perceived as more engaged, less distressed, and received a higher rating of service during a standard patient visit (Blanch et al., 2009). For a detailed discussion, see Hartigan (Hartigan, 2011).

Nonverbal skills also play a role in signaling trustworthiness in face-to-face interactions. Wood et al. highlighted the importance of the nonverbal communication between the sender and the receiver in the context of sales-people and customers (Wood, 2006). Wood et al. argue that customers rely heavily on nonverbal signals to assess the trustworthiness of sales-people. Thus, by focusing on the nonverbal behaviors, it is possible for business professionals to signal their trustworthiness and boost their business.

Nonverbal behaviors are also indicative of affinity in close relationships. For example, Gottman et al. studied verbal and nonverbal behaviors of married couples for many years (Gottman, Markman, & Notarius, 1977). In one of their studies, couples were asked to discuss and resolve a marital problem in front of a camera. Their facial, vocal, and bodily actions were manually coded by the researchers. Positive items included empathetic face, nodding, eye contact, a warm voice, leaning forward, and touching, whereas negative items included frowns, sneers; a cold, angry, whining or sarcastic voice; inattention; and hand movements such as points, jabs, and slices. The study concluded that couples under stress would verbally express agreement with the spouse, while demonstrating *channel inconsistency* by their negative nonverbal behavior. For a detailed overview of the role of nonverbal behavior in marriage, please see (Kluft, 1981).

The importance of social skills and their development is recognized beyond clinical research and practice, and includes areas such as professional development and placement. Research has shown that the interview is an effective construct for evaluating social skills [16],[17]. Through the employment interview, employers can gauge how potential employees

will interact with other employees. Studies found that the most important constructs for employment interviews are "personality traits and social skills" (Allen I Huffcutt et al., 2001).

Lack of nonverbal skills is also an important area of research. Feldman et al. (Feldman et al., 1991a) hypothesize that those who are socially deficient might suffer from a lack of nonverbal decoding (being able to understand others' cues) and encoding skills (being able to exhibit cues to others). This deficiency could lead to anxiety, depression, and lower self-esteem (Buss, 1989; McClure & Nowicki, n.d.; Nowicki & Carton, 1997; Nowicki & Mitchell, 1998), antisocial behaviors, increased incarceration, and symptoms of psychopathy (Hastings, Tangney, & Stuewig, 2008; Marsh & Blair, 2008). Given the importance of nonverbal behaviors, many social and computational techniques have been developed and researched towards enhancing human nonverbal behaviors.

2.2 SOCIAL BEHAVIOR INTERVENTIONS

Most of the previous research on designing counseling or social interventions included helping people diagnosed with social difficulties. One such condition is Asperger's Syndrome, which is an Autism Spectrum Disorder (ASD) and is characterized by difficulties in social and nonverbal interactions (Attwood & Lorna, 1998). People with Asperger's syndrome are known to suffer from lack of social skills. Another debilitating health condition is social phobia— the fear of being judged by others and of being embarrassed. Social phobia affects 13% of the population, and can contribute to difficulties in professional, academic, and social life, resulting in drug and alcohol addiction in an attempt to self-medicate (Craske, 1999)(Ruscio et al., 2008). Given the severity of conditions in Asperger's Syndrome and social phobia, most of the social interventions were designed to specifically benefit them.

There are several challenges in designing social interventions. First, people's sensitivity to nonverbal cues varies (Rosenthal, Hall, Matteg, Rogers, & Archer, 1979). Accordingly, response to the interventions is less likely to be consistent. Second, learned skills could be difficult to generalize beyond the training scenarios such as playground, lunchroom, classrooms, and or anywhere, with individuals not part of the training (Wilczynski, Menousek, Hunter, & Mudgal, 2007). Below are a few techniques that have been widely used by researchers, teachers, and therapists as social interventions.

2.2.1 Social Stories

Social Stories are short individualized stories designed to teach students how to behave appropriately in different social situations (McConnell, 2002).. The stories are presented to a student as a priming technique before the student actually practices the behavior. Social stories seem to be increasing in popularity as a child-specific intervention for students with

social difficulties. However, more research is needed to measure the overall effectiveness of Social Stories.

2.2.2 Social Scripting or Fading

Social scripting or fading is a process intended to increase students' social skills (Brown, Krantz, McClannahan, & Poulson, 2008)(Ganz, Kaylor, Bourgeois, & Hadden, 2008). Students are given scripts in a training environment and eventually are expected to begin using the scripted language in real-world interactions. While studies have shown that social scripts/fading could be used to increase social skills among students with autism, social scripts/fading has not had any effect on skills outside of the intervention environment (Ganz et al., 2008).

2.2.3 Grouping Interventions With and Without Typically Developing Pairs

Group Interventions provide a controlled environment to practice social skills (Krasny, Williams, Provencal, & Ozonoff, 2003). Students with social difficulties are taught social skills outside the general classroom environment in Group Interventions, which allows them to interact with other students with similar diagnose or with typically developing peers.

2.2.4 Integrated Play Groups

In Integrated Play Groups, adults mediate play between a student with ASD and their peers (Bass & Mulick, 2007)(DiSalvo & Oswald, 2002). The aim of Integrated Play Groups is to increase interaction and play between students with ASD and their peers. While no social skills are taught during an Integrated Play Group, the adult encourages using appropriate play skills.

2.2.5 Peer Tutoring

In Peer tutoring, social skills are taught to children in group and individual settings (DiSalvo & Oswald, 2002)(Rogers & Bennetto, 2000). Children are taught a variety of social skills such as how to initiate play or ask for help.

2.2.6 Peer Buddy

Peer Buddy interventions involve a typically developing peer helping a child with ASD to develop social skills. The typically developing peer plays and talks with the child with ASD and stays in the same playing area. Studies have shown that peer buddy interventions are effective in increasing ASD students' social skills (DiSalvo & Oswald, 2002)(McEvoy, Odom, & McConnell, 1992)(Rogers & Bennetto, 2000).

2.2.7 Video Modeling

In Video Modeling interventions, videos of desired social behaviors are shown to children with ASD. The goal is that the child with ASD will learn and imitate the desired social behaviors from the videos (Bellini, Akullian, & Hopf, 2007)(Buggey, 2005).

Video modeling has been found to be an effective way to treat patients with social phobia, by allowing them to directly observe themselves. Initially, some patients were unable to view their own video as an objective observer due to re-experiencing feelings they had during the experiment. In order to solve this problem, patients were asked to visualize how they appear in the video, ignore these feelings, and observe themselves from a stranger's perspective. As a result, patients were able to notice significant discrepancies between the video and their self-image (MacDonald, Clark, Garrigan, & Vangala, 2005).

Video Modeling has been shown to have an effect on increasing social interactions, but more research is required (Bellini et al., 2007)(Buggey, 2005)(MacDonald et al., 2005).

2.2.8 Class-wide Interventions

Class-wide Interventions involve the classroom teacher teaching social skills to the entire class, aiming to increase the skills of the child with ASD (Pollard, n.d.). Studies on class-wide interventions have been limited due to the difficulty in gaining information on effectiveness and lack of training programs.

2.3 COMPUTATIONAL INTERVENTIONS

Computerized interventions have also been widely used to help people with social difficulties. Researchers have given special attention to computer-based interactions for people with ASD (Kanner, 1943) and social phobia due to their predictability, safety, and specificity (Murray, 1997). Researchers emphasize the value of computer-based activities because they can be therapeutic and educational for people with social difficulties (Grynszpan, Martin, & Nadel, 2005)

2.3.1 Cognitive Bias Modifications

Researchers have also explored the use of computerized interventions in treatments for social anxiety. For example, Beard (Beard, 2011) demonstrated that, using a cognitive-bias modification, patients with social anxiety disorder exhibited significantly greater reductions in levels of social anxiety compared to patients in the control group. In addition, during the four-month follow-up, the patients who underwent the intervention continued to maintain their clinical improvement and diagnostic differences. This finding is encouraging, as one of the major challenges of automated behavioral intervention is to ensure that skills generalize and persist beyond the intervention duration.

2.3.2 Computer-Aided Psychotherapy

Meta-analytic research on computer-aided psychotherapy (CP) has shown that CP is just as effective as face-to-face psychotherapy (Cuijpers et al., 2009). In the study, researchers compared CP to non-CP in various anxiety disorders. CP provides many advantages, such as saving time in traveling and in therapist sessions, timely access to care, and accessibility through computers, phone-interactive voice response, DVDs, or cell phones. A meta-analysis of 23 randomized-controlled studies found that there was no significant difference between CP and face-to-face psychotherapy, which warrants that CP be used in routine psychotherapy practice (Cuijpers et al., 2009).

2.3.3 Virtual Reality Cognitive-Based Therapy

To help people with public speaking anxiety, Wallach et al. explored Virtual Reality cognitive-based therapy (VRCBT) to help people get over their anxiety (Wallach, Safir, & Bar-Zvi, 2009). Participants were randomly split into VRCBT (29 participants), cognitive-based therapy (CBT; 30 participants), and wait list control (WLC; 30 participants). Their results indicate that participants in VRCBT and CBT were significantly more effective than the participants in WLC in overcoming anxiety on certain dimensions, based on their self-reported measures. No significant differences were observed based on the ratings of independent observers. However, twice as many participants dropped out of CBT as from the VRCBT, providing further evidence that VRCBT could be as effective as CBT as a brief treatment regimen.

2.3.4 Use of Robots to Enhance Social Understanding

The Aurora project investigated the use of robots to enhance the everyday lives of autistic children, concentrating on narrative comprehension and social understanding (Billard & Dautenhahn, 1999). Results suggested that instead of being afraid of the robot, they felt motivated to interact with it, especially in the "reactive" model. Despite the large individual differences among the children, the ability of the robot to engage with autistic children in important social interactions was encouraging (Billard & Dautenhahn, 1999).

2.3.5 Virtual Environments (VE) for Social Skills Interventions

Another area of research for social skill interventions is virtual learning environments. The growing development of virtual environments has generated optimism on the possibility of transferring skills developed during the training to real-life interactions for individuals with ASD (Strickland, Marcus, Mesibov, & Hogan, 1996).

Similarly, there has been research done on the effectiveness of desktop VE's to improve social skills of people with ASD (Mitchell, Parsons, & Leonard, 2007). Performing a

qualitative study with two individuals with ASD, researchers report that even though there were concerns over repetitive behaviors, literal interpretation of the scenes, and VE being perceived as having no real-world relevance, interviewing the participants revealed something else. Participants mentioned that they were able to interpret the scenes meaningfully and enjoyed having a facilitator on the side to discuss appropriate social responses. They also provided specific examples of how the training with VE could possibly help in other social scenarios.

2.3.6 Computer Animations and Interactivity

Animations have been shown to be beneficial to students with special needs. Bernard-Opitz et al. posed eight social problems on a computer to eight typically developing and eight autistic children (Bernard-Opitz, Sriram, & Nakhoda-Sapuan, 2001). Examining the children's ability to produce solutions to a problem revealed that even though the autistic children provided fewer alternative solutions than their peers, animations increased the ability of the children to produce solutions to social problems (Bernard-Opitz et al., 2001).

Mayer and Chandler tested the possibility of benefits of using a computer-user interaction in computer programs (Mayer & Chandler, 2001). In the study, the same multimedia presentation was given to the participants in two groups. One group was given a continuous presentation while the other group could control the pace of a segmented presentation. Each segment included 8 to 10 seconds of animation. In subsequent tests of problem-solving, students who received the segmented, interactive presentation performed better than those who received a continuous presentation. In conclusion, the study found that interactivity can aid deeper learning (Mayer & Chandler, 2001).

Another study by Lahm aims to identify specific design features used in computer programs that aid learning (Lahm, 1996). In 48 trials, the study found that children had a preference for computer programs with higher levels of user-interaction (Lahm, 1996). In addition, the study also found that children preferred computer programs that used animation, sound, or voice features (Lahm, 1996). However, the research was unable to conclude that the participants gained social benefits or that animations improved the social skills of autistic children.

2.4 NONVERBAL BEHAVIOR SENSING

Real-time nonverbal sensing, interpretation, and representation involve significant technical challenges in affective computing. This area of research not only holds promise to reshape the ways we interact with machines today, but also helps us to think of innovative ways to help people with communication difficulties. However, nonverbal behaviors come in many varieties; some are intense and continual, while others are subtle and momentary

(Gunes & Pantic, 2010). Therefore, developing a computational model that can capture all the intrinsic details of human nonverbal behavior would require spontaneous examples of naturally elicited training data containing all the inherent details so that the model can learn from it.

2.4.1 Facial Expressions Processing

Basic Emotions and Beyond

Given the difficulty of collecting spontaneous natural data, most of the previous exploratory studies have attempted to classify so-called "basic emotions" (anger, disgust, fear, happiness, sadness, and surprise) from images (Keltner & Ekman, 2000) and videos (Juslin & Scherer, 2005) (as reported in (Gunes & Pantic, 2010)). Basic emotions are generally defined with Facial Action Coding Units (FACS), a technical guide that defines how to catalogue facial behaviors based on the muscles that produce them (Ekman & Friesen, 1978). Basic emotions are widely believed to be universally expressed, and their dynamics are typically much stronger than in spontaneous day-to-day facial expressions, which make them a natural place to start training expression recognition systems. Also, given that the majority of the available affective datasets contain basic emotions, it is desirable to work on them towards developing a common benchmark. As a result, there has been a trend to correlate FACS with affective states (e.g., associating "delight" with Action Unit 6 -cheek raise, and Action Unit 12 - lip corner pull). Most of that work has been conducted on posed or acted expressions or on expressions observed under less than spontaneous environments. Therefore, there is a need to experiment to see whether those findings hold when tested with more naturalistic, spontaneous, and congruent data.

Datasets

One of the major challenges in affect recognition is collecting datasets, which can be difficult, time-consuming and expensive to construct. In the past, there have been efforts to collect spontaneous sequences of basic and natural emotions while the participants were acting, reacting, or interacting. A few examples of such datasets include Rutgers and UCSD FACS database (RU-FACS) (Bartlett et al., 2006), Sensitive Artificial Listener (SAL) (Schroder et al., 2011), Spaghetti (Schroder et al., 2011), SEMAINE (McKeown, Valstar, Cowie, & Pantic, 2010), Mind-Reading (Baron-Cohen, 2004), and M&M Initiative (MMI) (M Pantic, Valstar, Rademaker, & Maat, 2005). Figure 2-1 demonstrates a graphical representation of each dataset in terms of whether it is acted vs. spontaneous and whether it contains basic vs. beyond basic emotions. It would be ideal to use a dataset that contains spontaneous natural emotion for affect analysis, and includes more than basic emotions, as shown in Figure 2-1.

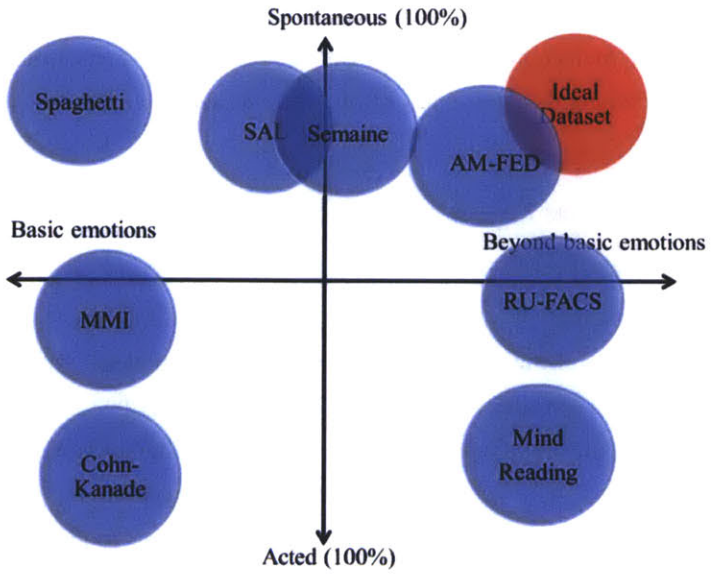


Figure 2-1. Comparison of existing datasets (in the last 7 years) in terms of spontaneous vs. acted and basic vs. beyond basic. An ideal dataset would be spontaneous and contain a complete set of expressions.

One of the early examples of facial expressions dataset includes Cohn-Kanade dataset (Kanade, Cohn, & Tian, 2000). The dataset includes 486 sequences of images of basic emotions from 97 posers. MMI is another publicly available dataset where 87% of the data are acted for basic expressions, whereas the remaining 13% are based on spontaneous basic expressions. Given the distribution of spontaneity vs. acted in the MMI dataset, MMI dataset gets positioned more towards acted than spontaneous, as shown in Figure 2-1.

In the RU-FACS database, participants were given a choice to lie about an opinion and receive $50 in return if they successfully convinced the interviewer. Otherwise, they would have to fill out a laborious and time-consuming questionnaire. Therefore participants were more inclined to lie, eliciting stronger emotions. Since the participants had to act to hide their true position, one could argue that the RU-FACS dataset is not fully spontaneous. Also, the RU-FACS dataset is not publicly available at this time. SAL and SEMAINE are publicly available datasets where participants are led through a range of emotional states through an interface. The interface is controlled by an operator who acts out one of four basic emotions (happy, sad, angry, and neutral) to elicit from the participants the same emotion as well. The SEMAINE dataset contains 578 labeled annotations, of these, 26% of them are "basic," 37% are "epistemic," 31% are "interaction process analysis," and the rest are instances of "validity." In the Spaghetti dataset, participants were asked to insert their hand inside a box that contained a warm bowl of Spaghetti. Since the participants didn't know what was inside

the box, they reacted strongly with disgust, surprise, fear, or happiness. The Spaghetti dataset only contains three participants with a total of 1 minute and 35 seconds of data, but it is highly spontaneous. The SAL data consists of audio-visual recordings of human-computer conversations. The conversations are elicited through an interface called "Sensitive Artificial Listener." The interface contains four characters with four different personalities – Poppy (happy), Obadiah (sad), Spike (angry), and Prudence (pragmatic). Each character has a set of responses that match their personalities. It is hypothesized that as the participants interact with Poppy/Obadiah/Spike/Prudence, the participants get drawn into the affect that those characters display. The Mind-reading dataset contains examples of more complex mental states, e.g., concentrating, thinking, confused, interested, agreement, and disagreement, and over a dozen others, but it has professional actors acting all the states. Affectiva-MIT Facial Expression Dataset (AM-FED) consists of 242 facial videos (168,359 frames) recorded through web cams as users watched Super Bowl ads in their own environment (Mcduff et al., 2013). As the users volunteered to watch an ad while agreeing to be video-taped and were not asked to act, AM-FED data is considered naturalistic and spontaneous. But the dataset may not necessarily contain the complete set of expressions.

Facial Analysis

Early efforts in modeling expressions include Yacoob et al. (Yacoob & Davis, 1996) and Cohn et al. (Lien, Kanade, Cohn, & Li, 1998), who have used optical flow computation to identify motions caused by human facial expressions. Using image sequences, Yacoob et al. were able to recognize eye blinking and six facial expressions (Yacoob & Davis, 1996). Cohn et al. used an optical flow-based approach to capture the full range of emotional expression (Lien et al., 1998). The system used a hierarchical algorithm to track selected facial features automatically.

Another approach to recognizing facial expression is probabilistic modeling. Cohen et al. experimented with different Bayesian network classifiers for facial recognition from a continuous video input (Cohen, Sebe, Garg, Chen, & Huang, 2003). The study proposed a new architecture of hidden Markov models for segmenting and recognizing facial expressions. Hoey et al. (Hoey, 2004) created a system that learned relationships between facial movements. The model is a Partially Observable Markov Decision Process (POMDP) and has video observations integrated through a dynamic Bayesian network. Pardas et al. developed a system that recognized facial expressions by creating Hidden Markov Models with Facial Animation Parameters. The system was used to effectively identify isolated expressions (Pardas, Bonafonte, & Landabaso, 2002).

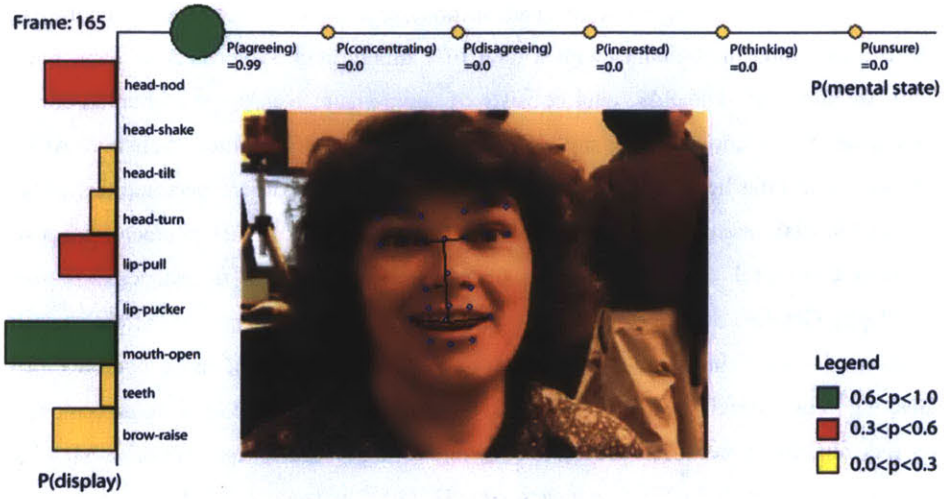


Figure 2-2. The MindReader Interface developed by el-Kaliouby et al (El Kaliouby, 2005)

The pioneering efforts of going beyond emotions and recognizing action units (AUs) were made by Bartlett et al. (Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999), Lien et al. (Lien et al., 1998), and Pantic et al. (Maja Pantic, Rothkrantz, & Koppelaar, 1998), as reported in (Jiang, Valstar, Martinex, & Pantic, 2011). Since then, a number of automated systems by Littleworth et al., (Littlewort et al., 2011) , el-Kaliouby et al. (El Kaliouby & Robinson, 2004) and Tian et al. (Tian, Kanade, & Cohn, 2001) have been developed that can recognize action units and complex mental states using frontal-view faces. In this thesis, I provide a detailed overview of el-Kaliouby et al. (El Kaliouby & Robinson, 2004) and Littleworth et al. (Littlewort et al., 2011), as their systems contain an interface to better understand and interpret the behavioral data.

El-Kaliouby et al. (El Kaliouby & Robinson, 2004) developed a real-time system called MindReader to recognize and visualize complex mental states by analyzing facial expressions and displaying the inferred states, as shown in Figure 2-2. The system was trained on acted video data from the Mind-Reading dataset (Baron-Cohen, 2004). The system consisted of a multilevel probabilistic graphical model that represents the facial features in a raw video stream at different levels of spatial and temporal abstraction. Dynamic Bayesian Network (DBN) was used to model observable head and facial displays, and corresponding hidden mental states, over time (El Kaliouby, 2005).

The Mind-Reader software has gone under several iterations to incorporate radial charts (Teeters, Kaliouby, & Picard, 2006) and traffic lights to better represent the inferred states. The system has been incorporated to help individuals with Autism Spectrum Disorder to capture, analyze, and reflect on a set of socio-emotional signals communicated by facial and head movements in real-life social interactions (Madsen et al., 2009)(Madsen, el Kaliouby,

Goodwin, & Picard, 2008). Analysis of whether the technology is improving the social interactions of those individuals remains for their future work. Mind-Reader focuses on a single modality and displays analysis results only in real-time, although results can be recorded for offline analysis.

CERT [4], another real-time behavior analysis system, was developed to recognize low-level facial features such as eyebrow raise and lip corner pull and graph them as a function of time. The system measures the intensity levels of Action Units by analyzing the SVM distance separating-hyperplane. A screen shot of the framework demonstrating its visualization parameters is displayed in Figure 2-3.

CERT was previously incorporated in a game called SmileMaze to help children with autism with their expression production skills (Cockburn et al., 2008). Participants were expected to navigate their character through a maze using a keyboard, and smile to move their game piece past obstacles at various points within the maze. Informal field study indicated that participants enjoyed tricking the system by posing odd expressions. Like MindReader, CERT also does not support multimodal behavior analysis.

Figure 2-3. Screenshot of the CERT framework (Littlewort et al., 2011)

Figure 2-4. SSI GUI that allows developers to record training data and build user-dependent or user-independent models out of it. The screen shot provides a snapshot of a session with video and audio data (J. Wagner, Lingenfelser, & Andre, 2011)

The most recent example of a real-time multimodal platform includes a Social Signal Interpretation (SSI) (Johannes Wagner, Lingenfelser, & Andre, 2011) tool. The framework was designed for developers to perform data recording, annotation, and classification using a Graphical User Interface (GUI). The developers also have the option to add their modules into the system using its flexible architecture. A screenshot of the system is shown in Figure 2-4.

While CERT, Mind-Reader, and SSI framework illustrate the promise of real-time, automated behavior sensing, there is a need for combining human-centric designs with multimodal nonverbal sensing that not only pushes the boundary of automated behavior analysis, but also empowers the users to understand and reflect on their own behaviors

2.4.2 Vocal Behavior Sensing

> *When most words are written, they become, of course, a part of the visual world. Like most of the elements of the visual world, they become static things and lose, as such, the dynamism which is so characteristic of the auditory world in general, and of the spoken word in particular. They lose much of the personal element...They lose those emotional overtones and emphases...Thus, in general, words, by becoming visible, join*

a world of relative indifference to the viewer – a world from which the magic "power" of the word has been abstracted.

- Marshall McLuhan in The Gutenberg Galaxy (1962), quoting J.C. Carothers, writing in Psychiatry, November 1959.

When an audio signal is transcribed into text, it becomes static and loses most of what may be its most important characteristics—the ability to signal underlying feelings. Characteristics such as harmonicity and dynamism contain emotional overtones and emphasize information that adds an extra layer on top of the words we say. The extra layer of information interacts with verbal content (e.g., sarcasm) and other nonverbal modalities (e.g., facial expressions, body language). The extra information is called prosody, which is conveyed as changes in pitch, loudness, speaking rate, pauses, and rhythm.

Diarization

Diarization is the process of identifying regions of time where a particular speaker is talking. Having such information makes it easier to automatically transcribe the speech, especially when it involves more than one individual. The methods for diarization can be categorized by the use of single or multiple modality - audio and audiovisual tracking. The use of one modality is reviewed in (Anguera et al., 2012). Attributes such as faster turn-taking and overlapping speech could make it difficult to transcribe conversational speech using audio only (Janin et al., 2004)(Wooters & Huijbregts, 2008). To overcome the difficulty, audiovisual speech diarization has been proposed and developed by leveraging the correlation between audio and video data (Friedland, Hung, & Yeo, 2009).

Transcription

Automatic speech recognition (ASR) is the next step after speech diarization. Speech recognition has been researched since the 1950's (see (Furui, 2005) for a complete review). State-of-the art speech recognizers' use previously recorded signals to train language models and extract language features to recognize speech. Unfortunately, error rates for ASR's are high for nonverbal behavior processing due to three challenges. 1) The discriminability of acoustic features is limited by far-field acquisition (to be more natural, we are not using a head-worn microphone), 2) acoustic variability is increased by speaking styles, and 3) the entropy of human expression is increased by emotion.

Intonation

Intonation is defined as the combination of tonal features into larger structural units associated with the acoustic parameter of voice fundamental frequency or F_0 and its distinctive variations in the speech process (Botinis, 2001). F_0, measured in Hz, is defined by the quasi-periodic number of cycles per second of the speech signal. F_0 is controlled by

muscular forces of the sublaryngeal (respiratory) system, and it is measured by the number of times per second that the vocal folds complete a cycle of vibration. The perception of intonation is defined by the perceived pitch, which roughly corresponds to F_0 realizations. In signal processing, F_0 means the greatest common divisor of the harmonics of a mathematically periodic signal, and pitch means the frequency of a sine wave that listeners judge to be equally as high as the signal.

Development of intonation models with predictive power has been formalized by many researchers (Gårding, 1979)(Thorsen, 1980)(Pierrehumbert, 1980). Those models have been tested and evaluated across many languages (Gårding, 1982)(Hirst & Di Cristo, 1984). In addition to modeling forms and functions, efforts have gone into labeling and transcription of intonation. Several systems have been proposed, including Tone and Break Indices (ToBI), applied initially in American English (Silverman et al., 1992), and International Transcription System for INTonation (INSTINCT), applied in several languages (Hirst & Di Cristo, 1998).

Loudness

Sound pressure level (or sound intensity) is an *objective* property of a sound, measured in decibels (dB). Perceptual loudness (or just loudness), on the other hand, is *subjective* and represents the magnitude of the physiological sensation produced by a sound. Perceptual loudness varies with intensity and frequency of sound, specifically the human ear's sensitivity to different frequencies. While loudness is closely related to sound intensity, they are by no means identical. Figure 2-5 shows an equal-loudness contour curve where the contours are lines on a graph of sound pressure level against frequency, joining all points that are equally loud. A 1000 Hz tone of 70 dB will, to most people, sound louder than a 200 Hz tone also at 70 dB.

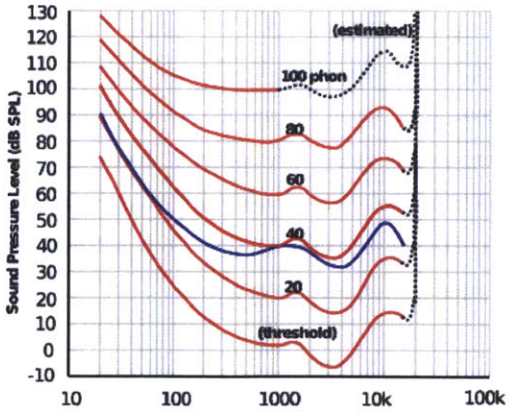


Figure 2-5. Equal-loudness contour curves are shown in red. Original ISO standard shown in blue for 40-phons. The x-axis is the frequency in Hz; the y-axis is the sound pressure level in dB (adapted from ISO 226:2003).

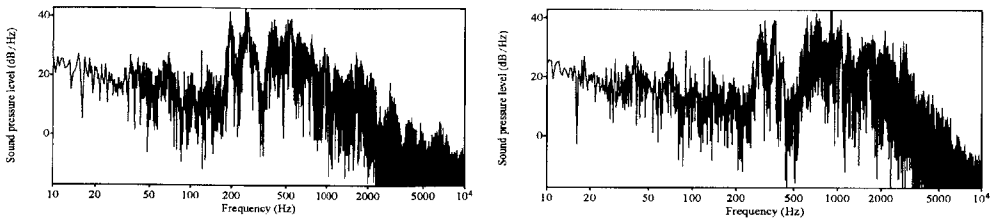


Figure 2-6: The same utterances said with different loudness levels have almost identical sound pressure levels. Left: Utterance with a soft tone with the microphone being close to the mouth. Right: Utterance with a soft tone yelled at a distance from the microphone.

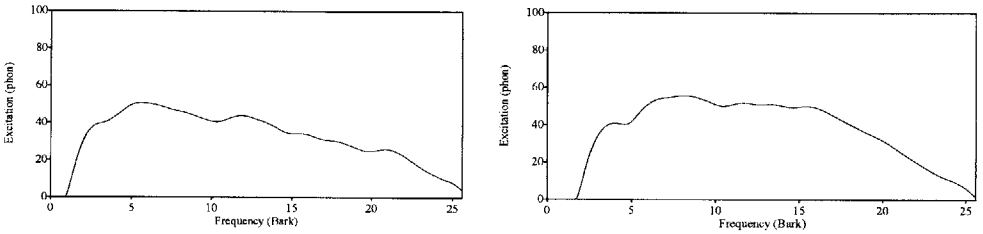


Figure 2-7: Excitation in phons for the same two utterances showing a difference in their loudness. Left: Utterance whispered close to a microphone. Right: Utterance yelled at a distance from the microphone.

In the context of Human-Computer Interaction, the difference between sound pressure and perceptual loudness means that a person could have the microphone close to the mouth with a soft voice or push the microphone away from the mouth and be somewhat loud. While the intensity of the two speech files might be the same (as shown in Figure 2-6 where the average sound pressure level is 60 dB for the two utterances), a human listening to the two files would easily discern the soft tone from the loud one. Figure 2-7 shows that the excitation in phons is able to capture this perceptual difference, where the first utterance (soft tone) has an average excitation of 21 sones and the second (loud tone) has an average excitation of 32 sones.

Speech rate estimation

When phonetic transcription is available, typically Automatic Speech Recognition (ASR) engines are used to align speech to the text symbol sequence to find a number of certain linguistic units per unit time (Katsamanis, Black, Georgiou, Goldstein, & Narayanan, 2011). Without speech transcription, speech rate information can still be estimated by techniques based on spectral sub-band correlation (Morgan & Fosler-Lussier, 1998) and techniques that include temporal correlation with prominent spectral sub-bands that help identify syllables (D. W. D. Wang & Narayanan, 2007). Speech rate could also be calculated by counting phonetic elements such as words, syllables (Morgan & Fosler-Lussier, 1998), stressed syllables, or phonemes (Nanjo & Kawahara, 2002) per second. In other studies, duration of

phoneme and comparison of measured and expected phoneme duration prove to be robust measures of long-term and short-term speech rate (Siegler, 1995). However, modeling the expected phone duration beyond a limited number of cases remains a difficult problem. Nooteboom (Nooteboom, 1997) showed that the syllable, defined as a unit of spoken sounds uttered with a single effort or impulse of the voice, helps provide a good estimate of speech rhythm, which is similar to speech rate. Syllables, by this definition, should demonstrate quite an even distribution under normal speed speech and their rate would depend on speech rate change (D. W. D. Wang & Narayanan, 2007). Therefore, syllables are widely used among researchers to measure speech rate (Shriberg, Stolcke, Hakkani-Tur, & Tur, 2000) (Siegler, 1995).

Word prominence estimation

In speech, some words are more prominent than others. Emphasizing certain syllables draws the listener's attention and can be automatically detected (Tamburini & Caini, 2005). Word prominence can be identified by different acoustic features such as spectral intensity, pitch, and speech rate (D. Wang & Narayanan, 2007).

Combining linguistic information with acoustic features

There are many techniques available to combine speech acoustic features with lexical, syntactic and discourse information. Examples include decision tree methods (Ross & Ostendorf, 1996; Wightman & Ostendorf, 1994), maximum likelihood classification (Chen, Hasegawa-Johnson, & Cohen, 2004), maximum *a posteriori* classification (Ananthakrishnan & Narayanan, 2005, 2008) and maximum entropy method (Sridhar, Bangalore, & Narayanan, 2008).

Interaction modeling

The social interactions between the participants often contain useful information so that their relationship could be automatically modeled. Tracking dynamics such as speaker activity and utterance patterns can yield information on the underlying behavior process. For example, patterns of similar pitch and energy are associated with affective dynamics or success in accomplishing tasks (Levitan, Gravano, & Hirschberg, 2011).

2.5 VISUALIZATIONS OF BEHAVIORAL DATA

Visualization of behavioral data for users to understand, interpret, and reflect on involves a few considerations. Should the data be mapped directly to the interface or should it be abstracted? Should the feedback be instant or gradual? Should the interface tell users what to do or should it empower users with data to make the right decisions? How do we even measure the effectiveness of such interfaces? This section provides a brief overview of

relevant projects with design considerations and tradeoffs. It demonstrates that solutions often depend on the applications themselves, and as well as contextual inquiry.

2.5.1 Visual Feedback for Self-reflection

UbiFit, an example of *persuasive technologies* (Fogg, 2003), was designed to help users reach physical activity goals set forth by themselves (Consolvo et al., 2008). The system contained the Mobile Sensing Platform (MSP) device (Choudhury et al., 2008) that collected data all the time to infer activities such as running, cycling, elliptical trainer, and stair machine. The interface also relied on the users to enter and categorize physical activities for better results. The system took the raw data and abstracted them into a flower garden to visually communicate whether the users are achieving their goals or not. It did not try to influence users' decisions by providing a recommendation. UbiFit was designed to visualize long-term abstract effects as opposed to empowering the users to understand, analyze, and reflect on their physical activity data.

Figure 2-8. UbiFit garden display showing butterflies and flowers (Consolvo et al., 2008)

Affective Diary was another application of visualizing sensed data to help users self-reflect (Ståhl, Höök, Svensson, Taylor, & Combetto, 2008). Unlike UbiFit, the diary not only captured data about the users, it also allowed them to reflect on their data by storing context. For example, a user would wear an armband sensor collecting Galvanic Skin Response (GSR) values (movement and arousal). There was a mobile phone logging system that collected received/sent text messages, photos taken, and other Bluetooth phones nearby. At the end of the day, users could upload their data to a mobile PC and explore using visual feedback. The visual feedback was designed to be ambiguous and organic with features that directly

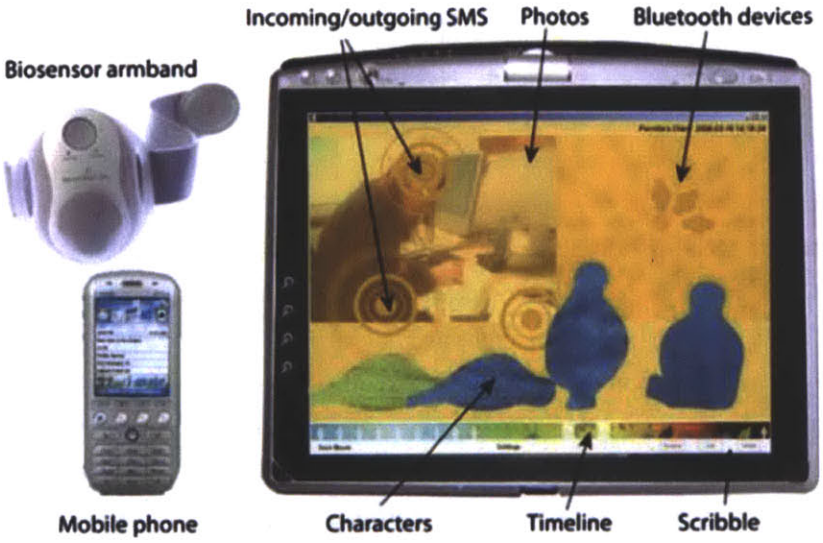


Figure 2-9. Affective Diary visual feedback on a PC tablet (Ståhl et al., 2008)

correlated to the data. For example, bodily activity was visualized in person-like figures. The figures automatically stood up when more movement was picked up, or they changed their color to red when the body was aroused. The abstract visual feedback was open to interpretation by the user. Many users reported that the visualized measurements of the body did not link with their subjective experiences of body movements. A few participants mentioned that the visual representation of their behavior felt foreign to them, making it difficult to incorporate the graphs in understanding themselves.

The two specific examples of UbiFit and Affective Diary highlight the difficulties in abstracting the human behaviors in a graphical format that is educational and self-explanatory.

2.5.2 Visual Feedback during Group Meetings

There has been work done to visualize the dynamics of nonverbal behavior during group meetings (T. Kim, Chang, Holland, & Pentland, 2008) (Sturm, Herwijnen, Eyck, & Terken, 2007). For example, Meeting Mediator, developed by Kim et al., is a system that uses mobile phones to display real-time visual feedback to aid group collaboration using sociometric badges (T. Kim et al., 2008)(Olguin Olguin et al., 2009). A Sociometric badge is a sensing device intended to be worn around one's neck. The badge can capture voice features (energy, speaking time, pitch), body movement using a 3-axis accelerometer, and

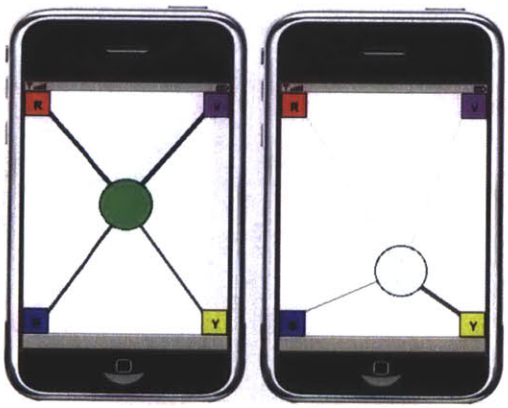


Figure 2-10. The visualization on the mobile phones emphasizes interactivity and balance. The four squares represent the participants and the position of the circle shows the balance of conversation. The color of the circle represents interactivity where green corresponds to higher interactivity and white corresponds to lower interactivity (T. Kim et al., 2008)

physical proximity to other people. In the study conducted by Kim et al., (T. Kim et al., 2008), each participant in the group wore a sociometric badge to collect data, which was sent to a smart phone paired with Bluetooth. The badges communicated with each other through radio and generated information on interactivity. Using the data, the system could visualize the balance of conversations on the phones to aid collaboration, as shown in Figure 2-10.

Sociometric badges are very useful in understanding the high-level structure of group interactions. For example, they show data representing the flow of meetings, speaking balance, and turn-taking patterns during face-to-face discussions. But due to its form factor, it can only sense a few high-level nonverbal features. Therefore, if one wanted to use it just to reflect nonverbal nuances of interaction, it might not be very effective. Also, sociometric badges were originally designed to understand the behavior of large organizations, and its direct effectiveness in helping individuals improve their nonverbal behaviors remains an empirical question.

2.5.3 Visual Feedback between Clinician-Patient Interaction

Using nonverbal communication as visual feedback to help health professionals has recently been explored. A system built by Patel et al. uses Wizard-of-Oz technique to capture the nonverbal behaviors and use them to drive reflective feedback through a scenario-based lab study (Patel, Hartzler, Pratt, & Back, 2013), shown in Figure 2-11. The visual feedback displays information on clinicians' and patients' reflection of warmth and connection as well as dominance and power. However, Patel et al. acknowledge that they might have provided potential users with feedback that captured the essence of researchers' perceptions of

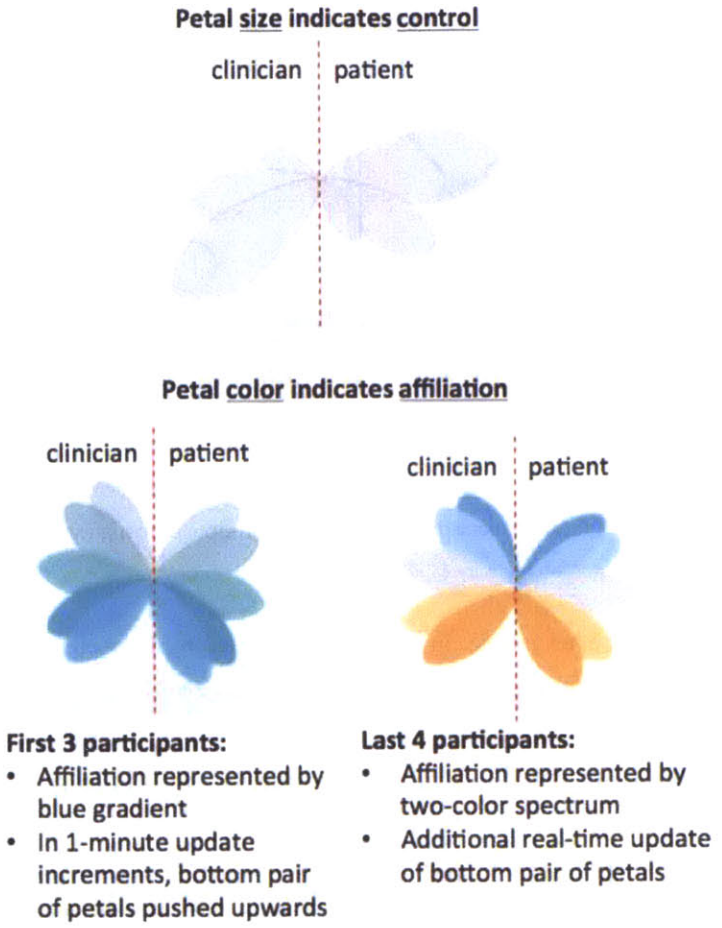


Figure 2-11. The lotus flower visualization uses size of the petals to represent control (dominance, power, etc.) and color to represent the affiliation (warmth, connection, etc.) (Patel et al., 2013)

nonverbal communication styles. Developing an automated system to recognize nonverbal behavior and following it up with real-world use in clinical settings remains part of their future work.

Based on the discussion above of UbiFit, AffectiveDiary, sociometric badges, and system developed by Patel et al., there seems to be an opportunity to make a contribution by leveraging the existing knowledge on visualization techniques. The inability of existing systems to help people improve their individual nonverbal behavior in conversations further motivates the research question, how would an interface automatically interpret, and represent conversational multimodal behavioral data in a format that is both intuitive and educational in the context of the interaction?

2.6 CONVERSATIONAL VIRTUAL AGENTS

> *It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly, it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it.*
>
> *-Alan Turing, "Computing Machinery and Intelligence," 1950*

Cassell describes conversational virtual agents as intelligent user interfaces (Justine Cassell, 2000). The virtual agent platform contains gestures, facial expressions, and speech to enable face-to-face interactions with the users, providing a paradigm in human-computer interaction.

Conversational virtual agents have been designed to support many communicative functions (e.g., learning environments, museum guides, and helping children with cognitive activities) The section below provides a summary of some of the characteristics of conversational virtual agents to further motivate the problem addressed in this thesis.

2.6.1 Virtual Agents as Training Interfaces

Steve (Soar Training Expert for Virtual Environments) was one of the early examples of animated agents designed for training students in the virtual world (Rickel & Johnson, 2000). *Steve* contained an interface, a simulator, commercial speech synthesizer (that did not emphasize important words), and a cognitive model. Steve trained students in how to maintain gas turbine engines on ships.

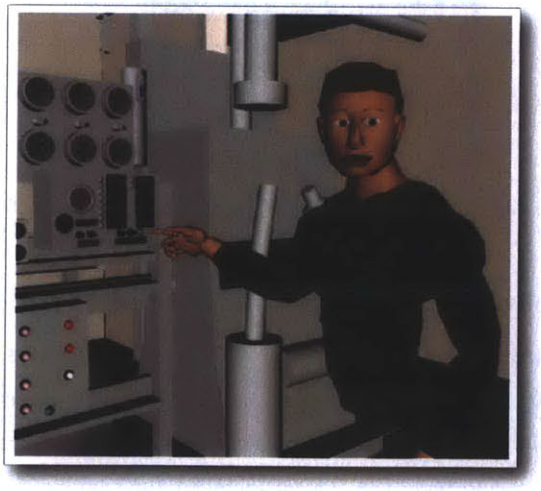


Figure 2-12. The agent *STEVE* (Rickel & Johnson, 2000)

Steve used gazes, pointing, and body orientations to draw attention to something and nodded or shook its head when the student did something correctly or incorrectly. The students could, however, interact with *Steve* only by "touching" virtual objects with a data glove. Even though *Steve* had the novelty factor of being used in a training scenario, it did not collect anything related to users prosody and facial expressions.

Other examples of training scenario include training clinical therapists (Kenny, Parsons, Gratch, Leuski, & Rizzo, 2007), training cross-cultural meeting skills (J. Kim et al., 2009), and culture and language skills (Johnson, Vilhjalmsson, & Marsella, 2009).

2.6.2 Discourse is Contingent on Conversational Cues

Unlike Steve, *Rea* was one of the early prototypes of conversational virtual agents where Cassell et al. made the observation that effectiveness of conversational interfaces is contingent on understanding the verbal and nonverbal conversational cues (J Cassell et al., 1999). *Rea* was designed to play the role of a real-estate agent, who could show properties to the user along with conducting a conversation with them. It sensed the users passively through cameras by responding to their speech, shifts in gaze, gestures and non-speech audio input. *Rea*'s synthesized behavior included using speech intonation, facial display, and gestural output through hard coded rules. *Rea* was a great step forward towards developing conversational agents that "*can hold up their end of the conversation*" (J Cassell et al., 1999). However, it was not designed or evaluated for a real-world training scenario nor it could process any prosody and facial expressions.

Figure 2-13. The Real Estate Agent – *REA* (Narayanan & Potamianos, 2002)

2.6.3 Interaction requires Social Awareness

Narayanan et al. (Narayanan & Potamianos, 2002) created conversational interfaces where children could interact with a computer using natural language, commands, or buttons. All the user actions would get translated into text strings, and the system responded through audio, graphics, animation, and text. The system included applications such as information retrieval from a personal directory, placing phone-calls, accessing the Internet, sending email)

and a computer game (spelling bee). The system was evaluated with eight children using qualitative methods. The authors acknowledged that flattening the nonverbal behavior as a text input, and using the text to drive the output, had negative implications. For example, many participants mentioned that the agent lacked social interaction capabilities. It further confirmed the notion that interaction or training is most effective when the agent appears socially aware of the affective states of the participants.

Cassell and Thorisson studied the use of nonverbal behaviors related to the process of conversation, also known as *envelope feedback* (Justine Cassell & Thórisson, 1999). The envelop feedback does not rely on the content of the conversation; rather it follows a regulatory function. For example, periodic nodding of the head to indicate that one is paying attention. Cassell et al. reported envelope feedback being extremely important in human-agent conversations.

Another interesting example of socially aware behavior was reported in *GrandChair* (Smith, 2000). In *GrandChair* participants sat in a comfortable rocking chair and told stories with the assistance of a conversational virtual agent that took the form of a child. The agent prompted the participants with questions, stories, and videos from previous interactions to help the participant tailor the stories for children. It was shown that embodied characters that nodded their head led to longer responses from the users than a tape-recorder asking the same questions.

2.6.4 Social Acceptance of Virtual Agents

Max, a museum guide, is one of the pioneering examples of using conversational virtual agents in a public environment (Kopp, Gesellensetter, Krämer, & Wachsmuth, 2005). The authors skillfully designed an interaction scenario given the challenges of being in a noisy environment. Max, an embodied virtual agent, allowed its users to engage in small-talk using a keyboard. Using a camera, Max could automatically recognize human faces, and use it to alter its affective states by changing its speech rate and voice pitch (e.g., spotting a human face in the camera automatically caused a positive stimulus). It did not, however, perform any affective analysis.

The authors reported that visitors used human-like communication strategies (e.g., greeting, farewell, small talk elements, insults) with Max. Visitors took the trouble to answer Max's questions, and adapted well to Max's human-likeness and intelligence. This was an example of agents being socially accepted by humans when the interactions could be carefully designed.

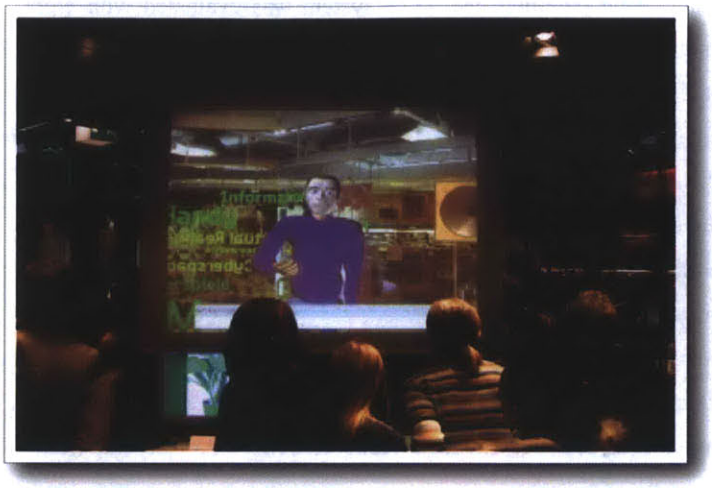


Figure 2-14. The museum guide – Max (Kopp et al., 2005).

2.6.5 Virtual Agents to Help with Social Difficulties

Cassell and Tartaro (Tartaro & Cassell, 2008) developed an intervention for children with social and communication difficulties using an embodied virtual agent called *Sam*. *Sam* did not perform any automated nonverbal sensing. Instead, it was controlled by a Wizard-of-Oz approach by letting the puppeteer select a pre-recorded set of speech and gestures. The authors ran two studies, children with autism engaging in collaborative discourse, 1) with another child, and 2) with a virtual agent, *Sam*. The findings suggested that unlike interactions with the human peer, contingent discourse increased over the interaction with the virtual agent. Also, topic management, such as introducing new topics or holding onto the current topic, was more frequent in interactions with the virtual agent. Even though the study was limited to six individuals, and it did not result in any automated technology, the evidence provided in their work demonstrated that conversational virtual agents may be useful interventions for individuals with social difficulties.

2.6.6 Social Interventions to Foster Social Connectivity

Ring et al. developed embodied conversational agents aimed to provide longitudinal social support to isolated older adults (Ring, Barry, Totzke, & Bickmore, 2013). The users are asked to enter an input out of a set of possible utterances using a touchscreen, and the agent responds with synthesized voice as well as nonverbal behavior. The multiple-choice menus are updated after each turn. Through the study, the authors claimed the following two things; 1) an in-home conversational agent has social acceptability for older adults; 2) intervention can influence on loneliness.

However, the authors do acknowledge two major limitations that they are planning to address in the future. First, the authors agree that the conversational agent could provide more nuanced and comprehensive counseling if it could sense the affective cues of the users. Second, their intervention does not facilitate social connectivity. For example, the goal of using agents is not to provide an alternative to human companionship to address loneliness. Instead, it should consider incorporating "human-in-the-loop" design so that the final outcome of the intervention could be measured by their effort to seek companionship through their social network to maintain and strengthen their relationships. This is a very important aspect of using conversational virtual agents for behavioral interventions so that the final outcome could be measured through a congruent real-world task.

2.6.7 Current State-of-the-Art

Two of the most advanced examples of real-time systems combining a virtual agent, incremental analysis of behaviors, dialogue management, and synthesis of behaviors are *Sensitive Artificial Listener (SAL)* (Schroder et al., 2011) and *Rapport Agent* (Gratch et al., 2006) (Huang, Morency, & Gratch, 2011).

Figure 2-15. The four SAL characters with 4 different personalities: aggressive, cheerful, gloomy, and pragmatic (Schroder et al., 2011).

SAL was designed to engage the user in a conversation by paying attention to the user's emotions and nonverbal expressions. The SAL characters have their own personalities; 1) Aggressive Spike; 2) Cheerful Poppy; 3) Gloomy Obadiah; 4) Pragmatic Prudence, as shown in Figure 2-15. Their objective is to drag the user into their dominant emotion through a set of verbal and nonverbal expressions. The agents have very limited verbal understanding. The system avoids task- driven conversations and tries to engage the users into conversations by exhibiting listening behaviors.

Rapport Agent 1.0 was designed to yield contingent nonverbal behaviors as the participants spoke. The initial system allowed participants to retell some previously observed stories from a recently watched video to a nonverbally attentive agent. Later, *Rapport Agent* 2.0 (Huang et al., 2011) was designed, allowing the agent to demonstrate positive affective information and reciprocity both verbally and nonverbally. The system was able to automatically model backchannels (L. Morency, Kok, & Gratch, 2008), predict end-of-turn, and provide enhanced

positive emotion during the conversation. The rapport agent was developed to establish rapport with its users during interaction as opposed to testing it on task-driven real-world scenarios.

Based on the literature search presented above, the following set of features stood out as necessary components to ensure *seamless* and *meaningful* interactions with conversational virtual agents.

- **Autonomous** – The agent does not require any manual input from the user. Once started, the agent is able to execute instructions on its own and terminate the session when necessary.
- **Embodied** – The research community has long been in debate over whether the conversational agents should take the physical form of humans. Mori first introduced the phenomenon of "Uncanny Valley" (Mori, 1970) where, the author argued, as robots take the form of humans, human observers' emotion becomes positive. However, there is a point when the positive emotion quickly turns into dislike. Therefore, many researchers take a stand on developing cartoonlike conversational virtual agents to minimize the expectation from the users. However, with the advancement of computer graphics and dynamic behavioral synthesis algorithms, there is a momentum towards using more realistic human-like agents. Many believe that if the agents are designed to simulate realistic interaction scenarios, a cartoonish look make it seem less serious.
- **Understand Speech** – The agent should understand what the user is saying. In other words, perform automated speech recognition.
- **Prosody Analysis** – The agent should be able to understand the prosodic contour of the speech from its users. In other words, the agent should not only understand what the users are saying, but also pay attention to how they are saying it.
- **Facial Expression Analysis** – The agent should have the capability of sensing the facial expressions of the user.
- **Training Scenario** – The area of conversational virtual agents, being a relatively new field, has focused on developing theories and techniques to support interactions with humans. With the establishment of relevant theories and techniques, there is a growing trend and motivation towards using agents for training scenarios using automated methods.
- **Affective Feedback** – If the agent is able to sense the nonverbal behavior of its user, perhaps it should exhibit behaviors to demonstrate its emotional intelligence. This could include real-time feedback (e.g., looking uninterested when the speaker's tone of voice falls off) or even post-feedback reflecting on the interaction.

- **Automated Backchanneling** – In order to support autonomous human-agent interaction, the agent should be able to synthesize its own behavior. One such example is backchanneling. For example, the agent could nod its head, share smiles, provide quick verbal acknowledgements (e.g., "uh-huh"), etc. while listening to a human.
- **Validation in Real Scenarios** – It is very important to think of new ways of using conversational virtual agents with real-world utility and implications. Along with proposing new interaction possibilities, it is also important to validate the proposed framework in real scenarios and with human users. This helps the community understand the practical limitations and promises of a particular technique to inform future work.

2.7 SUMMARY AND OPPORTUNITIES TO CONTRIBUTE

Nonverbal behaviors are the most important aspects of face-to-face interactions. The nonverbal research community has come a long way in developing theories and technologies towards understanding the nuances of nonverbal behaviors. While significant progress has been made, there is still a long way to go before we could reliably recognize and interpret the full range of nonverbal behaviors and have them fully incorporated with conversational virtual agents. Here are a few limitations that I have observed during the time of writing this dissertation.

	Current Limitations	Made an effort to address
1.	In terms of facial expression processing, there is an inclination to establish a one-to-one mapping between patterns of FACS with basic emotions. For example, occurrence of Action Unit 12 and Action Unit 6 means one is happy.	Addressed in *Chapter 4.2.3 Analysis: Acted vs. Elicited Feedback* with contradictory examples.
2.	Most of the established insights and theories about nonverbal behaviors are drawn from acted data. Those findings need to be re-validated with more naturalistic and spontaneous data, as they become available.	Addressed in *Chapter 4.2.4 Experiment 3: Temporal Models to Classify Smiles of Delight and Frustration.*
3.	Even though many systems were developed to sense real-time nonverbal behavior, none of them addressed the issue of how best to visualize the nonverbal behavior to the user in an intuitive format using user-centric approach.	Addressed in *Chapter 5:Training Interfaces for Social Skills*
4.	Before this thesis, there was no real-time multimodal platform	Addressed in *Chapter 6:My*

that allows users to understand and reflect on their nonverbal behavior and that has proven to be effective in enhancing nonverbal behaviors in real interaction scenarios.	*Automated Conversation coacH (MACH).*
5. Most of the computational interventions designed to improve nonverbal social skills find it challenging to demonstrate that the skills generalize while interacting with other humans.	Addressed in *Chapter 7:Evaluation of MACH in the Job Interview Scenario.*

Working with basic emotions has helped promote progress in expression recognition. But it is also important to push the boundary of working with spontaneous naturalistic data congruent with realistic tasks. For example, tools and techniques derived to correlate FACS with basic emotions may work well with acted or other limited forms of data; but it may not work with spontaneous data. Therefore, the differences between acted and spontaneous data need to be quantified, understood, and used to make systems that do work well for natural data.

Most of the conversational virtual agent systems successfully integrate aspects of affective analysis and interactive characters, but they do not include other components that are necessary for an automated coach such, as 1) a realistic task such as training real users to enhance nonverbal behavior, and 2) formative affective feedback that provides the user with useful feedback on their nonverbal behavior.

In summary, there is an opportunity to generate new knowledge in nonverbal behavior by working with naturalistic and spontaneous data. Second, it may be possible to use the knowledge thus generated to design interaction scenarios with conversational virtual agents to help people enhance their nonverbal social skills in many different situations.

2.8 CHAPTER SUMMARY

This chapter introduced some of the theoretical knowledge of nonverbal behaviors and provided evidence of why they are important. Lack of nonverbal skills introduces difficulties in social interactions. Many social and computational interventions have been developed to enhance human nonverbal behaviors. While some of them were reported to be effective for the given task, for many, how the learned skills generalize to other tasks remains a challenging question.

The preponderance of the theoretical knowledge presented in the field of facial expressions processing addresses six basic emotions. As we are able to collect more spontaneous and naturalistic datasets, there is an opportunity to computationally contribute to expanding, and to revalidating the existing theories.

Representation of behavioral data for humans to understand, interpret, and reflect on has involved many approaches. Most of the existing approaches have focused on displaying the data to the user either directly or indirectly, without explicitly trying to translate the patterns into action items.

There has been much research in automating the sensing of nonverbal behaviors in real-time. However, none of those frameworks were designed for people to practice and reflect to further enhance their nonverbal behavior. Similarly, in the field of conversational virtual agents, training users with real scenarios has not been fully explored. Therefore, there remains an opportunity to combine nonverbal behavior sensing with conversational virtual agents along with easy representation of behavioral data to enhance human nonverbal skills.

The next chapter defines a job interview as an interaction scenario. It identifies what really matters during job interviews through an extensive literature search and contextual inquiry.

Chapter 3: Job Interview – An Interaction Scenario

To develop the computational framework to enhance human nonverbal behavior, and to perform a thorough evaluation, I chose a job interview as a scenario. There were three reasons to choose a job interview as a scenario: 1) Job interviews follow a predictable structure of interaction pattern (e.g., turn taking) that are easy to model; 2) It is comparatively easy to recruit motivated participants in an academic setting to practice job interviews; and 3) it serves as an example for applications of nonverbal sensing in occupational therapy/training.

This chapter motivates, and contextualizes the study, and sets the stage to better understand the structure of job interviews. Through the background provided in this chapter, this thesis envisions simulating a scenario where a conversational virtual agent could play the role of the interviewer and help the interviewee with his or her nonverbal skills (to be discussed in Chapter 6).

This chapter provides an extensive review on "what really matters" during job interviews. Based on the findings from the relevant interview literature, I set up a mock interview study with 28 MIT students and 4 career counselors. The behavior of the interviewer and the interviewee were analyzed to inform the design process of the automated system to enhance human nonverbal behavior, to be explained in the later chapters.

3.1 JOB INTERVIEWS – WHAT MATTERS?

Several studies (A I Huffcutt, Conway, Roth, & Stone, 2001)(Campion, Campion, & Hudson, 1994)(Pulakos & Schmitt, 1995)(Allen I Huffcutt & Winfred, 1994)(McDaniel, Whetzel, Schmidt, & Maurer, 1994)(Schmidt & Rader, 1999)(Wiesnerf & Cronshaw, 1988)(Wright, Lichtenfels, & Pursell, 1989)(Conway, Jako, & Goodman, 1995) have provided evidence that performance during interviews is often used by interviewers to predict job performance. Given the limited interaction, interviewers often focus on an applicant's nonverbal cues to get an understanding of the applicant's personality, social skills and knowledge to determine the applicant's suitability. Along with nonverbal cues, there are also knowledge and skills, organizational fit, and preference that are considered towards hiring an applicant. Huffcutt et al. (A I Huffcutt et al., 2001) provides a complete taxonomy on interviews, as shown in Figure 3-1.

Based on an extensive literature search, Huffcutt et al. (A I Huffcutt et al., 2001) have identified mental capability, knowledge and skills, personal traits, social skills, interests and preferences, organizational fit and physical attributes to be the most important constructs of

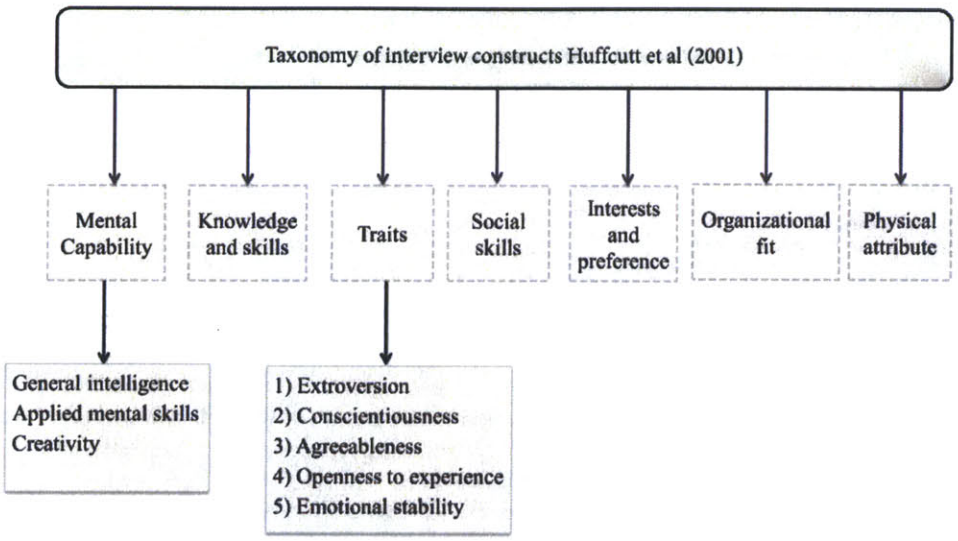


Figure 3-1. Taxonomy of interview constructs by Huffcutt et al. (2001) (A I Huffcutt et al., 2001)

interviews. After an analysis of 338 ratings from 47 actual interviews, Huffcutt et al. (A I Huffcutt et al., 2001) conclude that personality traits and social skills were the most frequently rated constructs, followed by mental capability and job knowledge and skills. Blackman et al. further argued in favor of social skills by stating that removing nonverbal behavior from a job interview (e.g., having a phone interview as opposed to conducting it in person) reduces accuracy (Blackman, 2002). Thus, Blackman (Blackman, 2002) argues that social skills certainly can contribute to accurate judgments during interviews, as cited in (Congreve & Gifford, 2006). In a study by McGovern, nonverbal behavior was shown to be important to interviewers' hiring impressions and decisions (McGovern & Tinsley, 1978). In the study, only participants with above average eye contact, speech fluency, and voice modulation were given a second interview.

According to DeGroot and Gooty (DeGroot & Gooty, 2009), examples of social skills during interviews include eye contact, frequent smiles, and modulation of speech (DeGroot & Gooty, 2009). DeGroot and Gooty (DeGroot & Gooty, 2009) describe that an interviewee who maintains direct eye contact with the interviewer appears more direct, honest and conscientious than one who looks away. Amalfitano & Kalt (Amalfitano & Kalt, 1977) also show that interviewees with more eye contact are perceived to be more responsible and have more initiative. Direct eye contact is also known as the reflection of an attentive listener during interviews (Shlenker, 1980).

Along with eye contact, in a study by Hollandsworth et al., fluency of speech was found to be an important factor in the hiring decision (J. G. . J. Hollandsworth, Kazelskis, Stevens, & Dressel, 1979). Speech fluency was ranked as the second most important variable

indicating appropriateness of content for a hiring decision (J. G. . J. Hollandsworth et al., 1979). Fluency of speech was defined as "[speaking] spontaneously," "[using] words well," and "[ability] to articulate thoughts clearly." "Pause-think-speak," a method that involves training interviewees to organize their response briefly before answering the interviewer's question, has shown to be an effective way to improve speech fluency (J. G. Hollandsworth, Glazeski, & Dressel, 1978).

Smiling, the "easiest recognized facial sign when greeting others," can also help increase likability (Shlenker, 1980). Interviewees are perceived more sociable and extroverted when they smile frequently and vary their voice intonation (Lord & Maher, 1993)(Schlenker, 1980). Smiling and gesturing were found to be correlated to each other and to the interviewer's perception of the applicant as motivated to work (Gifford, Ng, & Wilkinson, 1985). For example, human judges were given interview tapes and questionnaires about the social skills and motivation level of the interviewee. Applicants who used more gestures while smiling were perceived to have better social skills (Gifford et al., 1985).

Smiling while making eye contact is also believed to have positive influence on the interviewers' responses (Parsons & Liden, 1984). According to Edinger and Patterson, eye contact, smiles, gestures and head nods were found to yield better results in interviewers' judgment (Edinger & Patterson, 1983). Shlenker mentioned that "smaller interpersonal distances," "body orientation" and "facing towards the other person" show that one is listening (Shlenker, 1980).

Exhibiting the appropriate nonverbal behavior helps to establish rapport with the interviewer, which, according to Cogger (Cogger, 1982), is the heart of the interview process. Rapport is often equated with the fluidity and synchrony of the interaction. More specifically, as mentioned by Gratch et al. (Gratch, Wang, Gerten, Fast, & Duffy, 2007), Tickle-Degnen and Rosenthal (Tickle-Degnen & Rosenthal, 1990) define rapport-building behaviors as "behaviors indicating positive emotion (e.g., head nods, or smiles), mutual attentiveness (e.g., mutual gaze), and coordination (e.g., postural mimicry or synchronized movements)".

Given the importance of social nonverbal behaviors, many companies rely on technologies to filter out candidates with lack of social skills. Companies such as HireVue ("HireVue," n.d.) and JobOn ("JobOn," n.d.) now provide web-based technologies for employers to pre-screen potential employees without bringing them to the site. The technology allows candidates to go a website, enable their webcam, and answer a selected set of questions to be reviewed by the company later. Such practices highlight the changing nature of interviews and motivate the need for developing tools that can help people improve communication skills in such contexts.

3.2 CONTEXTUAL INQUIRY

To inform the design of an automated system for social skills training and contextualize the design in training for job interviews, I sought to understand how expert interviewers carry out job interviews. Those interviews were expected to inform our decision of how we could possibly have a conversational virtual agent play the role of the interviewer, conduct the interview and provide feedback to the interviewee. This study's first consideration was the appearance of the virtual agent. Should it look human-like or cartoon-like? What kind of feedback should the virtual agent provide to the user and how? Should the system provide real-time feedback to indicate how well the interview is going, or should it wait for the interview session to end before debriefing its user, just like in other standard mock interviews? If the latter, how should the system debrief the user? Should it provide feedback verbally or visually? If the latter, what should the visualizations look like? In addition, I was also interested to explore the dimensions of the interview behaviors that are easier for humans to measure and agree upon. For example, how easy it is for people to agree on dimensions such as "overall enthusiasm" or even "recommended for hire" given a video of a job interview. I approached these questions with a process of iterative design, system development, and a sequence of formative and summative user studies.

Figure 3-2. Experimental setup of the mock interviews. Camera #1 recorded the video and audio of the interviewee, while Camera #2 recorded the interviewer.

Figure 3-3. The experiment room where the experimenter controls the audio and video settings and captures data with perfect synchronization. The participant and the interviewer were oblivious of the existence of this room.

3.2.1 Study Setup

To better understand how expert interviewers facilitate mock interviews, I conducted a mock-interview study in a room equipped with a desk, two chairs, and two wall-mounted cameras that captured the interviewer and the interviewee, as shown in Figure 3-2 and Figure 3-3.

3.2.2 Participants

The study enrolled 28 college juniors (16 females and 12 males), all of whom were native English speakers from the MIT campus, and 4 professional MIT career counselors (3 females and 1 male) who had an average of over five years of professional experience as career counselors and advanced graduate degrees in professional career counseling.

3.2.3 Procedure

The students were recruited through flyers and emails. They were told that they would have the opportunity to practice interviewing with a professional career counselor and would receive $10 for their participation. They were also informed that their interview would be recorded. Male participants were paired with the male counselor, and female participants were paired with one of the female counselors in order to minimize gender-based variability in behavior, as discussed by Nass et al. (Nass, Moon, & Green, 1997).

3.2.4 Interaction

MIT Career Service has a set of 18 behavioral questions that they feel are relevant during the interviews. A list of those questions is provided below:

1. *Tell me about yourself.*
2. *Why do you want this job?*
3. *Why do you want to work for this company?*
4. *Describe a time when a team member came to you for help. What was the situation? How did you respond?*
5. *Tell me about a time when you had to deal with someone whose personality was different from yours.*
6. *Tell me about a leadership role you had in an extracurricular activity. How did you lead?*
7. *Describe a time when you saw a problem and took action to correct it rather than waiting for someone else to do so.*
8. *Give me examples to convince me that you can adapt to a wide variety of situations, people, and environments.*
9. *What makes you the best person for this job? Tell me why I should hire you*
10. *Tell me about your most successful presentation and what made it so.*
11. *Tell me about a time when you worked on a team. What role did you play on the team?*
12. *Tell me about a meeting where you provided technical expertise. How did you ensure that everyone understood?*
13. *Who or what has had the greatest influence in the development of your career interest?*
14. *What types of situations put you under pressure and how do you handle them?*
15. *What are your strengths?*
16. *Tell me a weakness of yours.*
17. *What are your long-term goals?*
18. *What questions do you have for me?*

I sought feedback from the 4 MIT career counselors to narrow the list to 5 questions. A visual example of the sequence of questions being asked during the interview by the MIT career counselor is provided in Figure 3-4.

3.2.5 Data Analysis and Findings

To better understand the agreement among the behavioral dimensions of interview, the data needed to be annotated. Many coding schemes have been developed to annotate non-verbal corpora. For example, Kipp et al. (Kipp, Neff, & Albrecht, 2008) have proposed a coding scheme for gesture phases. MUMIN multimodal coding scheme (Allwood, Cerrato, Jokinen, Navarretta, & Paggio, 2007) is an example of an exhaustive multimodal coding system, while Facial Action Coding System (FACS) (Ekman & Friesen, 1978) remains the most widely used. The developed coding scheme is shown in Table 3-1 and Table 3-2.

There were two goals to the annotations: 1) To understand the behavioral dimensions of interviews that are easy for others to agree on; 2) To understand the nonverbal and backchanneling behaviors of the counselor.

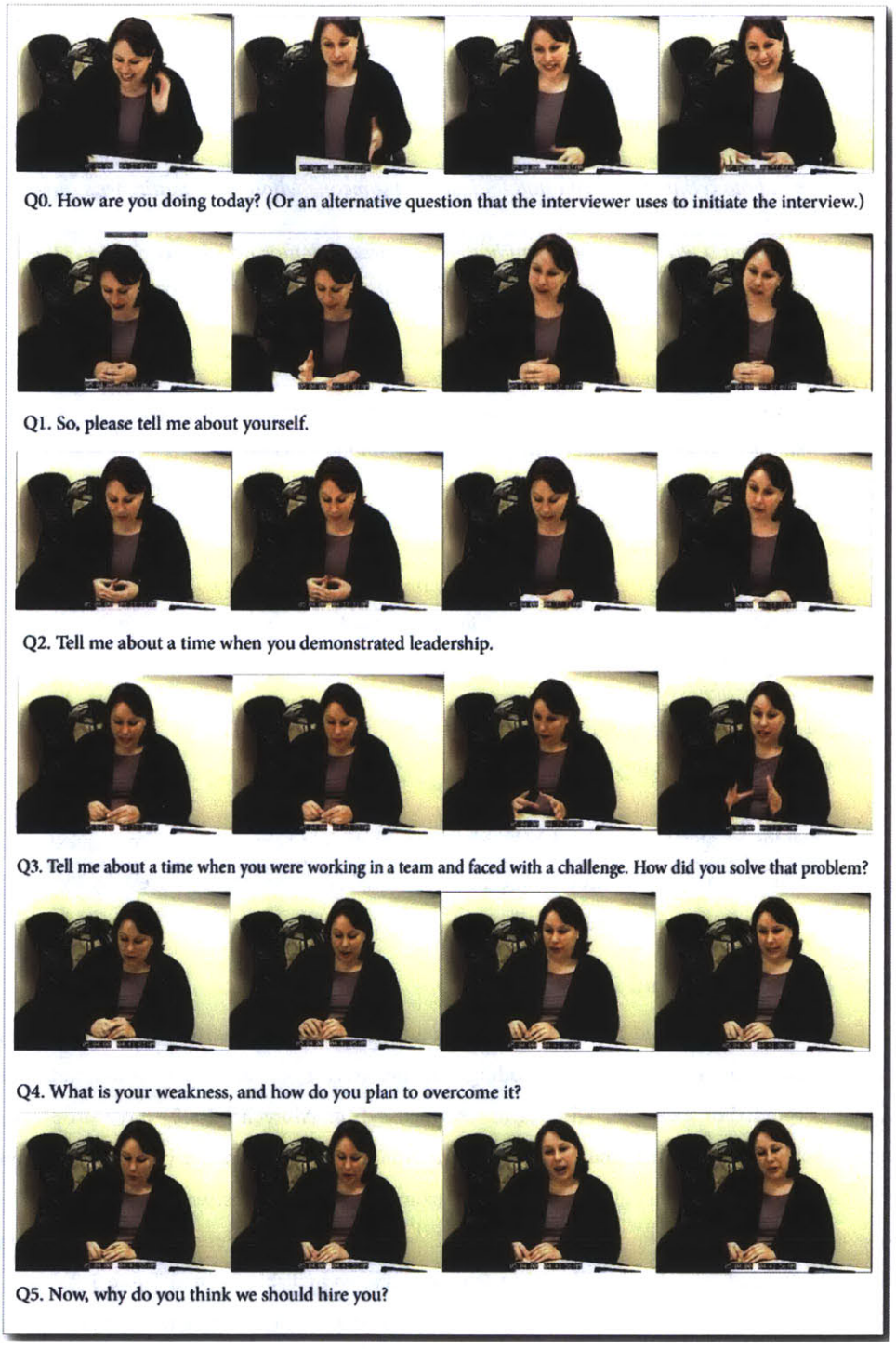


Figure 3-4. The sequence of interview questions being asked by the MIT Career Counselor.

Table 3-1. The coding scheme for the nonverbal behaviors. Coders used only video (no audio) to code on these dimensions for the entire video segment.

NONVERBAL RATINGS (USE ONLY VIDEO)	Description	Rating
Comfortable	extent to which P seems comfortable	1 = very uncomfortable, 7 = very comfortable
Anxious	extent to which P seems anxious	1 = very calm, 7 = very anxious
Eye contact	how much P makes eye-contact (looking at interviewer)	1 = avoidant, 7 = constant/intense eye-contact
Furrowed brow	extent to which P has a furrowed brow	1 = not at all, 5 = all the time
Pursed lips	extent to which P has pursed lips	1 = not at all, 5 = all the time
Face touching	number of times P touches face	COUNT
Neck touching	number of times P touches neck	COUNT
Fidgeting	rapid or repetitive movements, self-touching, playing with objects/clothes	1 = no fidgeting, 5 = constant fidgeting
□ *This is in contrast to gesturing. Fidgeting is awkward and distracting.*		
Gesturing	extent to which P gestures with hands and arms	1 = no gesturing, 5 = a lot of gesturing
Expressions	extent to which P shows facial expressions	1 = very few expressions, 5 = a lot of expressions
Laughing	Identify the time when P laughs	Timing
□ *Note whether this is happy laughter or nervous laughter.*		
Expansive Posture	extent to which P has an expansive vs. constrictive posture	1 = very constrictive, 7 = very expansive
- Expansive: Arms apart, gesturing a lot; Constrictive: Arms crossed and/or near torso, shoulders hunched		
- Make a note if the hands or legs are crossed in the "Notes" column of the spreadsheet.		
Body Orientation	did the P turn his shoulders or body away from the interviewer?	COUNT and 1-10 scale, 1 = facing directly, 10 = turned away

The data of 28 interviewees was manually analyzed and annotated by two Facial Action Coding System (FACS) (Ekman & Friesen, 1978) trained coders (one male and one female). The coders were not part of the job interview experiment and did not know the interviewees or the interviewers.

The agreement between the coders was measured using Cronbach's alpha (Cronbach, 1951), as shown in Table 3-3. The value of alpha ranges from 0 to 1, where $\alpha \geq .9$ means excellent, $0.8 \leq \alpha < 0.9$ means good, $0.7 \leq \alpha < 0.8$ means acceptable and $\alpha < 0.6$ is questionable (Kline, 2000).

The less ambiguous categories (categories with measureable easy to follow actions) such as hair touching, face touching, or neck touching elicited the highest agreement. The categories with highest disagreement included nervous laughter and lip movements.

Table 3-2. The coding schema for the high level behavioral dimensions. Coders used audio and video to code these following dimensions.

Factor	OVERALL RATINGS (USE BOTH AUDIO & VIDEO)	Description	rating
Overall	Performance	how good were the answers?	1 = awful, 7 = amazing
Overall	Authenticity	how authentic did the P come across?	1 = very inauthentic, 7 = very authentic
Overall	Recommended for Hire	do you think the P *should* get the job?	1 = definitely no, 3 = definitely yes
Warmth	Like to be friends	extent to which you would want to become friends with P	1 = no interest, 7 = a lot of interest
Warmth	Nice	(good natured, sincere)	1 = not at all nice, 7 = very nice
Warmth	Friendly	extent to which P seems friendly (this is different from "Nice". Someone can interact in a friendly way, but not come across as a good natured person)	1 = very unfriendly, 7 = very friendly
Content	Qualified	extent to which P seems qualified for the job he/she wants	1 = not at all qualified, 7 = very qualified
Content	Straightforward	how straightforward and clear were the answers and the points being made?	1 = not at all clear, 7 = very straightforward
Content	Structured	how structured, planned out, and logical were the answers?	1 = very unstructured, 7 = very structured
Competence	Competent	extent to which P seems competent and capable	1 = very incompetent, 7 = very competent
Competence	Intelligent	extent to which P seems smart and intelligent	1 = very unintelligent, 7 = very intelligent
Presence	Nervous/anxious	extent to which P seems nervous or anxious	1 = very calm, 7 = very nervous/anxious
Presence	Awkward	extent to which P seems awkward	1 = not at all awkward, 7 = very awkward
Presence	Captivating	extent to which P's answers captured your attention	1 = not at all captivating, 7 = very captivating
	- In other words, how easy or difficult did you feel it was to pay attention to the answers?		
Presence	Confident	extent to which P seems confident	1 = very timid/uncertain, 7 = very confident
Presence	Engaged	how interested P seems in the task (not bored/apathetic)	1 = very bored, 7 = very engaged/interested
Presence	Enthusiastic	extent to which P shows energy and excitement	1 =not at all enthusiastic, 7 = very enthusiastic

Table 3-3. Agreement between two coders on the nonverbal behaviors of the interviewees using Cronbach's alpha.

Criteria	Cronbach's alpha
Nervous laughter	-0.396
Pursed lips	-0.025
Expressions	0.141
Furrowed brow	0.185
Orientation	0.438
Body orientation	**0.679**
Fidgeting	**0.682**
Comfortable	**0.69**
Expansive posture	**0.698**
Eye contact	**0.745**
Gesturing	**0.816**
Structured	**0.826**
Neck touching	**0.926**
Happy Laughter	**0.93**
Face touching	**0.979**
Hair touching	**0.987**

Table 3-4. Agreement between the coders on the high-level behavioral dimensions using Cronbach's alpha.

Criteria	Cronbach's alpha
Recommended for Hire	0.507
Like to be friends	0.513
Authenticity	0.526
Straightforward	0.566
Nice	**0.604**
Competent	**0.635**
Intelligent	**0.64**
Anxious	**0.644**
Qualified	**0.727**
Friendly	**0.819**
Overall Performance	**0.83**
Engaged	**0.893**
Captivating	**0.924**
Confident	**0.928**
Enthusiastic	**0.931**
Nervous/anxious	**0.943**
Awkward	**0.944**

On the high-level behavioral dimensions, the highest agreements were achieved in dimensions of "Awkward," "Nervous," "Enthusiastic," "Confident," "Captivating," "Engaged," "Overall Performance," "Friendly," and "Qualified". The lowest agreements were in "Recommended for Hire," "Like to be friends," "Authenticity," "Straightforward," "Nice," and "Competence". It is interesting to see that the coders agreed the most in all the categories that came from the factor of *Presence*. All the categories in *Presence* involve nonverbal behavior including awkwardness, nervousness, enthusiasm, confidence etc.

The next highest agreements came in "Overall Performance" (alpha = .83) (part of the *Overall* factor), "Friendly" (alpha =.81) (part of the *Warmth* factor) and "Qualified" (alpha = .72) (part of the *Content* factor). Based on these data, it could be argued that it is possible for independent human judges to identify behaviors that could encompass factors such as *Presence*, *Overall*, *Warmth* and *Content*. It was not surprising to see that the coders disagreed the most on "Recommended for hire" as it could be very specific to a job description. This could also be used to explain why coders disagreed on the factor of *Competence*, as it is also dependent on the job. Coders also disagreed on 'Like to be friends" category.

Along with interviewees, two specific behaviors of interviewers were also coded. One of them was head nods, which was manually coded by one of the FACS coders. The other one was smiles, which was automatically coded. The following three behaviors on *listening*, *expressions* and *acknowledgements* emerged from the analysis of the interviewer's data.

Listening behavior

In almost all the interactions, the counselors of both sexes asked a question, carefully listened to the answer, briefly acknowledged the answer, and then moved on to the next question. The listening behavior of the counselor included subtle periodic head nods and occasional crisscrossing of the arms. Figure 3-5 provides cumulative frequencies of counselors head nods averaged over 28 sessions. All the videos were stretched to last between 0 and time 100 (*average duration of interviews*: 330 seconds, *SD:* 110.6). A "nod" is shown if 14 or more of the videos had a nod at that moment. Figure 3-5 demonstrates that counselors seemed to nod their head periodically (in every ~4.12 seconds, calculated before stretching the videos to 0:100) while listening to the participants during the mock interview.

Expressiveness

Our data and observations suggested that counselors maintained a neutral composition during the interviews; however, counselors also matched the expressions of the interviewees by reciprocating smiles and other behaviors. Figure 3-6 demonstrates the averaged smile intensity (smile intensity of 100 means a wide smile, 0 means no smile) of 28 participants.

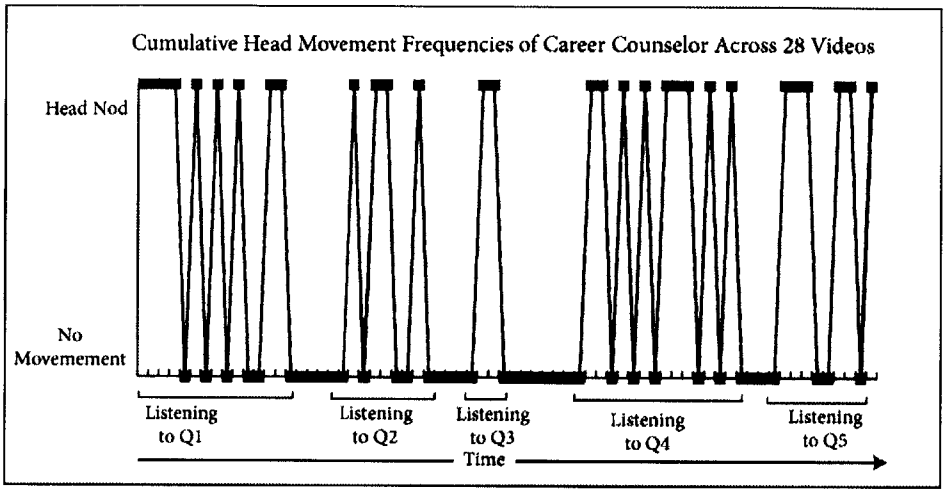


Figure 3-5. Cumulative Head Movement frequencies of Career Counselors Across 28 videos. Q1, Q2, Q3, Q4, Q5 correspond to questions 1, 2, 3, 4, 5.

The data were divided based on the *interview* and *feedback* segment. During the *interview*, the career counselor would ask interview questions and anticipate answers, and during the *feedback*, the career counselor would provide feedback on the overall interview. There was not much difference in the smiles of the participants during the *interview* (average smile intensity: 31.17, *SD* =4.36) and *feedback* (average smile intensity: 32.95, *SD* =7.46). However, counselors seemed to smile less during the *interview* (average smile intensity: 18.57, *SD* = 4.11), but smile more during the *feedback* (average smile intensity: 32.37, *SD* = 6.14), as shown in Figure 3-7.

Acknowledgements

The counselors used a similar set of acknowledgements at the end of each answer provided by the interviewee. Examples include "That's very interesting," "Thanks for that answer," "Thank you," and "I can understand that."

The 28 interview sessions were also manually transcribed using Mechanical Turkers. The transcriptions were manually verified for accuracy. A simple word frequency analysis was applied on the 28 transcribed interview sessions. The Figure 3-8 illustrates the most frequently occurred words being expressed with their font size (more frequent, the larger the font size). Words such as "like", "um", "uh", "just", which are known as *fillers*, were the most common words among the MIT interviewees.

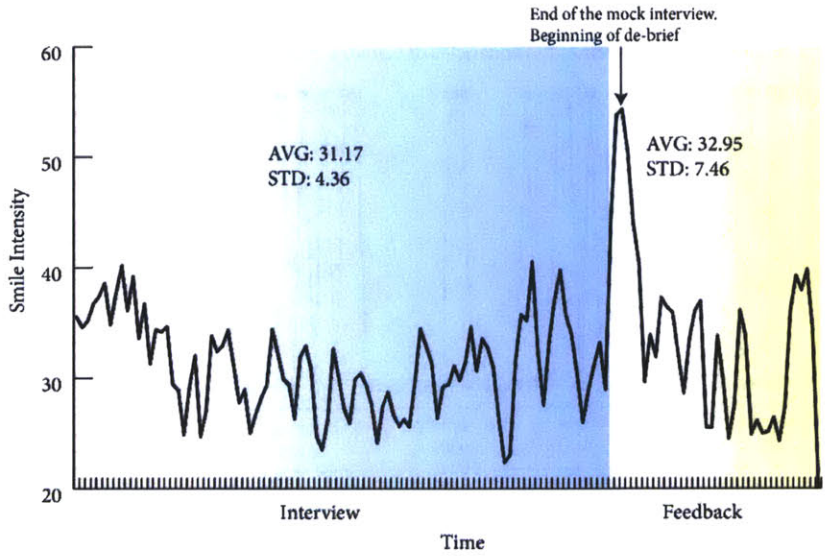


Figure 3-6. Average Smile Intensity of 28 student participants across 28 sessions. Participants seem to smile the same amount during the interview and during the feedback.

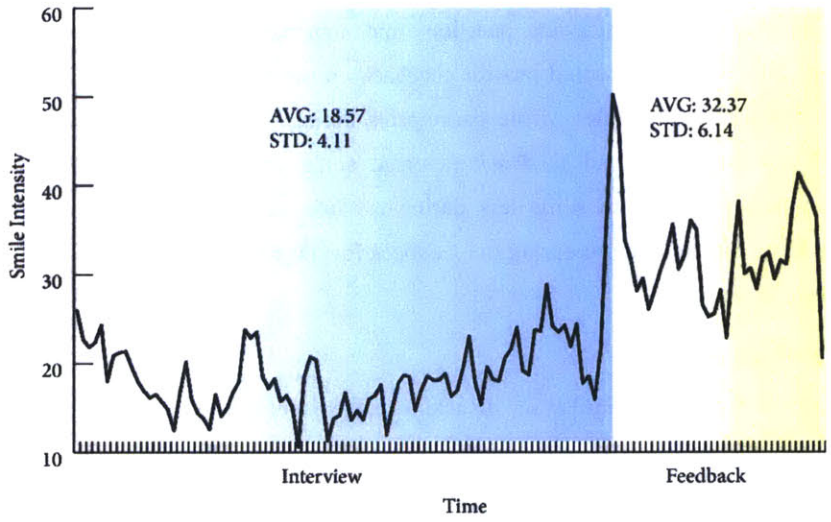


Figure 3-7. Average Smile Intensity of 4 Counselors across 28 sessions. All the counselors were less expressive with reciprocal smiles during the interview.

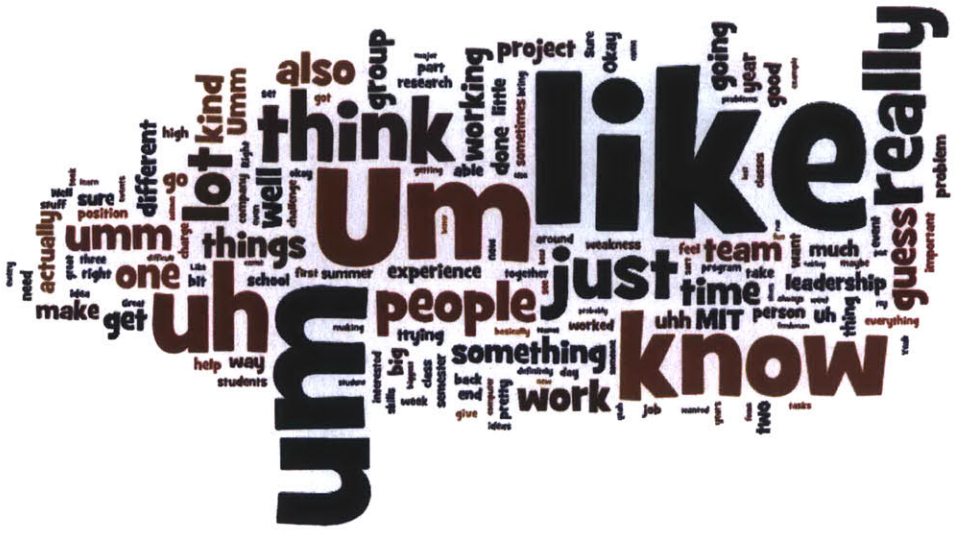


Figure 3-8. The most frequently occurred words across 28 interview sessions, being displayed as word bubbles. The most frequently words are "like," "um," know," "just," etc. This figure was generated using wordle.com

3.3 CHAPTER SUMMARY

In this chapter, I provided the contextual inquiry of job interviews. Through an extensive literature search, it has been shown that the most important aspect of a job interview is the social interaction.

I presented results from an experiment with 28 MIT undergraduate students and 4 career counselors from MIT career services. The annotations of the behavioral data by two FACS trained human judges indicate that it is possible for them to agree on behavioral factors of *overall* (e.g., "Overall performance"), *Content* (e.g., "Qualified"), *Warmth* (e.g., "Friendly") and *Presence* (e.g., "Engaged," "Captivating," "Confident," "Enthusiastic," "Nervous/anxious," "Awkward") in the context of a job interview. This process will be used to inform our next step of developing reliable measures to validate the effectiveness of interviews and behavior changes.

The analysis of the behavior of the interviewers and the interviewees demonstrated that the interviewer maintained a neutral face during the interview, and periodically nodded every ~4.12 seconds. Interviewers did, however, smile more during the debrief portion of the interview. The analysis revealed insightful information about backchanneling, which could be used to model the behavior of an automated virtual counselor, to be discussed in the later chapters.

The next chapter provides an exploration of developing relevant technologies to understand and interpret the nonverbal behaviors using natural data.

Chapter 4: Technology Design, Development, and Explorations

The design and implementation of an autonomous virtual agent that could "see," "hear," "respond," and "provide feedback" on nonverbal behavior combines areas of computer vision, speech processing (speech recognition and prosody analysis), machine learning, virtual agents, and automated behavior synthesis. This chapter provides exploratory details on extending our understanding of nonverbal behavior, along with practical considerations of implementing the deeper insights into working prototypes.

From the background section, we learned that a smile is an extremely useful expression in job interviews. Interviewees who smile more are perceived to have better social skills, and appear more likeable (Shlenker, 1980) (Gifford et al., 1985). That is why one of the considerations of technology development was to analyze smiles. First, the study zooms into recognition of smiles. But a smile is a multi-purpose expression. We smile to express rapport, delight, polite disagreement, sarcasm, and even frustration. Is it possible to develop computational models to distinguish among smiling instances when delighted, frustrated, or just being polite? Through a series of studies, I demonstrate that it is useful to explore how the patterns of smiles evolve through time, and that while a smile may occur in positive and in negative situations, its dynamics may help to disambiguate the underlying state.

Head gestures are known to convey information about personality and attention as argued by Mignault & Chauduri (Mignault & Chaudhuri, 2003). For example, the head being directed upward could mean dominance, while upward head tilts also could convey non-dominance (Mignault & Chaudhuri, 2003). Nodding of the head while the interlocutor is speaking is also considered a sign of paying attention (Tickle-Degnen & Rosenthal, 1990). Therefore, as part of the exploration, I investigated the process of automatically recognizing head gestures (e.g., head nods, shakes, and tilts).

In face-to-face interactions, information from vocal cues could contain evidence about emotional state and personality (Kramer, 1963). For example, a rapid speaking rate could be interpreted as confidence. However, it is also possible to equate a fast speaking rate with a nervous candidate rushing to wrap up. Pausing in between sentences is also very important. Depending on the context, a job candidate could use pauses to build up suspense. However, if not used well, awkward pausing could be a sign of embarrassment or lack of confidence. Pitch and volume of speech could also possibly aid in clarifying messages, set the stage to ask

further questions, or establish an ideal tone. Finally, fillers such "ah" and "um" are distractions that may express nervousness or lack of sincerity during a job interview. Motivated by these observations, in this chapter I provide theoretical background along with practical implications of developing technologies that can automatically measure features such as speaking rate, pauses, pitch, volume, and fillers in conversations.

Another important aspect of nonverbal communication during face-to-face interactions (e.g., job interviews) is establishing rapport (Tickle-Degnen & Rosenthal, 1990). It is the basis of understanding and communicating with the interlocutor. People can use nonverbal behaviors such as a smile, an open stance, and eye contact to establish rapport. Another technique for nonverbal communication that helps build rapport is mirroring. As we are developing an automated conversational virtual agent to play the role of the interviewer, it is very important that the agent is able to exhibit the appropriate behaviors during the interview. That includes being able to automatically use posture and arms, speak, follow the gaze, backchannel when necessary, and mirror certain behaviors to establish rapport. This chapter provides details on the process that we followed to develop those *behavior synthesis modules* in real-time under practical conditions.

All the technical details on the relevant components are provided so that the design and implementation process are repeatable. Implementation details for each sub-system are briefly outlined below.

4.1 SMILE ANALYSIS

This section describes how we recognize smiles of people given a video along with evaluation bench marks.

4.1.1 Algorithm

From the video of the user's face, we tracked smiles in every frame. We used the Sophisticated Highspeed Object Recognition Engine (SHORE) (Froba & Ernst, 2004) (Kueblbeck & Ernst, 2006)(Kublbeck, Ruf, & Ernst, 2009) API by Fraunhofer to detect faces and facial features in order to distinguish smiles. The classifier was trained using the Adaboost algorithm with sample images, considering smiling and neutral as the binary classes. Features from all over the face were used for boosting. The outcome of the classifier was a normalization function that projected the score onto a range, [0,100], by analyzing the entire face including mouth widening, zygomaticus muscles, orbicularis oculi, and other regions of the face in every frame. Thus, each face image was scored from 0 to 100, where 0 and 100 represented no smile and full smile, respectively. For the remainder of this thesis, the score is referred to as the *smile intensity*.

4.1.2 Evaluations

I evaluated this smile recognition system using the Cohn-Kanade dataset (Kanade et al., 2000), which includes 287 images of 97 individuals from the United States. The evaluation was expressed in terms of precision, recall and F-scores. The F-score (the ratio of geometric mean and arithmetic mean of precision and recall) provides the coherence between the precision and recall values of the model and is a very good indicator of the reliability (higher F-score implies a better and more reliable model) of the predicted values.

$$F-score = 2 * \frac{precision * recall}{precision + recall}$$

Out of these images, 62 were labeled as happy, where happiness is defined as smiling at various levels. The testing of the smile analysis module for classifying images labeled as happy and the remaining images in the Cohn-Kanade dataset yielded precision, recall, and F-score values of .90, .97, and .93, respectively (see Table 4-1).

In addition, I tested the smile module on the JAFEE 178-image dataset (Lyons, Akamatsu, Kamachi, & Gyoba, 1998) of happiness, sadness, surprise, anger, and disgust, including 29 instances of happy faces from 10 Japanese women. The results from the testing with the JAFEE dataset yielded precision, recall, and F-score values of .69, 1, and 0.81, respectively.

Table 4-1. Evaluation results of the Smile Analysis Module on Cohn-Kanade and JAFEE dataset

Dataset	Precision	Recall	F-score
Cohn-Kanade	.90	.97	.93
JAFEE	.69	1	.81

4.2 FACIAL ANALYSIS: CHALLENGES FOR COMPUTERS

This section is an exploration involving four consecutive experiments. The basic motivation behind this section was to understand the details of two fundamental affective states that are important to correctly distinguish in general interactions: frustration and delight. This section illustrates how the exploration led to answering questions of how acted data are different than spontaneous data, the morphological differences of smiles, and an algorithm to automatically classify them.

Experiment 1 (section 4.2.1) and Experiment 2 (section 4.2.2) describe two experimental situations to elicit two emotional states: the first involves recalling situations while expressing either delight or frustration; the second experiment tries to elicit these states directly through a frustrating experience and through a delightful video. I find two significant differences in the nature of the acted vs. natural occurrences of expressions. First, the acted

ones are much easier for the computer to recognize. Second, in 90% of the acted cases, participants did not smile when frustrated, whereas in 90% of the natural cases, participants smiled during the frustrating interaction, despite self-reporting significant frustration with the experience (this is explained in section 4.2.3). To further motivate Experiment 3 (section 4.2.4), the reader is invited to look at Figure 4-2 and guess in which frames the participants were frustrated and delighted. Answers are provided at the caption of the Figure 4-2.

It is possible for people to smile under both natural frustration and delight conditions. Is there something measurably different about the smiles in these two cases, which could ultimately improve the performance of classifiers applied to natural expressions? Experiment 3 (section 4.2.4) attempts to answer that question.

Finally, Experiment 4 (explained in section 4.2.5) explores the morphological patterns of polite and amused smiles in the context of a banking environment, and provides new findings about smile dynamics and smile symmetry for shared vs. unshared smiles.

4.2.1 Experiment 1: Acted Data

The experiment took place in a well-lit empty room where participants were expected to interact with a computer program. The participants interacted with the computer program, which consisted of a 2D image of an avatar (Figure 4-1). During the interaction, the avatar would ask a sequence of questions. The questions would appear in the form of text on the interface (Figure 4-1). The participants would wear a headset and speak directly to the avatar to answer the questions. Additionally, there was a video camera to capture the face of the participant. The exact interaction between the avatar and the participant was as below:

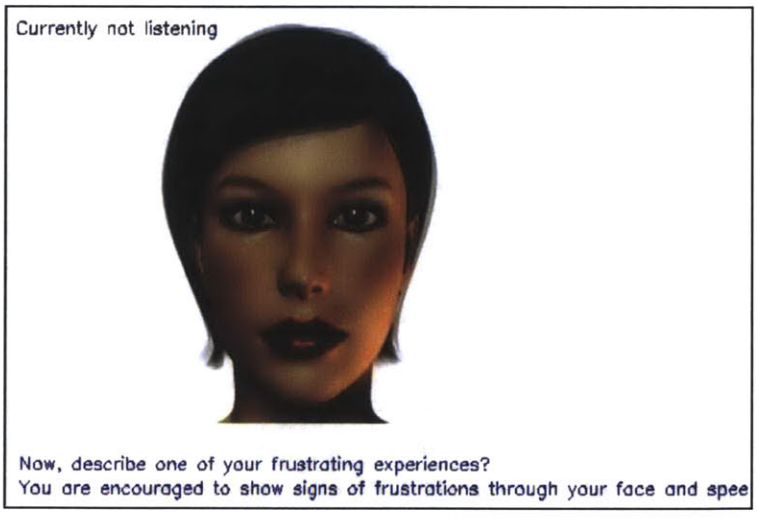


Figure 4-1. 2d image of the computer program used in the "Acted data experiment"

Figure 4-2. Four participants from elicited dataset, each smiling while being in either a (i) frustrated or (ii) delight state. (a), (d), (f), (h) are taken from instances of frustration; (b), (c), (e), (g) are from instances of delight. I conducted an independent survey of 12 labelers of these, and all scored at or below chance (4 out of 8, or 50%) in labeling images in this Figure.

Avatar: Hi There! I am Sam. I hope to be a 3D character someday. But today, I am just a 2D image who would like to interact with you. (Pause for 15 seconds)

Avatar: I hope you have signed the participant agreement form. If yes, please say your participant number. Otherwise, just state your name. (Avatar waits for the participant to speak and finish).

Avatar: Please briefly say a few sentences about why you are interested in this study? (Avatar waits for the participant to speak and finish)

Avatar: Now describe one of your most frustrating experiences. You are encouraged to show signs of frustration through your face and speech. (Avatar waits for the participant to speak and finish)

Avatar: Now describe one of your most delightful experiences. You are encouraged to show signs of delight through your face and speech. (Avatar waits for the participant to speak and finish).

Participants

The "Acted Data Experiment" consisted of 15 participants – 10 male and 5 female. All of them were employees at a major corporation and their age ranged from 25-40. From 15 participants, we gathered 45 clips of frustration, delight and neutral expressions (3 clips from each participant). The average duration per clip for delight and frustration was over 20 seconds, whereas the average duration for neutral was around 10 seconds. Participants wore a Logitech ClearChat Comfort USB Headset to communicate with the avatar. The frontal face of the participant) was recorded using a Logitech 2 MP Portable Webcam C905. Logitech webcam software was used to connect the webcam with the PC providing 30 frames per second. The participants were left in an empty office to finish the task, as shown in Figure 4-3.

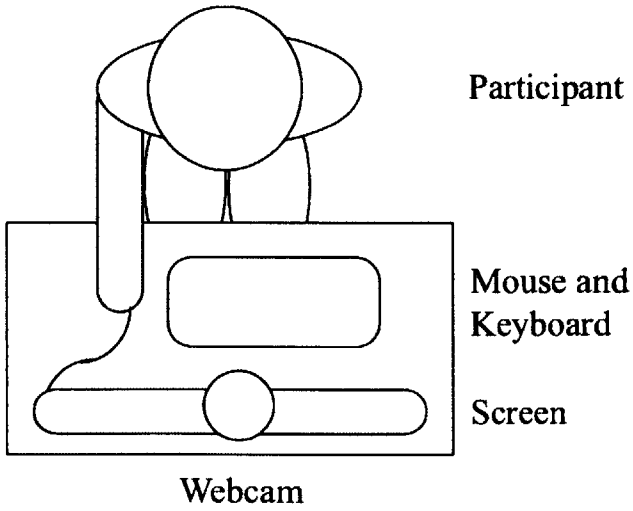


Figure 4-3. Experimental set up for Acted and Elicited Data Experiment.

4.2.2 Experiment 2: Elicited Data

For this study, 27 new participants were recruited. The participants were not part of the "Acted data experiment" and were blind to the hypothesis. The participants were told that they would have to evaluate the usability of a web form, and provide suggestions for improvement, if necessary. After the participant entered the room, the participant was told that s/he would have to fill out a web form. They were also instructed that based on how the task progressed; the participant may or may not be asked to speak to the camera to provide feedback on the form. The form contained 10 biographical questions (details in Table 4-2), including a field for date and current time without instructions on the format. The participants were instructed not to leave the experiment room until they navigate to the confirmation screen of the form (screen 16 of Table 4-2). The exact sequence of interactions between the form and the participant is provided in Table 4-2.

All the text messages in Table 1 were converted into .wav files. As the participants navigated from one screen to another, the interface would read the text message out loud. The texts were converted into .wav files using ATT's publicly available text to speech engine with a female American accented voice. Initially, the participants are asked two questions (screens 3 and 4 of Table 4-2), one after another. The purpose of those questions was to elicit expressions that were more likely to be neutral. The reason I opted for two consecutive questions is because during the pilot study I noticed that a lot of participants felt awkward looking at the camera for the first time. As a result, they either laughed out of embarrassment or provided a very brief answer, when asked, "Why are you participating in this study?" Adding a follow up question in the next screen helped them to loosen up, which resulted in a more neutral answer for the second question. I have seen this "first time recording an expression" effect dominate expressed emotions regardless of which emotion the stimuli were designed to elicit, and I encourage scientists to consider this when designing emotion elicitation experiments.

The biographical forms (screens 5, 7, 9 in Table 4-2) contained a timer that started counting the elapsed time. I intentionally put the timer in the middle of the screen in large font. Right mouse click and CTRL keys of the keyboard were disabled to prevent participants from copying content from one screen to another. The claim that 94.5% of the previous participants were able to finish this study in less than 2 minutes was a made up number to put more pressure on the participants. After three attempts to submit the form, the participants eventually reach screen 10 where, they are asked to solve a CAPTCHA to move forward. I used Google images (images.google.com) to select a few nearly impossible CAPTCHAs for this study. Therefore, regardless of whatever the participants typed, the interface kept presenting an error message asking participants to solve another CAPTCHA. After 3 trials,

Table 4-2 The sequence of screens for the natural experiment. The same sequence was maintained for all the participants.

Screen	Purpose	Message
1	Welcome screen	Click here to move on with this study.
2	Greetings to welcome the participant	Hi there! I hope you are doing well. Please click here to move forward with this experiment.
3	Elicit a neutral expression (Neutral)	Can you look at the camera and say a few sentences about why you are participating in this study? Please click here when done.
4	Elicit a neutral expression (Neutral)	Thank for your kind participation in this study. Before we move on, there is one more thing. Can you again look at the camera and say a few sentences about your regular activities in this department? Please click here when done.
5	Biographical form	Before you move on with this study, fill out the form below. 94.5% of the previous participants in this study were able to do this in less than 2 minutes.
6	ERROR	Error: You either did not enter the date or entered it in wrong format (correct format is: Month/Day/Year, Hour: Minute, AM/PM)
7	Biographical form	Before you move on with this study, fill out the form below. 94.5% of the previous participants in this study were able to do this in less than 2 minutes.
8	ERROR	Error: Your "About Me" section did not contain the minimum of 500 characters.
9	Biographical form	Before you move on with this study, fill out the form below. 94.5% of the previous participants in this study were able to do this in less than 2 minutes.
10	Confirmation	Your form has been submitted. Since you took a few trials to submit this form, please solve the following CAPTCHA to move forward.
11	ERROR	ERROR: Wrong values entered. Please solve this CAPTCHA to move forward.
12	ERROR	ERROR: Wrong values entered. Please solve this CAPTCHA to move forward.
13	Feedback (Frustration)	Since you are one of those participants who could not finish the form within 2 minutes, we want your feedback. Look at the camera and say a few things about why you could not finish the form within 2 minutes, unlike most of the participants.
14	Prepare for the next phase	Wonderful!! Thanks for your honest feedback. For the next phase of the experiment, you will be asked to share an experience from your past that you think is funny and delightful. To help you get started, I am sharing a click from youtube which hopefully will put you in the right mood. When ready, click here to move to the next screen and share the experience.
15	Share an experience (delight)	Now please look at the camera and share a funny experience from your past.
16	Thank you	Thank you! Your study has been completed!

the participants would reach screen 13, where the interface would prompt them to provide feedback on what they had done wrong and why they were unable to finish the form in less than 2 minutes unlike most participants. In this phase of the study, I expected the participants to be somewhat frustrated and demonstrate signs of frustrations either through their face, speech or both.

In screen 14, participants begin the second phase of the study. In this phase, participants were given time to relax a bit and think of a funny experience that they would have to share momentarily. To help them transition into a relaxed state of mind, the interface shows them a funny YouTube video of a baby laughing uncontrollably. This particular video has more than

11 million views since 2006 and can be viewed through this link http://tinyurl.com/tac-affective. This video was picked because I felt that laughing is contagious and it may help to distract the participants from their frustrating experience of filling out the form. At the end of the experiment, majority of the participants mentioned even though they had watched the video before, they still found it funny and exhilarating. After the end of the interaction with the web form, we set up a post de-briefing session asking each participant to self-report how frustrated and delighted they were using a scale of 1-10, while they were filling out the form and watching the funny video. The entire interaction was recorded using a Canon 3.89 MP VIXIA HF M300 Camcorder and an Azden WMS-PRO Wireless Microphone. The Canon VIXIA HF M300 captured video at 30 frames per second.

The recorded footage was split into two different categories: 1) "The Feedback" (contains both audio and video) 2) "Interaction" (contains only video, but no audio). Feedback dataset consisted of facial expressions and speech data of participants as they directly spoke to the camera with their feedback regarding the form and sharing a funny experience (e.g., screens 4, 13, and 15 of Table 4-2). The Interaction dataset consisted of clips of participants when they were either filling out the form or watching the YouTube video (e.g., screens 5, 7, 9 and 14 of Table 4-2).

Participants and dataset

There were a total of 27 graduate students who participated in this study. Five of them were female and 22 male. All of them were blind to the hypothesis of this study. In post-experimental de-briefing, three participants informed us that they were able to figure out that the forms were intentionally designed to be buggy to provoke frustration from them. Since they were able to determine the objective of the study, we eliminated their data, resulting in 24 clips of frustration for the "feedback" dataset. Four of our participants were unable to remember a funny experience from their past during the experiment. Two of the participants told us in the de-briefing that they were so frustrated filling out the form that they were reluctant to share a delightful experience to the camera. As a result, from 27 participants, I ended up having 21 clips of delight for the "feedback" dataset. For neutral expressions, I only considered expressions from screen 4, as indicated in Table 4-2, and ignored the expressions elicited in screen 3. Therefore, I had 27 instances of neutral expressions for the "feedback" dataset. The average length of each clip in the "feedback" dataset for frustration and delight was a little over 30 seconds, and for neutral it was around 15 seconds.

4.2.3 Analysis: Acted vs. Elicited Feedback

In analysis 1, I take acted instances of frustration, delight and neutral from experiment 1 and naturally elicited instances of frustration, delight, and neutral from the "feedback" dataset

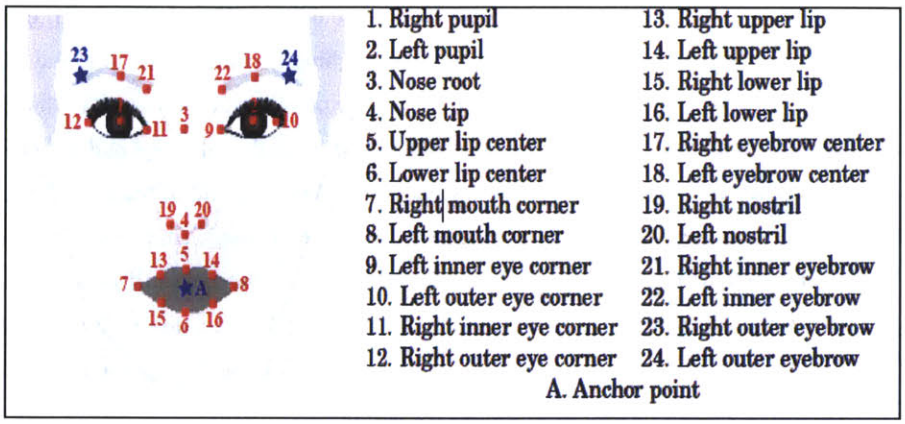


Figure 4-4. Extracted feature points of the face using Google Tracker

of experiment 2. The goal was to allow for a comparison of recognition results on both acted and elicited data, where both facial expressions and speech were present. Below are the descriptions of the facial and speech features used for classification.

Face Analysis

Google's facial feature tracker (formerly known as Nevenvision) ("FaceTracker. Facial Feature Tracking SDK.," 2002) was used to track 22 feature points: 8 points surrounding the mouth region, 3 points for each eye, 2 points for each eye-brow, and 4 points for two nostrils, nose tip, and nose root. Points 23 and 24 shown in Figure 4-4 were extrapolated.

Raw distances (in pixels) as well as their standard deviations across facial feature points were calculated. For example, distances and standard deviations between 12 and 11, 9 and 10, 2 and 18, 1 and 17, 11 and 21, 9 and 22, 7 and 8, 5 and 6 etc. were calculated.

The local distances among those points as well as their standard deviations were measured in every frame and used as features. Additionally, I have used *smile intensity*. The features extracted per clip were averaged to form a feature vector per clip. In the first experiment with acted data, while trying different techniques, averaging all the features across each clip yielded satisfactory results. Therefore, in the second experiment with naturally elicited "feedback" data, we also averaged all the features across each clip to allow for a valid comparison.

Speech Analysis

Prosodic features related to segmental and supra-segmental information, which were believed to be correlates of emotion were computed. Using *Praat* (Boersma & Weenink, n.d.), an open source speech processing software package, I extracted features related to pitch

(mean, standard deviation, maximum, and minimum), perceptual loudness, pauses, rhythm and intensity, per clip. A details analysis of how they computed is provided in section 4.4.

Final Feature Set

There were 45 clips from experiment 1 and 72 clips from the "feedback" dataset from experiment 2. For each individual clip, I extracted audio and video features and concatenated them in a vector such that each clip's feature vector was as follows: $V_{clip} = \{ A_1,...A_n, F_1,...F_m\}$, where $A_1,...A_n$ are n speech features, and $F_1,...F_m$ are m facial features. In this study, n was equal to 15 and m was equal to 25; features are described below.

Results

Five classifiers (BayesNet, SVM, RandomForest, AdaBoost, and Multilayer Perceptron) from the Waikato Environment for Knowledge Analysis (WEKA) toolbox (M. Hall et al., 2009) were used, to compare the classification accuracy between the elicited face + voice data and the acted face + voice data. There were 45 instances of acted data and 72 instances of naturally elicited feedback data. One sample was removed for each dataset and held out as the test sample. Leave-one-out K-fold cross validation (K=44 for acted, and K=71 for naturally elicited feedback) was applied. The model was trained on K-1 samples, while testing its parameters on the remaining sample, and repeating leaving a different one out each time. Through this iterative process, optimal parameters were chosen and then tested on the unseen test sample. This was repeated for all samples in the dataset yielding 45 test results for acted and 72 test results for feedback dataset for each classifier. Figure 4-5 shows all the classifiers performed significantly better with acted data compared to elicited data (using a leave-one-out test). The highest accuracy for acted data was 88.23% (chance for each category was 15 out of 45 or 33%) while the highest accuracy for naturally elicited feedback data was only 48.1% (chance for delight was 21 out of 72 or 29%, chance for neutral was 27 out of 72 or 38%, and chance for frustration was 24 out of 72 or 33%). The higher accuracy for the acted data held across the models with the average accuracy across all the classifiers for acted data around 82.34%, a value that dropped to 41.76% for the three-class classification of the elicited data.

Additional analysis on the feature vectors for participants from experiment 1 and experiment 2 revealed that in the acted data, close to 90% of the participants did not smile when they were encouraged to show frustration while recalling being frustrated. On the contrary, in the elicited data, close to 90% of the participants did smile when they were frustrated.

The results shown in Figure 4-5 demonstrate significant differences in correctly classifying instances when the frustration, delighted and neutral states are acted as opposed to being elicited. One possible explanation is that acted states seem to contain prototypical facial

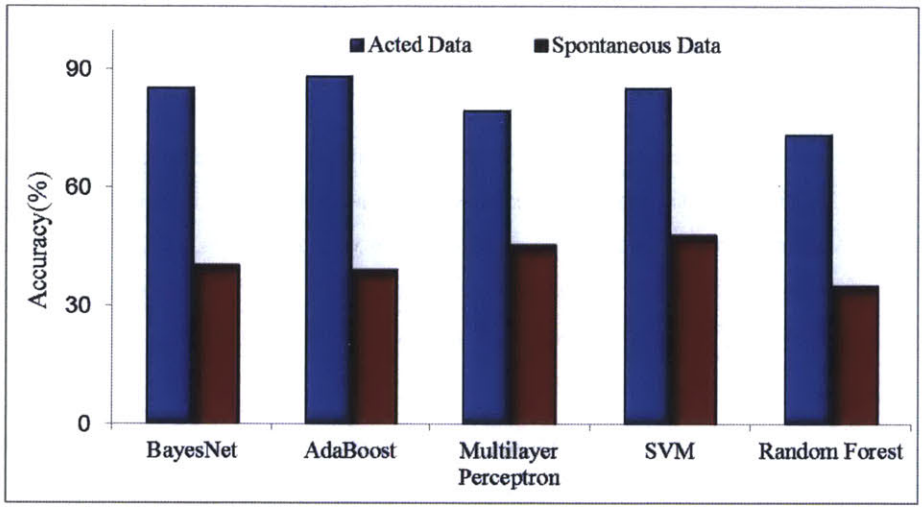


Figure 4-5. Classification accuracy for recognition of frustration, delight and neutral states using various classifiers with elicited and acted data. The accuracy is reported using the leave-one-out method.

features, whereas elicited data may not contain similar facial and speech attributes. That might be why recognizing unique features of affective states and feeding them in a classifier worked fairly well with acted data, but the performance degraded significantly when applied on elicited data. To further stimulate the findings, along with reporting the average, I also conducted an examination of subtle individual differences in terms of expressions. As part of post-analysis, I went through the analysis of each individual to get more insights on whether there are sub-categorical patterns among the participants. Specifically, I zoom into a narrow set of smiles to analyze the intrinsic dynamics of the expressions.

I analyzed each individual clip from the Feedback dataset of Experiment 2 for all the participants was analyzed and revealed interesting findings. I noticed that almost all of the participants, despite self-reporting to be extremely frustrated, did not show the prototypical signs of frustration. In fact, in 90% of the cases, participants showed signatures of delight (e.g., smile) while providing their unpleasant feedback of filling out the form. One possible explanation is that all the participants were MIT colleagues and therefore, they refrained from being impolite given the dynamics of everyday social interaction. However, they were in a room alone during the study. Another possible reason for the greater smiling might be that the population in this study uses smiling to cope with frustration and to keep going. The participants in the second study, MIT graduate students, are all very accomplished and part of what might have helped them get where they are today is that they may have great coping abilities that perhaps use smiling to make them feel better when things go wrong. However,

the participants in the first study, while none were students, were all also accomplished professional researchers at a top industrial research lab and one could argue that they would have similar excellent abilities for coping with frustration, and probably even more experience in doing so. Yet only 10% of them smiled when asked to recall frustrating experiences in Experiment 1.

The occurrences of frequent smiling in elicited frustration may help explain why some people diagnosed with an Autism Spectrum Disorder (ASD) find it hard to make precise sense out of spontaneous facial expressions. If one were taught that smiles mean happiness then it would be easy to mistake smiles from a frustrated person as evidence that things are going great. Subsequently, walking up and smiling to share that person's "happiness" could be misconstrued as insensitivity or worse, and lead to numerous problems.

Almost all of our participants from Experiment 2, whether frustrated or delighted, demonstrated signatures of smile (AU 12) during their interaction. This is problematic data for those who promote that a smile is a strong disambiguating feature between delight and other affective states. To better understand this phenomenon, I analyzed and compared the smiling patterns of each participant when they were frustrated and delighted. Some of the interesting characterizing patterns are plotted in Figure 4-6. A small subset of the participants, as shown in Figure 4-6 (a, b, c), have clear separation of their smiles in terms of magnitude or intensity when they were frustrated and delighted. However, the pattern dissolves immediately when averaged with the rest of the participants. This phenomenon, once again, motivates the need to look at individual differences rather than reporting the average. In the context of delight, the intensity traces in Figure 4-6 (d, e, f, g) demonstrate that some participants gradually progressed into peaks in terms of smile. This finding is very insightful because now it supports the need to analyze the temporal dynamics of the smiles. Another interesting occurrence to observe, especially in Figure 4-6 (g) and Figure 4-6 (f), is that some people could initiate a frustrating conversation with a big social smile and then not smile much for the rest of the conversation. The prevalence of smiles when the participants were frustrated could likely be the social smile that people use to appear polite or even to cope with a bad situation by trying to "put a smile on."

Smiling under the condition of frustration or failure, even though surprising, is not a new phenomenon that I am reporting in this thesis. Paul Ekman mentioned that people often smile when experiencing unpleasant emotions in the presence of others (Ekman, 1982). It has been shown in earlier work (Schneider & Josephs, 1991) that preschoolers tended to demonstrate more true smiles in the sense of a "Duchenne" smile (Lip Corner Pull or AU 12, and cheek raised or AU 6) when they failed as opposed to when they succeeded. In this study, I observe that people smile in unpleasant and frustrating situations when they interact with computers. As it has been argued that interactions between people and computers are social

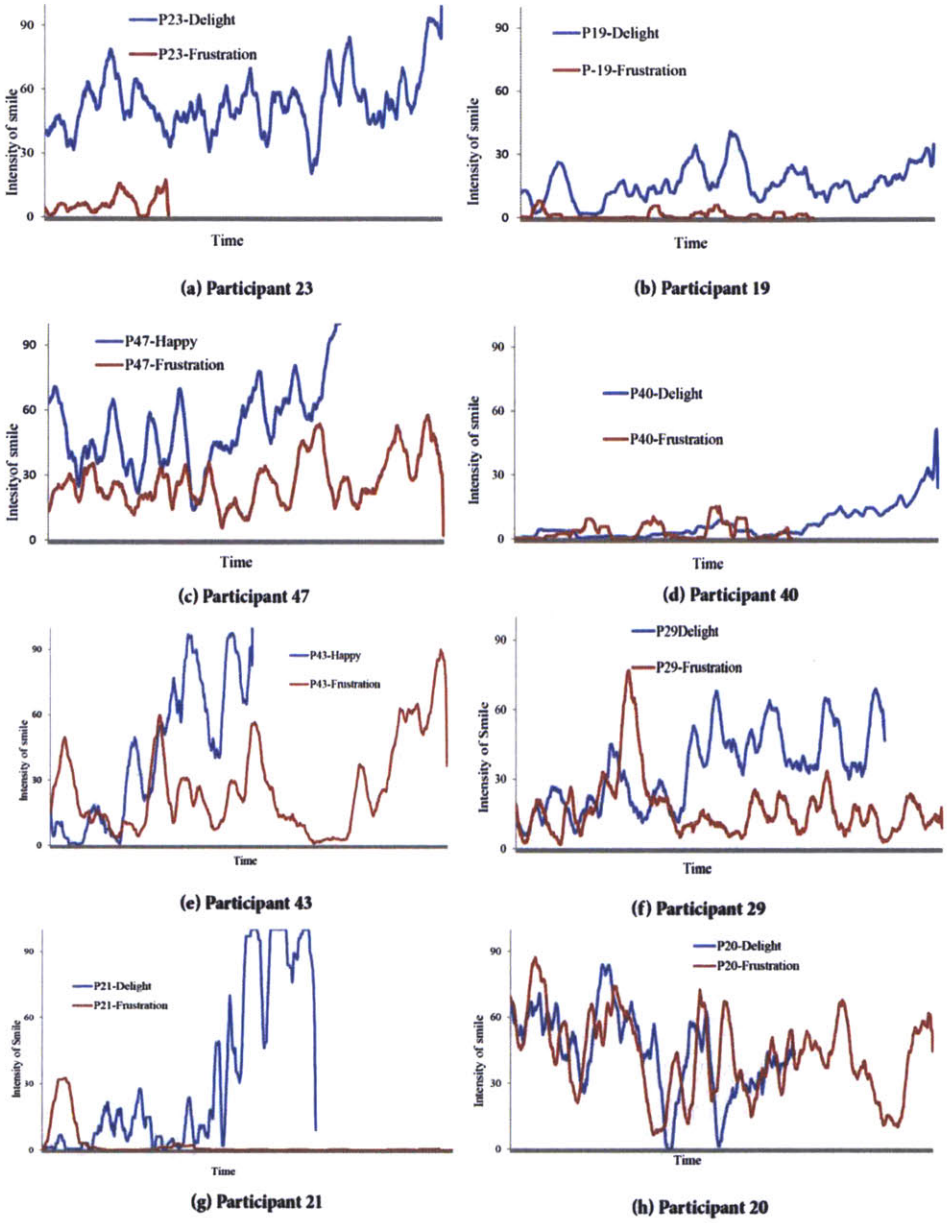


Figure 4-6: Graphs (a-h) of 8 participants whose patterns are representative of the rest of the participants. X axis is the time in seconds and y axis is the smile intensity/strength. Graphs (a, b, and c) are examples of participants who have distinct patterns of smile intensity when they are frustrated and delighted. Graphs (d, e, f, and g) provide examples of how the state of delight builds up in terms of smile intensity through time. Graph f, g are examples of participants who initiated their frustration with a social smile. Graph (h) is an example of one person who exhibited similar smile patterns regardless of whether delighted or frustrated.

and natural (Martin, 1997), it is possible that the participants under frustrating situations were trying to communicate their aggravation and acceptance of the situation, and trying to communicate that they were being put upon - to an imaginary interactant. This explanation does not come as a surprise since Fridlund (Fridlund, 1991) demonstrated that people who watched a pleasant videotape with friends smiled the same amount as people who watched a video with the belief that their friends were also watching the same in another room. In other words, it is possible for people to experience relevant social context even if they are alone, and enable it to guide the interaction patterns.

Is it possible for smiles under delighted and frustrated stimuli to have different temporal patterns? Messinger et al. (Messinger, Fogel, & Dickson, 1999) demonstrate that in the context of face-to-face interactions between adults and infants, contrasted types of smiles (e.g., Duchenne and non-Duchenne) can happen one after another in similar situations. But they usually occur in different temporal phases of a continuous emotional process. All these previous findings further strengthened the observation that it might be useful to analyze the temporal patterns of smiles as they occurred under delighted and frustrated stimuli as opposed to equating the presence of smiles with delight and the absence of smiles to be expected with a state of frustration.

4.2.4 Experiment 3: Temporal Models to Classify Smiles of Delight and Frustration

In the previous section, I demonstrated that smiles elicited under frustrated and delighted stimuli may look very similar as a snap shot, but their temporal patterns are usually different. Therefore, in this section, I propose a new algorithm to model the temporal patterns of frustrated and delighted smiles by analyzing their smile intensity. To test and validate the algorithm, I use the naturally elicited "Interaction" dataset from Experiment 2 (described in the previous section) with the goal of automatically separating them into appropriate classes.

Dataset

The Interaction dataset contained instances of smiles under frustrated and delighted stimuli, as the participants were either filling out the forms or they were watching the YouTube video. Since the participants were asked to hold their natural posture to elicit natural interaction during the experiment, in the post data-analysis stage, I noticed a lot of participants moved out of the camera frame as a result of natural movement. This resulted in 14 sequences of smiles under delighted stimuli, and 20 sequences of smiles under frustrated stimuli. The examples of smiles under frustrated stimuli were 7.45 seconds (*SD*: 3.64) and smiles under delighted stimuli were around 13.84 seconds (*SD*: 9.94) long on average.

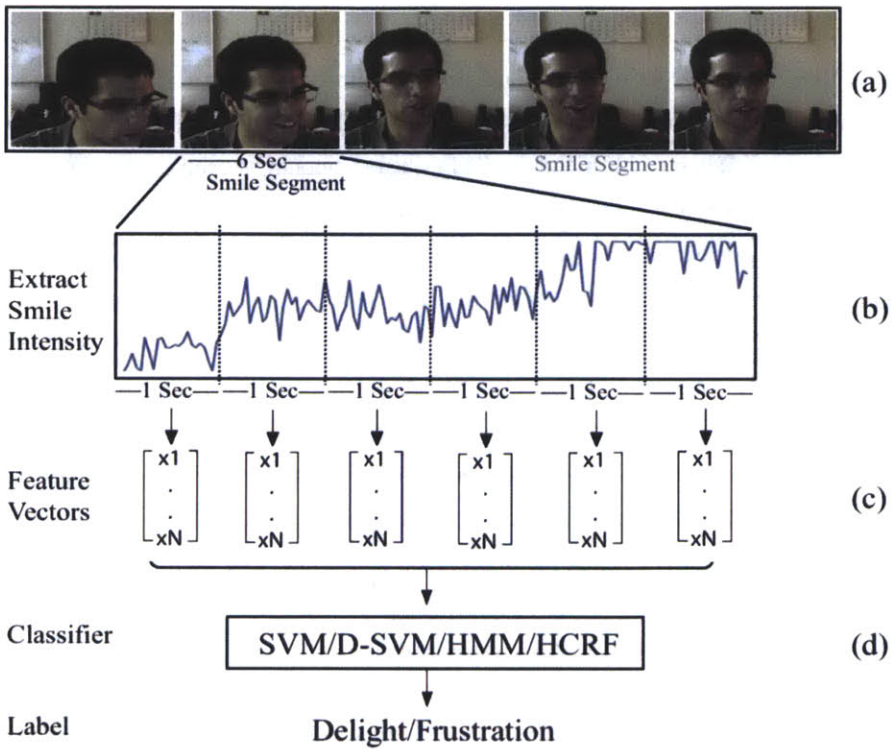


Figure 4-7: Methodology for smile classification. a) Segment smile sequence from clip, b) extract smile intensity from frames of smile segment, c) Form feature vectors from 1-second segments of smile intensity, d) classify input vector using SVM, HMM, or HCRF.

Algorithm

The system diagram of the algorithm to distinguish between the smiles under frustrated and delighted stimuli is provided in Figure 4-7. Figure 4-7 (a) refers to the video stream which was segmented into smaller sequences based on the rule described in Figure 4-8.

if (movement of the lip entails a smile)
mark the beginning of clip
***else if** (lips retract to a neutral position)*
mark the end of a clip

Figure 4-8: Logic of clip extraction from a larger file

Each sequence is then run through a feature extraction algorithm to determine the intensity of the participant smiling in each frame. The resultant graph per sequence looks like

Figure 4-7 (b), where the x-axis represents time and the y-axis represents the intensity of the smile. I split each sequence into smaller segments (30 frames, 1 second) and extract

local features per segment, as shown in Figure 4-7 (c). The feature sets are then classified to distinguish between the smiles under frustrated and delighted stimuli.

Features Extraction

As mentioned in the previous section, each smile sequence gets broken into smaller segments. I measure global peak and the global gradient across the entire segment. From the smaller segments, I only extract local mean and local peak, as shown in Figure 4-9. Given all the extracted local and global features, I infer the following 4 features that compare each local segment with the entire segment.

1) Percentage of local frames above global mean.
2) Local mean: *mean value within the segment*
3) Gradient across the segment: *change in smile intensity per frame along with the x axis*
4) Peak comparison = $\frac{LocalPeak}{GlobalPeak}$

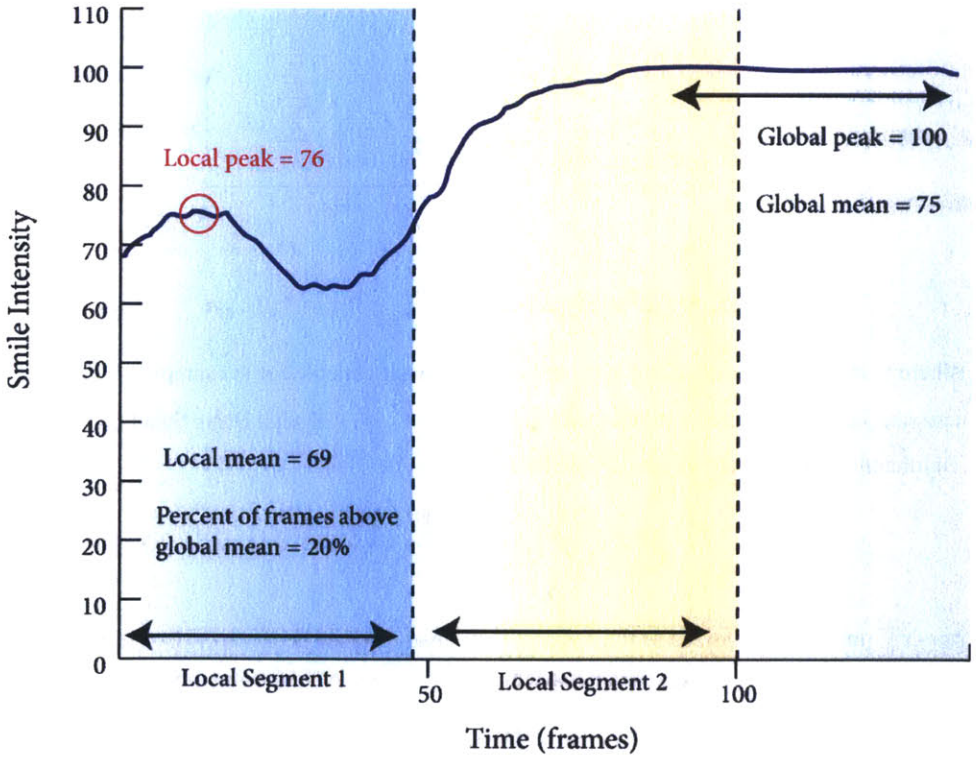


Figure 4-9: Description of the local and global features

Classification

The feature vectors and labels were used to train, validate, and test four models — Support Vector Machine (SVM), D-SVM (variation of SVM), Hidden Markov Models (HMM), Hidden Conditional Random Fields (HCRF). These experiments were carried out in order to evaluate the performance of classifiers with different dependence assumptions and to compare the performance of static vs. dynamic and generative vs. discriminative classifiers. The SVMs were implemented using LIBSVM (Chang & Lin, 2001). The HMMs were implemented using the HMM toolbox for MATLAB (Murphy, 2001). The HCRF classifiers were implemented using the HCRF toolbox (Quattoni, Wang, Morency, Collins, & Darell, 2007).

Support Vector Machine: A Support Vector Machine (SVM) classifier (Cortes & Vapnik, 1995), a static discriminative approach to classification, was used as the first benchmark. SVM works to find the hyper plane w that maximizes the margin between two different data points (e.g., frustrated smiles, delighted smiles). The standard formula for SVM is as follows:

$$\min_{w} \underbrace{\frac{1}{2}\|w\|^2}_{regularization} + \underbrace{\frac{C}{n}\left(\sum_{i\in\{y=+1\}}^{n_+}\varepsilon_i + \sum_{j\in\{y=-1\}}^{n_-}\varepsilon_j\right)}_{loss\ function},$$

$$s.t.\ \ y_i(w^T x_i) \geq 1-\varepsilon_i \ and\ \varepsilon_i \geq 0, i = 1,2,\ldots n$$

Where C is the misclassification cost, and ε_i is the slack variable for the sample x_i. For any new sample $\overline{x}$, prediction is performed through $\overline{y} = w^T \overline{x}$. A Radial Basis Function (RBF) (Buhmann, 2001) kernel was used. The RBF kernel is defined as:

$$K(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$$

$\|x-x'\|_2^2$ may be recognized as the squared Euclidean distance between the two feature vectors. The value σ is a free parameter which could be simplified using the following: $\gamma = -\frac{1}{2\sigma^2}$.

$$K(x,x') = \exp\left(\gamma\|x-x'\|_2^2\right)$$

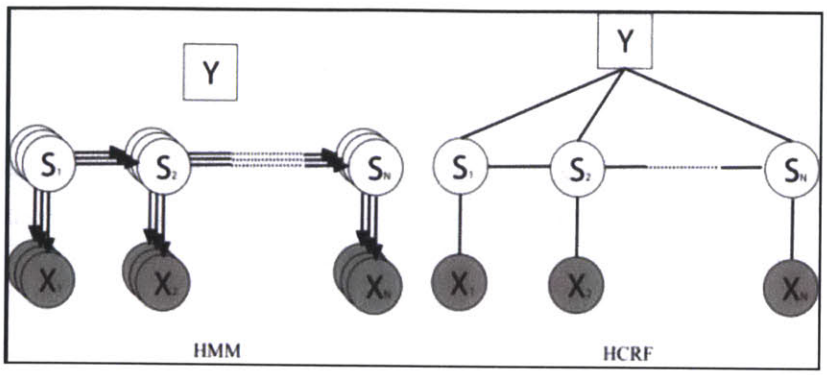


Figure 4-10: Structure of models. Xj represents the jth observation, Sj the jth hidden state and Y is the class label. The HMM requires a chain to be trained for each class label.

During the validation the penalty parameter, C, and the RBF kernel parameter, γ, were each varied from 10^k with k = -3,...,3.

For the SVM, all the local features for the 1 second long segments were averaged over the entire sequence to form a 4-d vector. These inputs were used to train an SVM. The D-SVM was a pseudo-dynamic model in which the time samples were appended. As the video samples were of varying lengths, zeros were appended to the end to form input vectors of equal length, 128 (32 seconds *4 features/second). After subtracting the mean from the data matrix, Principal Component Analysis (PCA) (I T, 2002) was used to reduce the dimensions. The four largest principal components were used to build the model. This process was repeated for all iterations of the validation and training scheme.

Hidden Markov Model (HMM) - HMMs are one of the most commonly used methods in modeling temporal data. An HMM is a first order Markov model, where the state sequence is not directly observable. Its observations are dependent on the state of the model. An HMM is specified by its first order Markov State transition probability, initial state probabilities and observation/emission distribution. For more information on the HMM, and how to learn the parameters of an HMM from observations, please see (Rabiner, 1990). I trained one HMM each for the delight and frustration classes. This is a dynamic generative approach to modeling the data.

In testing, the class label associated with the highest likelihood HMM was assigned to the final frame of the sequence. During the validation the number of hidden states (2,...,5) was varied, with two states being the most frequently chosen as performing the best.

Hidden Conditional Random Fields (HCRF) - In contrast to HMMs, Conditional Random Fields (CRFs) and CRF variants are discriminative approaches to modeling temporal data. The CRF model removes the independence assumption made in using HMMs and also avoids

the label-biasing problem of Maximum Entropy Markov Models (MEMMs) (Lafferty, McCallum, & Pereira, 2001). The dynamics of smiles are significant in distinguishing between them (Maja Pantic, 2009); as such, I hypothesized a potential benefit in removing the assumption that current features are solely dependent on the current valence label. During validation, the regularization factor (10^k with k = -3,...3) and number of hidden states (0,...,5) were varied, with a regularization factor of 10 and two states being the most frequently chosen as performing best.

Human Performance: In addition, I also had 10 individuals label all the clips elicited under frustrated and delighted stimuli. Users watched the clips and were asked to force a label: frustration or delight.

Results

In this section, I present the performance of a static model (SVM), a pseudo-dynamic version of SVM (D-SVM), and two dynamic models (HMM, HCRF). I had 34 samples in the dataset. First, one sample was removed from the dataset and held out as the test sample. Leave-one-out K-fold cross validation (K=33, training the model on 32 samples, testing its parameters on the 33rd, and repeating leaving a different one out each time) was performed to find the optimum parameters. The best of these was tested on the test sample. This was repeated for all samples in the dataset (34), providing 34 test results for each model. The HMM models required no more than 30 iterations during training. The HCRF needed no more than 300 iterations in training. Table 4-3 provides a comparison of the performance of the models.

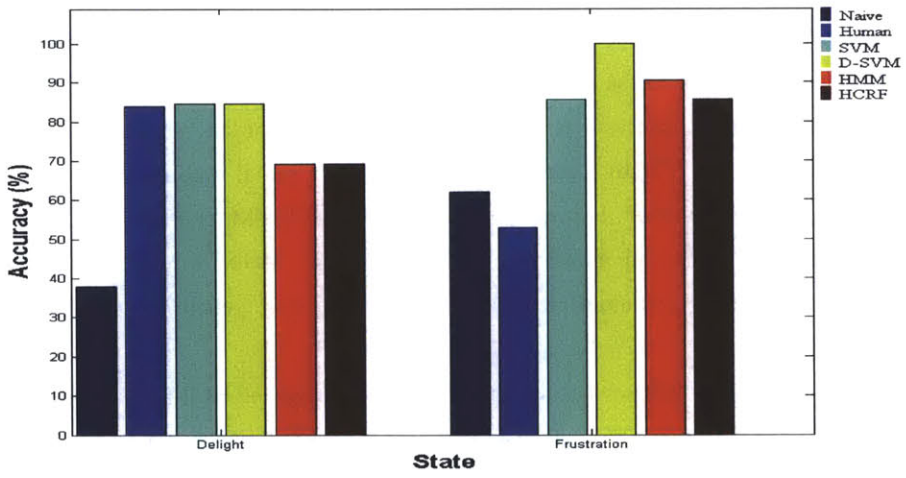


Figure 4-11: Bar chart comparing the performance of the human and computer labeling of 34 delighted and frustrated smile sequences.

Table 4-3: Performance statistics for SVM, D-SVM, HMM, HCRF towards binary classification

Model	SVM	D-SVM	HMM	HCRF	Human
Accuracy(%)	**85.30**	**92.30**	**82.40**	**79.40**	**68.98**
Sensitivity	0.89	0.91	0.83	0.82	0.81
Specificity	0.69	1.00	0.81	0.75	0.51
F-Score	0.82	0.92	0.75	0.72	0.68

In describing the following measures, I consider delight to be the positive class and frustration to be the negative class. Sensitivity measures the proportion of actual positives that are correctly identified, and specificity measures the proportion of negatives that are correctly identified.

In order to compare the machine performance with human performance, I asked 10 individuals, who were not part of this experiment and were oblivious of the experiment objective, to label the 34 video clips without using sound, "for frustrated smiles and for delighted smiles." The labelers were instructed to watch each clip and predict whether the participant in the clip was in a happy or in a frustrated state of mind. During the labeling process, the average accuracy among 10 labelers towards labeling the delighted smiles was 84% (chance was 14 out of 34 or 41%), and the accuracy for the frustrated smiles was 54% (chance was 20 out of 34 or 59%), with an overall accuracy across both categories of 69%.A detailed performance comparison between humans and the classifiers to recognize smiles under frustrated and delighted stimuli is provided in Figure 4-11. Figure 4-12 demonstrates visual sequences of smiles under delighted stimuli (Figure 4-12 A [I-III]) and frustrated stimuli (Figure 4-12 B [I-III]). Careful observation does reveal the fact there is a stronger smile signature in the frustrated smile compared to the delighted smile, which may explain why most people got it wrong. However, all of the classifiers (except for HCRF for the instance of delight) were able to classify the instances, shown in Figure 4-12, correctly. This demonstrates that the algorithm not only properly utilizes the signatures of smile (e.g., lip corner pull, cheek raiser etc.), but also the dynamics of the pattern in which they appear in time.

Discussion on Smiles Elicited under Frustration and Delight

I demonstrate in this section that it is useful to explore how the patterns of a smile evolve through time, even over many seconds (smiles under frustrated stimuli averaged 7.5 sec. and smiles under delighted stimuli averaged 13.8 sec.). The average smile intensity per clip under delighted stimuli was 76.26% (*SD*: 17.8) and for frustrated stimuli, it was 47.38% (*SD*: 28.9). While a smile of similar intensity may occur in positive and in negative situations, its dynamic patterns may help to disambiguate the underlying state.

Figure 4-12. A (I-III) sequences of images while a user is subjected to a delightful stimuli. B (I-III) sequences of images while a user is subjected to a frustrating stimuli. Only 5 out of 10 of the human labelers were able to label the video sequence containing images A (I-III) as a delighted smile, and only 1 out of 10 of the human labelers was able to label the video sequence containing images B (I-III) as a frustrated smile. However, all of the classifiers (except for HCRF for the instance of delight) were able to classify the instances.

Smiles are not only a universal, but also a multi-faceted expression. We smile to express rapport, polite disagreement, delight, favor, sarcasm and empathy. Being able to automatically recognize and differentiate the different types of smiles could fundamentally change the way we interact with machines today. Moreover, it is very important that a machine discern the difference between a frustrated customer and a delighted one and not just assume that a smile means the customer is happy.

Analysis on the feedback datasets collected from Experiment 1 and Experiment 2 revealed that in the acted data, close to 90% of the participants did not smile when they were frustrated. On the contrary, in the naturally elicited Interaction dataset of Experiment 2, close to 90% of the participants did smile when they were frustrated. I was surprised to see a lot of participants smile despite self-reporting to be frustrated. This further motivated to develop algorithms to distinguish between the spontaneous naturalistic examples of smiles under delighted and frustrated stimuli. To do this, I have automated the process of extracting temporal facial features in real-time that are believed to be correlates of smiles. Among the 4 classifiers, the most robust classification was achieved using D-SVM with an accuracy of

92% and F-score 0.92. It is a little surprising that D-SVM outperformed HMM and HCRF for the dataset, especially when HMM and HCRF have been shown to perform well modeling temporal data. However, with the addition of more classes and training samples, the best model might change. All the classification models that I have used in this thesis could be implemented as part of a real-time system. Also, it is worth noting that given the limited set of smiling instances, I used leave-one-out method as opposed to k-one-out, where k>1. Leave-one-out methods could provide optimistic results on unseen data. However, with the availability of more data, the system could scale to recognize a wide variety of smiles.

In the dataset, the gradient across the entire smiling instance was the most important feature towards distinguishing between delighted smiles. While this is an important finding, it needs to be further validated across larger dataset and individuals.

How good are we at differentiating the patterns of delighted smiles and frustrated smiles if we can only look visually at videos of facial expressions? The results, as plotted in Figure 4-11, show that the human ability to identify spontaneous frustrated smiles by looking at the facial cues is below chance, whereas we perform comparatively better in identifying the spontaneous delightful smiles. Therefore, one may question if we can build systems that perform better than the human counterpart in disambiguating between naturally occurring smiles under delighted and frustrated stimuli by only analyzing facial expressions. Results demonstrate that the automated system offers comparable or stronger performance in recognizing spontaneous delighted smiles. However, the system performs significantly better by correctly labeling all the spontaneous smiles under frustrated stimuli compared to the below-chance human performance.

It is interesting to note that even though it is possible for people to smile under frustration, we usually have a pre-defined mindset of not associating smiles with frustration. This mindset was reflected in the study through the human's inability to label the frustrated smiles correctly, as well as the human posers who posed frustration without smiles. Presumably, many of them would have smiled if they had actually been frustrated instead of just being asked to recall being frustrated.

One would wonder, and rightly so, why would a machine perform better than the humans in recognizing instances of spontaneous frustrated smiles? One possible explanation is that humans usually rely on additional information that they sense using other modalities (e.g., prosody, spoken words, context) to disambiguate among different kind of smiles. Unavailability of such information could reduce a person's ability to understand emotions. Machines, however, could utilize the local intrinsic structures of the temporal patterns in the context of the entire sequence discovering unique patterns that are typically not seen by humans. Another possible explanation is that we have used a skewed number of samples (62% instances of frustrated smiles, and 38% instances of delighted smile) in the training

process. Therefore, the classifier is more likely to do better in categories where it has seen more examples. However, humans have seen examples of these smiles throughout their life in everyday interactions, so this does not explain why they are not better still.

4.2.5 Experiment 4: Understanding the Morphology of Polite and Amused Smiles

In this section, I expand the research of smiles by looking at polite and amused smiles. Motivated by the previous work (Cohn & Schmidt, 2004)(Hoque & Picard, 2011)(Krumhuber, Manstead, & Kappas, 2007)(Valstar, Gunes, & Pantic, 2007) on disambiguating different kinds of smiles (e.g., deliberate vs. genuine), I focus more on understanding the morphological and temporal patterns of polite and amused smiles in the context of face-to-face interactions.

Even with one category of smile, there are ways to vary the dynamic and morphological properties of the smile to indicate the scale and sincerity of that expression. How are the properties of smiles different when they are shared vs. solo? Previously, these kinds of questions have been very difficult to answer, especially since it is not trivial to collect large sets of labeled spontaneous expression data from quality-recorded natural conversational interactions.

In the past, Ambadar et al. (Ambadar, Cohn, & Reed, 2009) investigated morphological and dynamic properties of deliberate and genuine smiles. They collected data on a study where participants were brought to the lab to act various facial expressions. Between acting and data collection, the participants voluntarily looked at the experimenter and smiled. Those examples were then tagged by judges and were used to analyze the properties of deliberate and genuine smiles. In another study (Ochs, Niewiadomski, & Pelachaud, 2010), Ochs et al. investigated the morphological and dynamic characteristics of amused, polite, and embarrassed smiles displayed by a virtual agent. A web platform was developed for users to provide smile descriptions (amused, polite and embarrassed) for a virtual agent. While these studies have been extremely useful to motivate the problem with initial exploratory results, none of them really address the issues of understanding those smiles in contextual face-to-face interactions when those smiles are shared and not shared.

In this next study, I utilize a dataset collected by Kim et al. at MIT (K. Kim, Eckhardt, Bugg, & Picard, 2009), which contains spontaneous face-to face interactions in a banking environment, where smiles were labeled by both participants after the interaction. While the dataset is labeled for various expressions, for this study, I focus on understanding the differences between polite and amused smiles and how these smiles change when the smiles are shared or occur from just one participant. In particular, I focus more on understanding the

difference in durations, occurrences, and dynamic properties of polite and amused smiles. The remaining sections describe the dataset and experimental set up, research questions, findings and discussions.

Banker Dataset

This section describes how the data was collected and is largely an excerpt from the work of Kim (Kim et al., 2009). In Kim et al.'s MIT study, young adults interested in learning about banking services were invited to meet with a professional banker in a conference room (Figure 1). The bankers provided information about two kinds of financial services just as they did at the retail branches where they worked during the day. The first service was to cash a $5 voucher from the participant as compensation for participating in the study. This part was designed to simulate a cashing a check scenario. The participants were recruited in the study with the incentive of getting $10 for their participation. However, after each arrived, the banker told him or her that they could only get $5 for now and would need to fill out additional paper work to claim their remaining $5. This manipulation was made to elicit a slightly negative state in the customer in order to mitigate the "it's fun to be an experiment" phenomenon and also to approximate the real-world situation where a customer often goes to a banker feeling a little negative because of a need to fix a problem. After the experiment ended the participant received the rest of the money without additional paperwork.

The second service was for the banker to explain one of four financial services that a customer chose to learn more about. This part was designed to allow the customer to ask questions and receive information about the financial product just as they would in a real bank visit.

Participants - Two professional personal bankers were hired, each with over two years of career experience as a personal banker, to do what they usually do at work - explain financial services. One banker interacted with seventeen participants, while the other interacted with forty-four. Each experiment included one banker with one customer.

Before hiring, the bankers were asked if they would be willing and able to manipulate the type of facial expressions displayed during interaction with the customers. Each banker agreed to alter his facial expressions in three different ways, following these exact instructions.

- Manipulation 1 – Neutral facial expressions: Sustain neutral facial expressions over the entire interaction.
- Manipulation 2 – Always smiling: Sustain smiling over the entire interaction.

- Manipulation 3 – Complementary facial expressions, i.e., empathetic: Understand the customer's feeling and respond to it appropriately by smiling when the customer seems to feel good.

Throughout the experiment, the bankers interacted with the customer normally in addition to maintaining one of the three manipulations. This included greeting the customer, providing proper information, and thanking the customer for their time. The facial expressions and the voices of the banker and of the customer were unobtrusively recorded using a video camera from the moment they met and greeted to the end when the customer left. Customers were not told about the banker's facial expression manipulations and all the interactions appeared to proceed very naturally.

Forty one males and twenty females (n=61) who were interested in receiving information about different financial services were recruited through flyers. Before the experiment started, they were told that their face and voice data would be recorded as banks normally do for security reasons. However, they were not told that their facial expressions would be analyzed until after the study. Afterward, they were told about the need to analyze expressions and they were asked to help label them.

Participants Experimental setup

The experiment was conducted in a room equipped with a desk, two chairs, bank service advertising pamphlets and two cameras to make the appearance alike to a personal banking service section at banks (Figure 4-13). One camera was used to record the banker's facial expressions and the other was used to record the participant's facial expressions.

Prior to the participant entering the room the banker was told which expression manipulation to conduct. The participant was then allowed into the experiment room where they would interact with the banker and learn about specific financial services. At the end of the experimental interaction, which took about 10 minutes on average, both the banker and participant filled out 9-point Likert scale surveys evaluating the quality of the service based on the most comprehensive and popular instrument SERVQUAL (Parasuraman, Zeithaml, & Berry, 1988) and the attitude of the banker.

Facial Expressions Coding

After the banker and participant finished the surveys, they were debriefed and asked to label the video data for their facial expressions. First they labeled their own video information, and then they labeled the videos containing the person they interacted with (e.g., banker coded customer & customer coded banker). Therefore, for each conversation, there are two videos containing the facial expressions of banker and customer.

Figure 4-13. Experimental set up for banker and the customer interaction. The camera that is visible behind the banker is capturing the facial expressions of the customer. There is another camera, not visible in the image, behind the customer, capturing the facial expressions of the banker.

Bankers and customers used custom labeling software to label their expressions and affective states. The label interface contained two parts: the upper part displayed the video and the lower part provided the entity for the banker and the participant to enter the time when a certain facial expression was observed and seven affective labels to select. These seven labels were: smile, concerned, caring, confused, upset, sorry, and neutral. If there was no proper label to choose from, the user could press "Other" and enter another label that they thought was appropriate for the expression. The labelers were instructed to stop playing the video and click on the label button when they saw a facial expression, and then to continue to play the video until they saw a change in the facial expression. On the right side of the user interface, there was a text box displaying the time and the labeling result and it was editable so that the user could annotate the reason for each facial expression, e.g. "smile – he made me laugh". The labelers were instructed to group every smile as either polite or amused. These extra labels were entered manually in the text box.

Research Questions

In this dissertation, I focus primarily on understanding the differences between the polite and amused smiles. I anticipate that polite smiles in the context of our dataset are more likely to be social, masking and controlled smiles while the amused smiles in the context of

our dataset are more likely to be genuine, and felt. In this study, I focus the attention towards exploring the differences between polite and amused smiles in face-to-face interactions.
I was primarily interested to explore three questions in this study. When people exhibit amused and polite smiles in task-driven spontaneous face-to-face interactions:

1. Are there any differences in terms of durations between polite and amused smiles?
2. Do amused and polite smiles get shared by the conversation partner? Do people share them equally or does one type get shared more often?
3. Are there any differences in dynamic features between polite and amused smiles? How can we quantify the difference of dynamic?

Experiments

To directly address the research questions, I performed a series of experiments. This section describes the experimental setup and results from the analysis.

Smile Annotations and Segmentation

In this study, I do not measure the dynamics of the bankers' smiles since they were manipulated by the experimental condition; I only analyze the dynamics of the customers' smiles and whether or not their smiles occurred in conjunction with a banker smile or solo.

As mentioned in the previous section, each customer video was labeled for polite and amused smiles by the banker and by the customer him/herself. I did not use the labels produced by the customers since, after looking at them and seeing huge variation in how the labels seemed to be applied, I realized different customers seemed to interpret the labels differently. Using the two bankers' labels led to significantly more consistent labels as judged by outward appearance of the expressions. I therefore chose to use the labels produced by the bankers, which are more likely to be consistent. Using a third party coder to code the smiling instances remains for future work. One significant advantage of using the banker's labels is that they are automatically taking conversational context into account when interpreting the smiles.

The labelers (bankers) indicated individual points in the video where they observed polite and amused smiles. Therefore, extra work was needed to be able to approximate the exact beginning and end points of each marked smile. Given the variability in the data, I manually annotated the beginning and end of each smile given the initial labels produced by the bankers. Through this process, I gathered 227 clips of amused smiles and 28 samples of polite smiles encompassing 61 participants playing the role of customers. I was also interested to find out which of those samples of smiling instances were also shared by the banker. Therefore, I separated the smiling instances of customers where the banker also self-labeled himself to be exhibiting the same kind of smiles (polite or amused).

Durations and Timing

The average duration of customers' shared polite, shared amused and unshared polite and unshared amused smiles are shown in Table 4-4.

Table 4-4. Comparison of durations for customers' polite and amused smiles. The labels are produced by the bankers.

	Average duration	standard deviation
shared amused smiles (n=44)	6.1 sec.	4.6
non-shared amused smiles (n=183)	4.7 sec.	3.0
non-shared polite smiles (n=21)	3.7 sec.	1.2
shared polite smiles (n=10)	3.2 sec.	0.77

It is evident that the amused smiles are usually longer when shared as opposed to unshared. Comparatively, the durations of polite smiles are usually the same regardless of whether the smile is shared or not. Thus, it is not simply the case that sharing happens because the smiles have longer duration. The high standard deviation for un/shared amused smiles also indicates that the distribution of durations for amused smile is pretty widespread, as shown in Table 4-4 .

It is evident that the amused smiles are usually longer when shared as opposed to unshared. Comparatively, the durations of polite smiles are usually the same regardless of whether the smile is shared or not. The high standard deviation for un/shared amused smiles also indicates that the distribution of durations for amused smile is pretty widespread, as shown in Table 4-4.

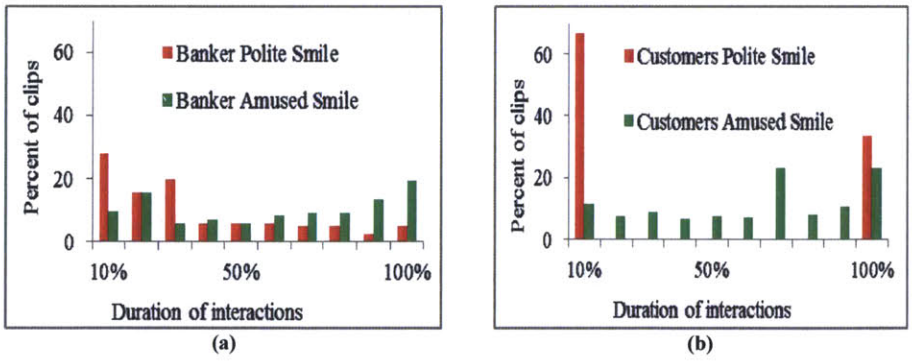


Figure 4-14. Position of polite and amused smiles relative to the entire conversation. (a) Bankers yielded polite and amused smiles consistently throughout the interaction. (b) Customers yielded polite smiles only at the beginning and end of conversations, and amused smiles throughout the interaction.

I have also investigated the positions in respect to the entire conversations for amused and polite smiles, for both bankers and customers, as shown in Figure 4-14. It is evident from Figure 4-14 that bankers seem to display polite and amused smiles throughout the interaction, whereas the customers seem to display polite smiles at the start and end of the conversations. On the other hand, about 1/3 of the 31 polite smiles were shared, while only about 1/5 of the 227 amused smiles were shared in these data.

Smile Dynamics

Along with duration and position parameters, I was also interested in exploring the dynamics of smile. I define three parameters to better analyze smile dynamics: rise, sustain and decay. Note that in our natural data, there was often not one clear "apex" or "peak" to the smile. Thus, I do not use the usual definition of onset time = "time to the highest peak", while, offset= "decay from that highest peak", because for spontaneous smiles, they often had a sustained region with multiple peaks, as in Figure 4-15. Therefore, in this study, I refer to onset as rise time, offset as decay, and apex as sustain.

Careful observations indicated that the time stamps produced by the bankers were mostly the beginning of peak (L) of the smile without any further information on the rise and decay time as well as its sustain period.

The manual labeling process thus provided us with the beginning of rise times (R) and end of decay times (D). A visual example of where points L, R and D are more likely to be located is shown in Figure 4-15.

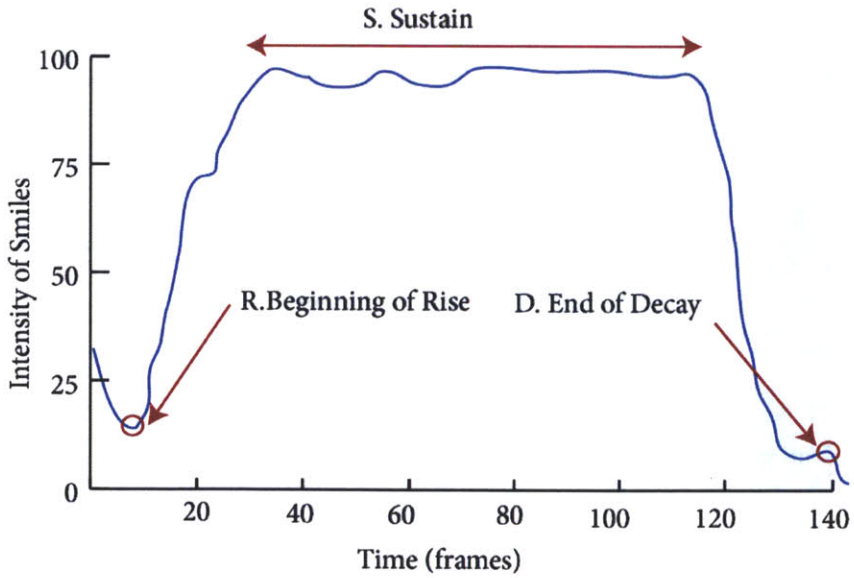


Figure 4-15. A visual example of where points such as R (beginning of rise), D (end of decay) and S (sustain) could be located given the time stamp label, L, given by the labeler.

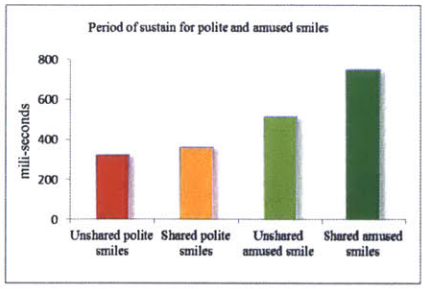


Figure 4-16. Comparison of the period called sustain for (un)shared polite/amused smiles. The period of sustain for instances of shared amused smiles is the highest.

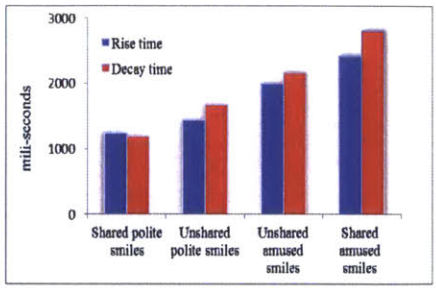


Figure 4-17. Comparison of rise, and decay time for (un)shared polite/amused smile instances. The ratio between rise time and decay time for all the categories seem very symmetrical.

The task was to automatically identify the region, S, which defined the time frame when the participants are more likely to be holding their smiles. I automated an algorithm to identify the locations where the probability of smiling is the highest. Then it traverses left and right looking for deviations that are higher than a pre-determined threshold to mark the start or end of the sustain period. For clips with multiple peaks spread over the signal, the algorithm is biased towards selecting an initial point that is closer to the point labeled by the labeler.

Figure 4-16 provides the comparison of sustain period among shared polite/amused smiles and unshared polite/amused smiles. In these data amused smiles have a longer sustain period than polite smiles. Additionally, shared amused smiles have longer duration for sustain compared to unshared amused smiles, whereas the duration of sustain for both shared polite and unshared polite is almost the same. This finding appears to be consistent with the popular notion that shared joy multiplies joy, here manifest by the extended duration of an amused smile.

In addition to sustain, I also analyzed the rise and decay times of amused and polite smiles, as shown in Figure 4-17. It is evident for both amused and polite smiles, regardless of whether they are shared or not, the difference between rise time and decay time is not statistically significant, and they are somewhat symmetric. Given this result, I decided to look more closely at the velocity for both rise and decay.

I analyzed the velocity of rise and decay signals for polite and amused smiles when they are shared vs. not shared. The velocity of rise (Vr) and decay (Vd) were defined as displacement in y axis divided by the elapsed time.

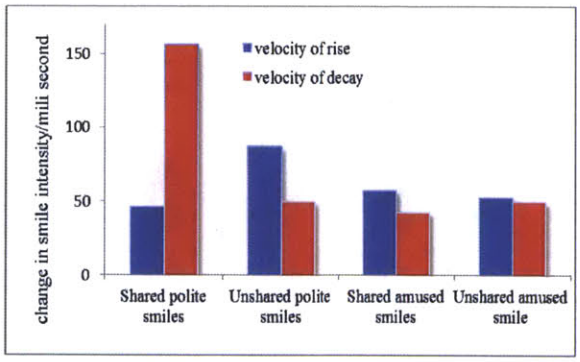


Figure 4-18. Comparison of shared polite/amused smiles with unshared polite/amused smiles in terms of velocities

$$\text{Vr} = \frac{Y_s - Y_r}{T_s - T_r} \text{ and Vd} = \frac{Y_d - Y_s}{T_d - T_s}$$

where Y_s, Y_r and Y_d represent the smile intensity at the middle of the sustain period, the beginning of rise and at the end of decay, respectively. T_s, T_r and T_d represent the time at the middle of sustain, at the beginning of rise and at the end of the decay, respectively.

As shown in Figure 4-18, the analysis suggests that the amused smiles have the most symmetric velocities of rise and decay, whether shared or unshared, $V_d \approx V_r$. However, for polite smiles, these velocities were more asymmetric. Shared polite smiles decayed the fastest: $V_r < V_d$ while the polite smiles that rose the fastest were unshared $V_r > V_d$ As shown in Figure 4-8, for shared and unshared polite smile instances, the ratio between T_s - T_r (time difference between sustain and rise) and T_d - T_s (time difference between sustain and decay) remains almost same. It is the smile intensity (Y) that is contributing to the difference in velocities between shared polite and unshared polite instances.

Discussion of Experiment 4

In experiment 4, I have investigated the phenomenon of polite and amused smiles in a corpus of Banker face-to-face interactions. The results suggested that duration of amused smiles are higher than the duration of polite smiles, which is consistent with what has been reported in the literature so far, although under different data gathering conditions. I additionally report that the duration of amused smiles are more likely to be higher when they are shared as opposed to solo. However, for polite smiles, the duration does not seem to change much regardless of whether the smile is shared or not (in fact, slightly higher duration when not shared).

In this face-to-face banking dataset, I notice that when bankers labeled their polite and amused smiles during all the interactions, they seem to indicate that they have displayed polite and amused smiles consistently during the entire interaction, as shown in Figure 4-14

(a). However, when the same banker labeled the corresponding customer video, they indicated the occurrences of polite smiles only in the beginning and end of the interactions, as shown in Figure 4-14 (b). In other words, customers were viewed as less likely to share polite smiles with the bankers unless it happened at the beginning or end of the interactions. For amused smiles, the customers were more likely to share the smiles with the banker during the entire interaction. These data support a view that it is socially acceptable not to share a polite smile when it occurs in the middle of the discussion.

One of the key findings is that whether shared or not, amused smiles are more likely to be symmetrical in rise and decay velocities. I additionally report that the duration of amused smiles are more likely to be higher when they are shared as opposed to solo. However, for polite smiles, the duration does not seem to change much regardless of whether the smile is shared or not.

Hopefully, the reported findings will further motivate the development of automated systems that can differentiate between polite and amused smiles under natural conditions.

4.3 HEAD NOD AND SHAKE DETECTION

Detecting natural head nods and shakes in real-time is challenging, because head movements can be subtle, small, or asymmetric. My implementation tracked the "between eyes" region, as described by Kawato and Ohya (Kawato & Ohya, 2000). The head-shaking detection algorithm is described below.

$$i = \begin{cases} stable & if \sum_{n=-2}^{2} \max(x_{i+n}) \le 2 \\ extreme & if\ x_i = \sum_{n=-2}^{2} \max(x_{i+n})\ or\ x_i = \sum_{n=-2}^{2} \min(x_{i+n}) \\ transient & otherwise \end{cases}$$

Where, *i=current frame*

ALGORITHM. HEAD NOD AND SHAKE

If *max* (X_{i+n}) – *min* $(X_{i+n}) \le 2$ $(n = -2 \ldots + 2)$
then *frame i is in stable state*

If X_i = *max* (X_{i+n}) $(n = -2 \ldots + 2)$
or X_i = *min* (X_{i+n}) $(n = -2 \ldots + 2)$
then *frame i is in extreme state*

else
frame i is in transient state

Head nods are also detected using this algorithm, except that only movements in the y-axis are considered. The algorithm assigns one of three states to each frame: 1) stable, 2) extreme, and 3) transient.

At every frame, the system checks whether the state has changed from stable to extreme or transient in order to trigger the head shake evaluation process. Therefore, the system has a two-frame delay. If there are more than two extreme states between the current stable state and the previous stable state, and all the adjacent extreme states differ by two pixels in the x-coordinate, then the system records the movement as a head shake. When a user looks to the left or right, the system logs these as stable states instead of extreme states between or after the transient states, and therefore, these head turns do not get mislabeled as head shakes.

In order to evaluate the head nod and shake detection algorithm, four videos from the study reported in Chapter 3 (section: 3.2) were randomly selected. The average duration per video was 370 seconds and the total number of frames in all the videos was 44,460 (~24 minutes). The head nod and shake detection algorithm was run through all the frames to detect the regions where participants either nodded or shook their head. In order to validate the output of the algorithm, a human annotator also watched the all the videos, and identified the regions where head nods or shakes occurred. Given that it might be difficult for humans and algorithm to agree on frame level accuracy, I used a buffer window of 30 frames. In other words, if the regions identified by human annotator and the algorithm were within 30 frames (1 second), then I considered it as an agreement. The precision, recall and F-score for head nod and shake are reported in Table 4-5.

Table 4-5. Precision, recall and F-score of head nod and shake algorithm

	Precision	*Recall*	*F-score*
Head nod	0.75	0.88	0.81
Head shake	0.74	0.96	0.84

4.4 PROSODY ANALYSIS

This section provides details on the real-time framework that was developed to recognize prosodic properties of speech. Specifically, the algorithms to track pitch, loudness, pauses, speaking rate and weak language (e.g., fillers) are described below.

4.4.1 Intonation

To measure pitch, I use the autocorrelation method towards short term analysis of fundamental frequency proposed by Boersma (Boersma, 1993). Boersma argued that the position of the maximum of the autocorrelation function of the sound carries information about the acoustic pitch period, while the relative height of the maximum contains information on the degree of periodicity (harmonics-to-noise ratio) of the sound. However, sampling and windowing introduce errors in accurately measuring the position and height of the maximum. Therefore, the successful approach relies on the exclusive use of frequency-domain techniques for the determination of the harmonics-to-noise ratio.

For a stationary time signal $x(t)$, the autocorrelation $r_x(\tau)$ as a function of the lag τ is defined as:

$$r_x(\tau) \equiv \int x(t)x(t+\tau)dt \qquad \text{(i)}$$

When $\tau = 0$, the function has a global maximum. If there are also global maxima outside 0, the signal is called periodic and there exists a lag T_0, called the period, so that all these maxima are placed at the lags nT_0, for every integer n, with $r_x(nT_0) = r_x(0)$. The fundamental frequency F_0 of this periodic signal is defined as $F_0 = 1/T_0$. If there are no global maxima outside 0, there can still be local maxima. The signal is said have a periodic chart and its harmonic strength R_0 is a number between 0 and 1, equal to the local maximum $r'_x(\tau_{max})$ of the normalized autocorrelation, if it satisfies the following two conditions: 1) If the highest of the local maxima is at a lag $\tau_{max} = T_0$ and, 2) if its height $r_x(\tau \text{ max})$ is large enough.

$$r'_x(\tau) \equiv \frac{r_x(\tau)}{r_x(0)} \qquad \text{(ii)}$$

For dynamically changing signals, which is of interest in this thesis, the *short-term* autocorrelation at a time t is estimated from a short windowed segment of the signal centered around t. This gives estimates $F_0(t)$ for the local fundamental frequency and $R_0(t)$ for the local harmonic strength. When performing short-term analysis on a non-stationary signal, the estimates should be as close as possible to the quantities derived from equation (1), for them to be meaningful. A detailed step-by-step algorithmic description of how this is achieved is provided by Boersma in(Boersma, 1993). This particular formulation of F_0 is closer to perceived pitch than to mathematical periodicity. An implementation of this algorithm is available through *Praat* – an open source speech processing software (Boersma & Weenink,

n.d.). *Praat*'s pitch analysis settings do the following five things that bring the analysis towards something that is close to how we perceive intonation in language.

1. The "octave cost" setting makes sure that if even-numbered periods (of duration T_0) have slightly different shapes from odd-numbered periods (of duration T_0 as well), then the pitch will be $1/T_0$ rather than the mathematically expected $2/T_0$, which humans do not hear.

2. The "octave-jump cost" setting makes sure that upward or downward octave jumps are not analyzed. Octave jumps do often occur mathematically, but tend not to be heard by humans if reversed within a short time.

3. The "voiced-unvoiced cost" setting makes sure that stretches tend to be analyzed as voiced or unvoiced rather than alternating in every 10 milliseconds;

4. The "silence threshold" makes sure that very quiet background noises are not incorporated into the pitch analysis.

5. The "voicing threshold" setting makes sure that a human-like criterion is set for what parts of the signal should be considered voiced or voiced non-human sounds.

4.4.2 Speech Recognition and Forced Alignment

For real-time speech recognition, I used the Nuance speech recognition software development kit (SDK)("Nuance Communications," n.d.). While the speech recognition system captures the entire transcription of the interaction, it does not perform any natural language understanding, i.e., there is no assessment of the semantics of speech, as the current application context focuses on nonverbal training.

The system was designed with the aim to perform word level prosody analysis, since this will enable obtaining information on prosodic properties per word (e.g., which words are being more emphasized or stressed). There are two main challenges to this: 1) Real-time recognition of text from the recorded audio file; 2) once the text is extracted, information on the boundary for each word is needed to accomplish word level prosody analysis. To solve these issues, the Nuance speech recognition Software Development Kit (SDK) and a forced aligner (Yuan & Liberman, 2008) were used. I used the Nuance SDK to first recognize the text, and then the forced aligner to recognize the beginning and end of each word. An example of how a raw speech signal gets mapped into words and its prosodic properties is shown in Figure 4-19.

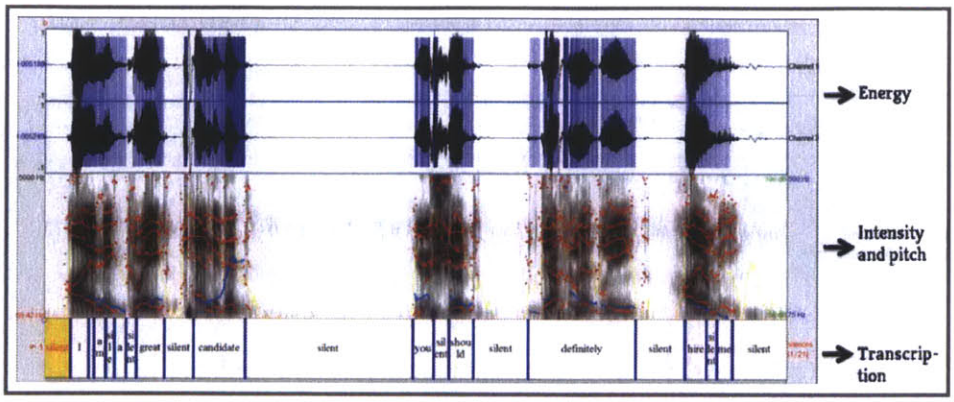


Figure 4-19. The text and the speech signal are aligned using forced alignment to perform word level prosody analysis.

4.4.3 Speaking Rate

In this thesis, I focused on articulation rate, which is defined as syllables per second in segments of speech where there are no pauses (De Jong & Wempe, 2009).

The peak of a syllable is most commonly known as the vowel in the syllable. It can be assumed that a vowel within a syllable has higher energy compared to its surrounding sounds. Therefore, using a threshold for intensity, I locate the peaks corresponding to vowels only. In order to discard multiple peaks in one syllable, I use an intensity contour to confirm that the intensity between the current peak and the preceding peak is very low. Finally, I eliminate surrounding voiceless consonants that have high intensity by excluding unvoiced peaks. Once the syllables are located, one can compute the number of syllables per second, or speech rate. A visual example of how speech rate is calculated is shown in Figure 4-20. The speaking rate algorithm was adapted from Jong et al. (De Jong & Wempe, 2009), and further details on evaluation of the speaking rate algorithm is available in their paper.

4.4.4 Weak Language

I would like to be able to automatically spot the filler words (e.g., "like," "basically," "umm," "totally," etc.) in conversations and give users feedback on its use and frequencies. I worked with the Blue Planet Public Speaking Company ("Blue Planet Public Speaking," n.d.) to define the weak language dictionary (provided in Appendix A: Weak Language Dictionary). Given the transcription of what the user has said from the speech recognizer, this module performs a "string matching" algorithm to spot the weak words. At the end, it calculates the percentage of weak words as [(weak words/total number of words) * 100] as well as what they are and how many times they have occurred.

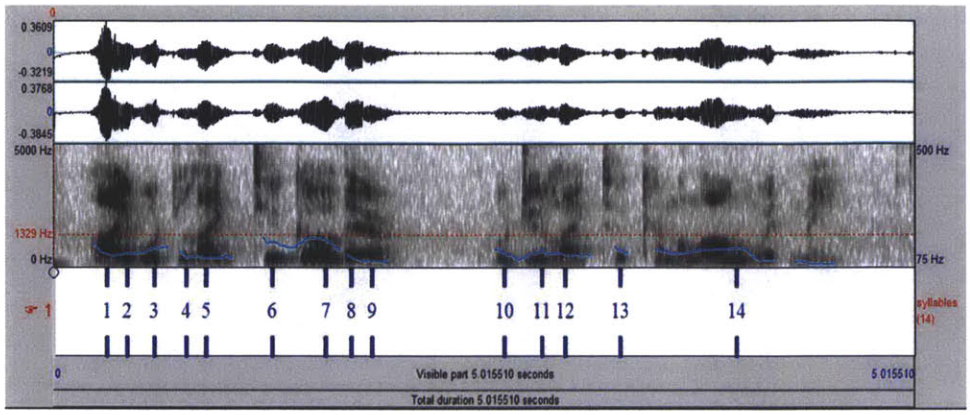


Figure 4-20. Computation of speech rate from raw speech. Top: Raw speech. Middle: Pitch contour. Bottom: Syllables (depicted as blue lines) located using the peaks. Speech rate is computed as syllables per second.

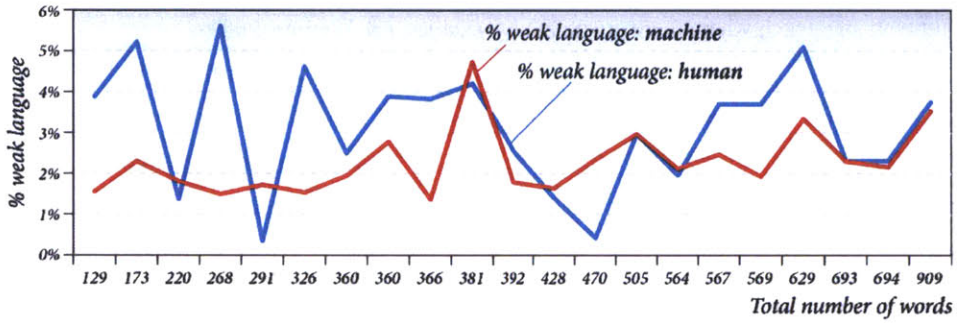


Figure 4-21. Comparison of system's automated detection of weak language with human performance across 21 audio files.

The weak language detection algorithm was validated using the job interview dataset collected from (Hoque, Courgeon, Martin, Mutlu, & Picard, 2013). The corpus contains speech data of 90 college students being interviewed. I randomly took 21 audio files from the corpus and applied the algorithm to measure the percentage of weak language, what words they were, and their frequencies. I had a human annotator listen to the 21 audio samples and manually annotate the weak language that occurred during the interview sessions. Considering the human label as ground-truth, I measured the precision, recall, and F-measure of the automated weak language module to be .91, .53, and .65, respectively. Most of the mistakes in the algorithm were found to stem from the errors of the speech recognition engine. Figure 4-21 provides a visual illustration comparing the human performance in spotting the weak language with the machine performance. The algorithm performs nearly as well as humans as the number of words increases.

4.5 CHAPTER SUMMARY

This chapter is an exploration of technologies to understand and recognize nonverbal behaviors that are deemed relevant in job interview scenarios, as well as in many other kinds of face-to-face interactions. For example, smiling during interviews to be perceived as social, using head gestures to appear confident, pausing between sentences to build suspense, enunciating each word and modulating volume to increase clarity in speech, and minimizing the use of fillers to appear more knowledgeable, are all examples of nonverbal skills.

The first exploration started with the effort to analyze smiles. But a smile is a multi-purpose expression with many different variations. To further understand the phenomenon, as part of an exploration, I conducted a new study to collect acted (Experiment 1) and elicited (Experiment 2) expressions of frustration and delight, and ran analysis with multiple methods to automatically classify them. I found two significant differences in the nature of the acted vs. natural occurrences of expressions. First, the acted ones were much easier for the computer to recognize. Second, in 90% of the acted cases, participants did not smile when frustrated, whereas in 90% of the natural cases, participants smiled during the frustrating interaction, despite self-reporting significant frustration with the experience. As follow up (Experiment 3), I proposed, implemented, and evaluated an automated system that can correctly extract and classify sequences containing smiles elicited under delighted and frustrated stimuli. The best classifier distinguished between the patterns of spontaneous smiles under delighted and frustrated stimuli with 92% accuracy. Moreover, individually the classifier was able to identify all the instances of smiles under frustrated stimuli correctly, compared to below-chance performance (53%) of humans. Meanwhile, the performance of recognizing delighted smiles was comparable between humans and machines. The final study (Experiment 4) presented in this chapter offered new findings on the dynamic properties of polite and amused shared/unshared smiles in the banking context. I hope that the findings presented in this section will motivate progress beyond teaching people that "smiles mean happy" and lead to developing methods to interpret challenging spontaneous data that contain complex patterns of expression.

Along with facial expressions, I surveyed and described the algorithmic details of some of the selected prosodic features intonation, pitch, weak language, and loudness. The algorithms were motivated in the context of Human-Computer Interaction and could be implemented as part of a real-time system.

The next chapter provides details on the iterative process that was followed to design training interfaces for social skills training. It addresses the difficulty of displaying multimodal nonverbal data to the user in an intuitive format and proposes a solution using a user-centric approach.

Chapter 5: Training Interfaces for Social Skills

Sensed nonverbal data are nothing but a noisy multidimensional array to a computer. How would one take these numbers and present them to a user with very little training? Some of the nonverbal behaviors are interaction-specific and could be perceived as complex by the user. For example, smile intensity is a continuous stream of data. There are prosodic details in the speech including intonation, loudness, pauses, speaking rate, word level prominence, and pitch. The transcription of the interaction could contain useful information (e.g., number of filler words). Given that the interactions could be as long as a few minutes, a fine-grained analysis would entail a large number of multidimensional arrays. Do we let the interface display the behavior, convert it into action items, and provide the results to the users? Should the interface abstract the behaviors, visualize them, and let the users interpret them on their own (Consolvo et al., 2008)(Patel et al., 2013)(Lindström et al., 2006)? Should the interface summarize the entire interaction and provide a brief summary so that the users don't get overwhelmed (T. Kim et al., 2008)? What happens if the user wishes to watch the entire interaction details through an interface that is intuitive and generalizable? Can the interface ensure that the user is able to generalize the data in the context of the interaction, making it a fun and educational experience?

This chapter attempts to answer all the questions raised above through an iterative design process to develop the training interfaces to enhance nonverbal skills. The design considerations were guided by intuition, relevant literature, and findings from the previous iteration. In this chapter, four such iterations are visually illustrated.

5.1 ITERATIVE DESIGN OF FEEDBACK

Based on the findings presented in *Chapter 2.1.2: The Codebook of Nonverbal Behaviors,* interpretation of nonverbal behaviors is a subjective process. Let us consider one of the previously stated examples on speaking rate. For many, a rapid speaking rate could mean a confident speaker. However, for others, speaking fast could mean a nervous speaker rushing to conclude. Similar phenomena exist for pauses and other nonverbal behaviors. There are also gender differences in nonverbal behaviors. For example, Mast and Hall found that women exhibit higher status with a downward head tilt while males exhibit higher status with "formal dress and a forward lean" (Mast & Hall, 2004). Therefore, in this thesis, I leveraged methodology from (Consolvo et al., 2008) (Lindström et al., 2006) (T. Kim et al.,

2008) (Patel et al., 2013), and took the approach of designing novel interfaces that could represent nonverbal behavior in an educational format, making it easy for the users to interpret them on their own. My goal was to design visualizations that were easy for users to understand and interpret and that enabled them to make comparisons across sessions.

I performed four iterative design exercises to understand how we might best visualize the transcribed speech, prosodic contour, speaking rate, smiles, and head gestures in an interface that is appealing and insightful. This process and the resulting design are summarized in the paragraphs below.

5.1.1 Iteration 1: Prosodic Interpretation

Design

Motivated by ProsodicFont (Rosenberger, 1998), the initial sketch had the idea of adding shape, motion and orientation to the alphabetical letters of the participant's spoken words by prosodic interpretation of the speech signal. If the user puts more emphasis on certain words, they could be indicated by a darker color or larger font size. The intonation of each word could also be displayed by either orienting the words upward or downward to indicate their pitch characteristics (rising, falling or neutral). The distances between each word could represent the pause time in between them. These characteristics of the text could represent the entire interaction through embedded emotional overtones on a timeline. A preliminary example is demonstrated in Figure 5-1.

Evaluation

To test whether this format of representation is intuitive to the user or not, I created examples of all the possible variants of the graph shown in Figure 5-1. The variation parameters are shown in Table 5-1. All the examples were sent out to several mailing lists of our university; around 350 people (55% male, 45% female), mostly in the age range of 21-40, participated in the study. Here is the link to the survey that users filled out.

https://www.surveymonkey.com/s/visualizationparameters

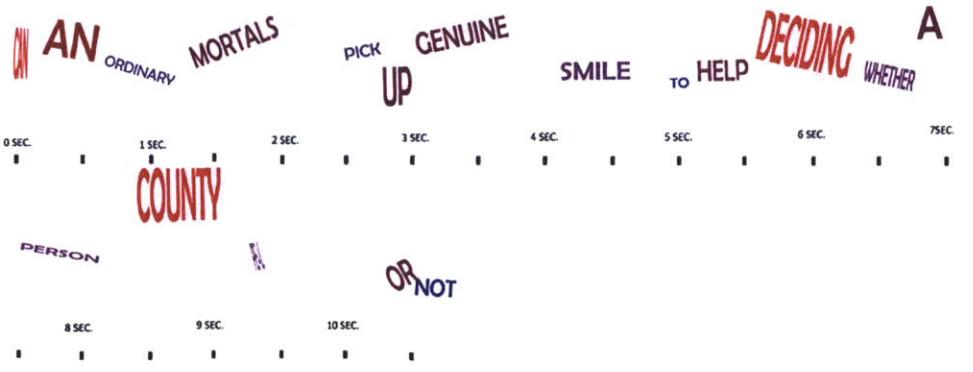


Figure 5-1. Feedback interface of iteration 1. This version adds shape, motion, and orientation to the alphabetical letters of the participant's spoken words by their prosodic interpretation.

Findings

The results were somewhat inconclusive with a handful of participants mixing up the color schema, and other design artifacts. Some of the comments from the questionnaire include:

- *"The color seems to also suggest loudness. I kept going for big red words. Slope and positions were irrelevant to me, I think."*
- *"Red louder than blue, usually."*
- *"Color could be tonal quality: red is sharp/edgy, blue is soft/fuzzy."*

Table 5-1. The design parameters and options for the interface

Choices	Recognition Accuracy (%)
Red to blue gradient, without blending (words in red are loud, blue are soft, black are average)	Less than 10%
Red to blue gradient with blending (words in red are loud, blue are soft, average are purple)	58%
Simple red gradient (words in light red are loud, dark red are soft)	57%
Orientation of each word was mapped to its intonation	50% (20% confused it with pitch)
Font size as loudness (e.g., louder words are displayed with larger font)	75% (10% thought it was mapped to intonation
Vertical displacement of words was mapped to pitch	35% (37% thought it was loudness)

5.1.2 Iteration 2: Combining Prosody with Facial Expressions

Design

Figure 5-2 shows the second iteration of the system. In this version, instead of putting artifacts on the text itself, I created form factors that would contain the embedded affective overtones. Each word was wrapped around a blob and the size of the blob corresponded with the loudness of the word. The vertical placement of the blobs was determined using their pitch characteristics (e.g., words that contained high pitch would be at the top and the words that contained low pitch would be at the bottom). This enabled the user to view his/her

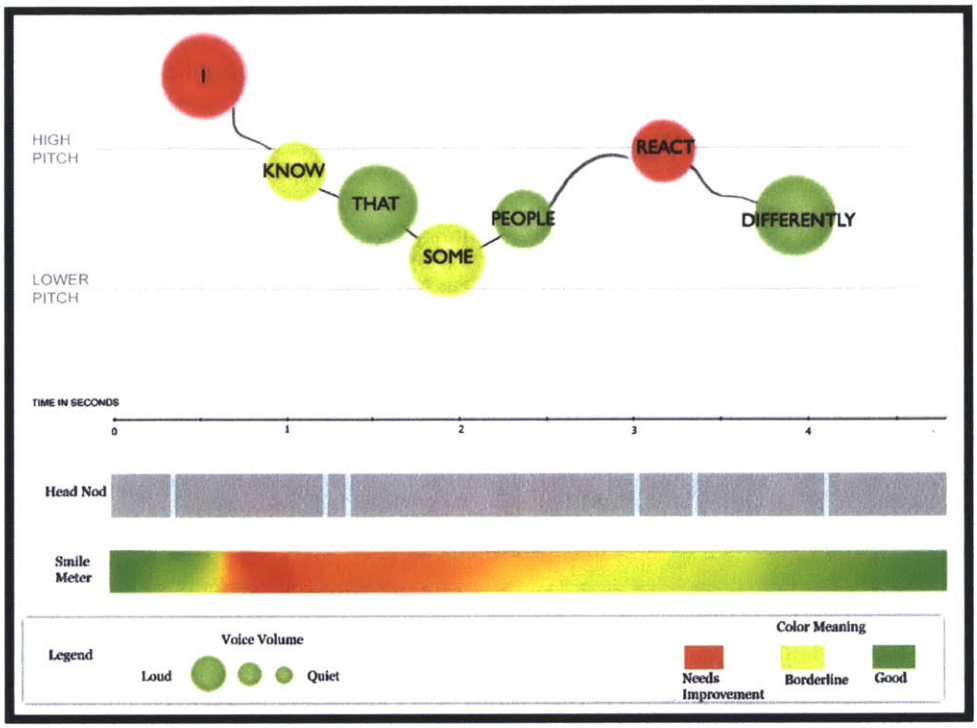


Figure 5-2. Iteration 2 of the interface. The transcribed speech is surrounded by blobs, where the size of a blob corresponds to its volume. The interface also captures smile and head gesture information.

transcribed speech along with pitch variation (i.e., if the blobs almost form a straight line through time, the speaker is more likely to be perceived as monotonous). Based on the findings from iteration 1, I used colors such as green, yellow, and red to symbolize present/absent, or good/bad.

Evaluation

This interface was evaluated through a limited field trial with 10 individuals. They all came to the lab, interacted with the system, and looked through their transcriptions embedded with the affective overtones, as shown in Figure 5-2. At the end, I debriefed the participants.

Findings

- The colors were easy to interpret since the participants were able to relate it to the colors of traffic lights (e.g., red, yellow, green).
- The labels and the legends were also helpful; however, the placement of the legend was deemed confusing.
- The correlation between smile meter gradient and the intensity of smiles was not clear.

- Participants mentioned that they enjoyed interacting with the interface, but they were, to some extent, unable to relate the graphs to what they really did during the interview.

5.1.3 Iteration 3: Prosody, Video, and Analysis

Design

The major change in this iteration was adding the video of the participant during the interaction as part of the interface. This feature was motivated by social-phobia treatment (Clark, 2001), in which people with social phobia are often made aware of how they appear in front of others by having them watch their own video. The interface that I developed, as shown in Figure 5-3, tried to emulate and elicit those features and feelings by juxtaposing the nonverbal data with their video.

Specifically, the following features were implemented:

- The system was redesigned such that it would automatically capture the video of the participant during the interaction, as shown in Figure 5-7.
- The legend was put on the upper-right corner of the interface with additional information.
- Color gradient in the smile meter was eliminated and smiley or "not-smiley" faces were added in the graph itself to indicate the regions where the participants were either smiling or not smiling.
- Head nods and head shakes graphs were combined.
- The occurrences of head shakes were indicated with green and head nods were indicated

Figure 5-3. Iteration 3 of the interface. A video of the participant is added to the interface. The smile meter now contains smiley faces.

with blue.

Evaluation

The same set of participants from Iteration 2 was used for the evaluation of Iteration 3, and I used the same user study methodology as Iteration 2.

Findings

- Some participants were not motivated to watch their entire video and expressed a desire to have an additional interface which would summarize or provide a bird's-eye view of the interaction.
- More than half of the participants, despite not enjoying watching their own video, advised us to enlarge the size of the video for the feedback experience to be more compelling.
- Participants felt that there were too many colors in the interface, and that they moved very fast, making it difficult to follow through.

5.1.4 Iteration 4: Summary and Focused Feedback

In iteration 4, the feedback interface was split into two different phases: 1) Summary feedback, and 2) Focused feedback, shown in (Figure 5-4, Figure 5-5, Figure 5-6 and Figure 5-7).

Summary Feedback

In the Summary feedback phase, as shown in Figure 5-4, participants could track their progress across many sessions (Figure 5-5, Figure 5-6). The upper side of the interface contains the smile track of the entire interaction for each interview session (smile results of multiple sessions are marked in different colors). Through that information, the participant could easily get the idea of how much he/she smiled during the interview and exactly when. The bottom part of the interface contained a configurable set of four dimensions of nonverbal cues:

Total pause duration: The percentage of the duration of the pauses in the user's speech.

Speaking Rate: Total number of spoken syllables per minute during the entire interaction.

Weak Language: Filler words such as "like," "basically," "umm," "totally," calculated as the percentage of the interviewee's spoken words. The complete list of words that I considered as weak was obtained from Blue Planet Public Speaking ("Blue Planet Public Speaking," n.d.) and provided in Appendix A: Weak Language Dictionary.

Pitch Variation: The fourth dimension was the variability in the pitch of the speaker's speech.

Focused Feedback:

Once the user is shown the Summary feedback, he or she is given the option to view the Focused feedback, as shown in Figure 5-7. The design of the Focused feedback was informed by elements of treatment programs developed for social phobia (Clark, 2001). For example, these treatment programs inform individuals with social phobia of how they might appear to others by asking them to watch videos of their own behaviors. One drawback of this approach is that many such individuals view their video appearance negatively. To resolve this problem and to maximize the discrepancies between the individual's self-image and the video, the individual is asked 1) to imagine how they will appear before viewing the video, 2) to create a picture of what their negative behaviors will look like, and 3) to ignore their feelings and watch the video as if it is someone else's. Our design seeks to emulate this strategy and elicit those features and feelings by juxtaposing the nonverbal data with their video. The different variables for which data are visualized are described below and illustrated in on the right.

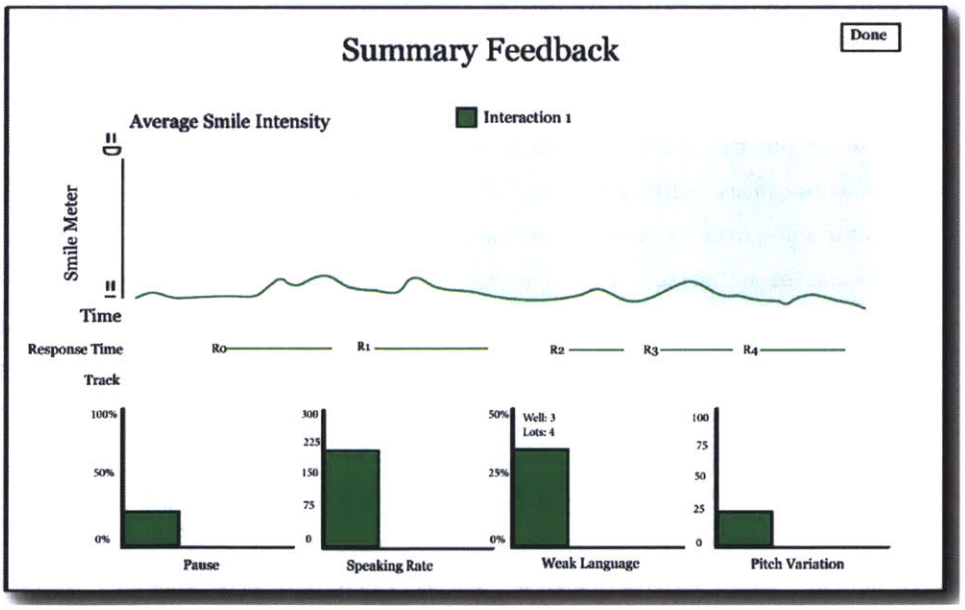


Figure 5-4. Summary feedback captures the overall interaction. Participants can practice multiple rounds of interviews and compare their performance across sessions. This snapshot provides the data from the first round.

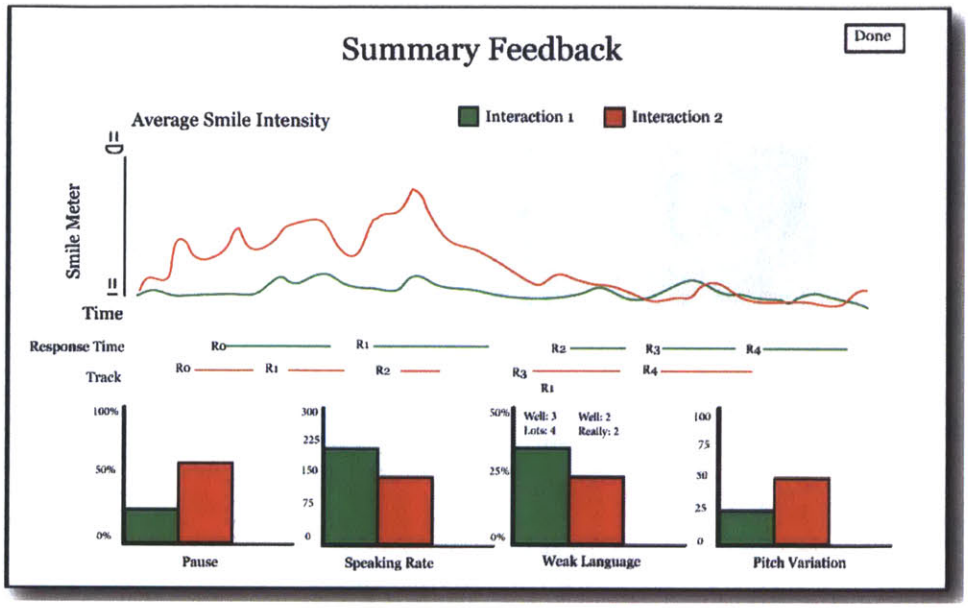


Figure 5-5. Summary feedback captures the overall interaction. This is the snapshot of the second round, where participants can see the data from the last two sessions, side by side.

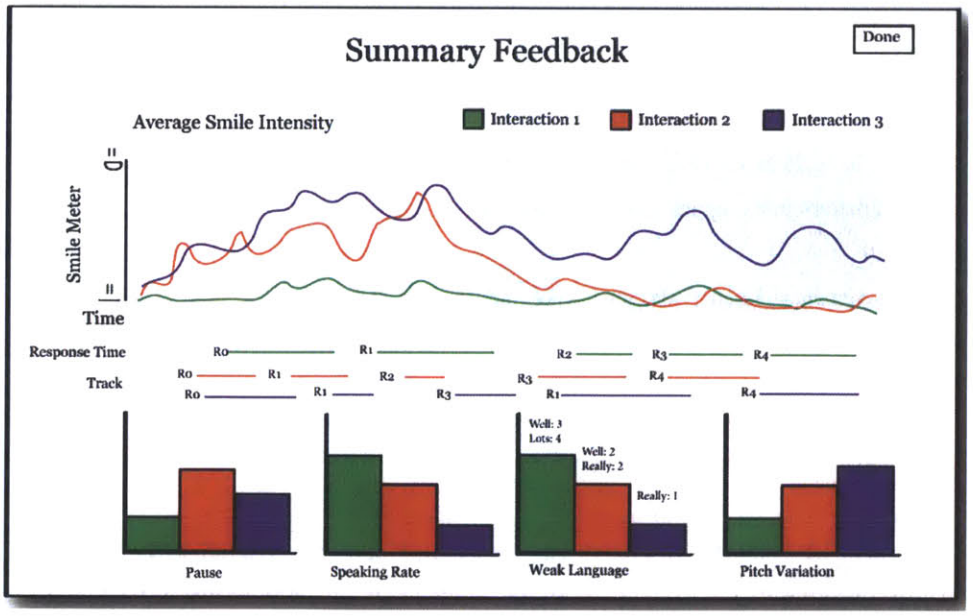


Figure 5-6. Summary feedback captures the overall interaction. This is the snapshot of the third round, where participants can view how the nonverbal properties of their interactions are changing across sessions.

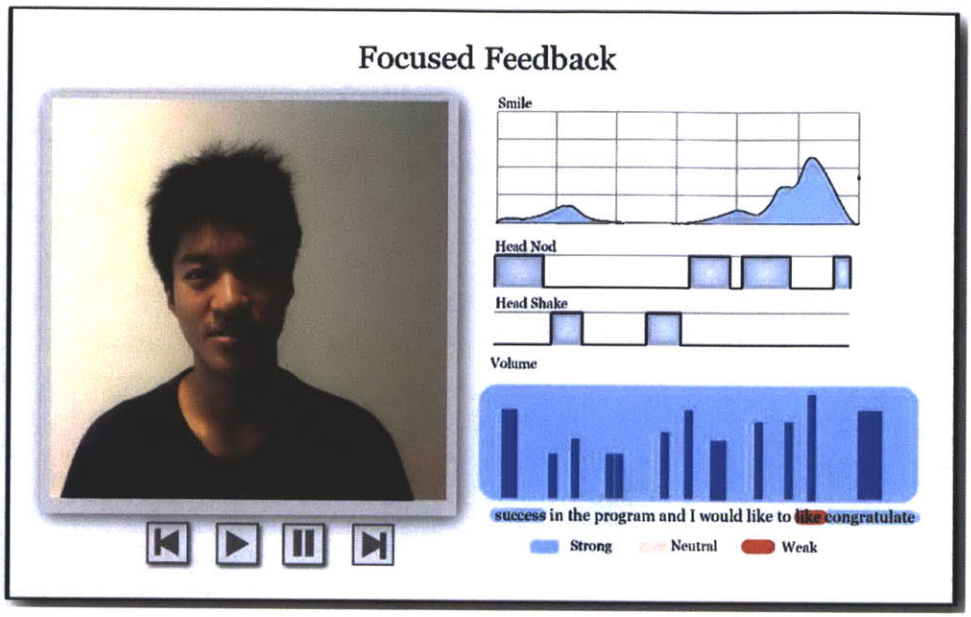


Figure 5-7. The Focused feedback enables participants to watch their own video. As they watch the video, they also can see how their nonverbal behaviors, such as smiles, head movements, and intonation change over time. Participants could watch the focused feedback after they watched the summary feedback at the end of each interaction.

Video: As the participant interacts with an automated system, it captures the video of the interaction using the webcam and displays it to the participant, as shown in Figure 5-7.

Smiles: The system captures intensity of smiles at each frame and displays intensity as a temporal pattern in the upper pane of the interface.

Head Movements: Head nods and shakes are recognized per frame as an output of 0 (not present) or 1 (present). Therefore, they are plotted as binary patterns, as shown in Figure 5-7.

Spoken Words: The spoken words are plotted at the bottom of the interface with weak and strong language marked in red and blue, respectively.

Loudness: The loudness of each word was plotted as a bar, providing the user with a comparative view of the loudness of all the words in an utterance, shown at the lower right part of Figure 5-7.

Emphasis and Pauses: The space that each word occupies in the interface corresponds to the amount of time the user took to enunciate it. Thus, the visualization conveys the emphasis that

the speaker puts on each word with elongation, which corresponds to enunciation time, and height, which corresponds to loudness. The space between each pair of words represents the length of the pause between them.

The focused feedback display provides an opportunity for participants to view both their interview video and data on various nonverbal behaviors as a function of time. This allows the users to identify behaviors across multiple modalities that are out of sync, such as using emphasis improperly or smiling inappropriately, with fine resolution and quantifiable patterns.

In summary, the timeline provides an opportunity for participants to view their own interview video along with various nonverbal behaviors as a function of time. The idea is to look for behaviors across multiple modalities that are out of sync (e.g., putting emphasis on the wrong words, or smiling at odd moments) and make an effort to improve.

Figure 5-8 demonstrates a visual example of how the interface is able to visualize the nonverbal nuances of conversations. Figure 5-8 graphs three different videos with the speaker saying the phrase, "*I am a great candidate; you should definitely hire me,*" while varying certain aspects of his nonverbal behavior. For example, in Figure 5-8 (a), the speaker showed no expression, no head gestures, and flat volume. In Figure 5-8 (b), the speaker seems to increase his volume when he says "strong", and then he emphasizes the word "definitely" by stretching it. In Figure 5-8 (c), the speaker seems to start smiling when he starts to say, "... you should definitely hire me." The user also happens to nod his head when he says "definitely," which would be equivalent to aligning the nonverbal behaviors to appear more compelling.

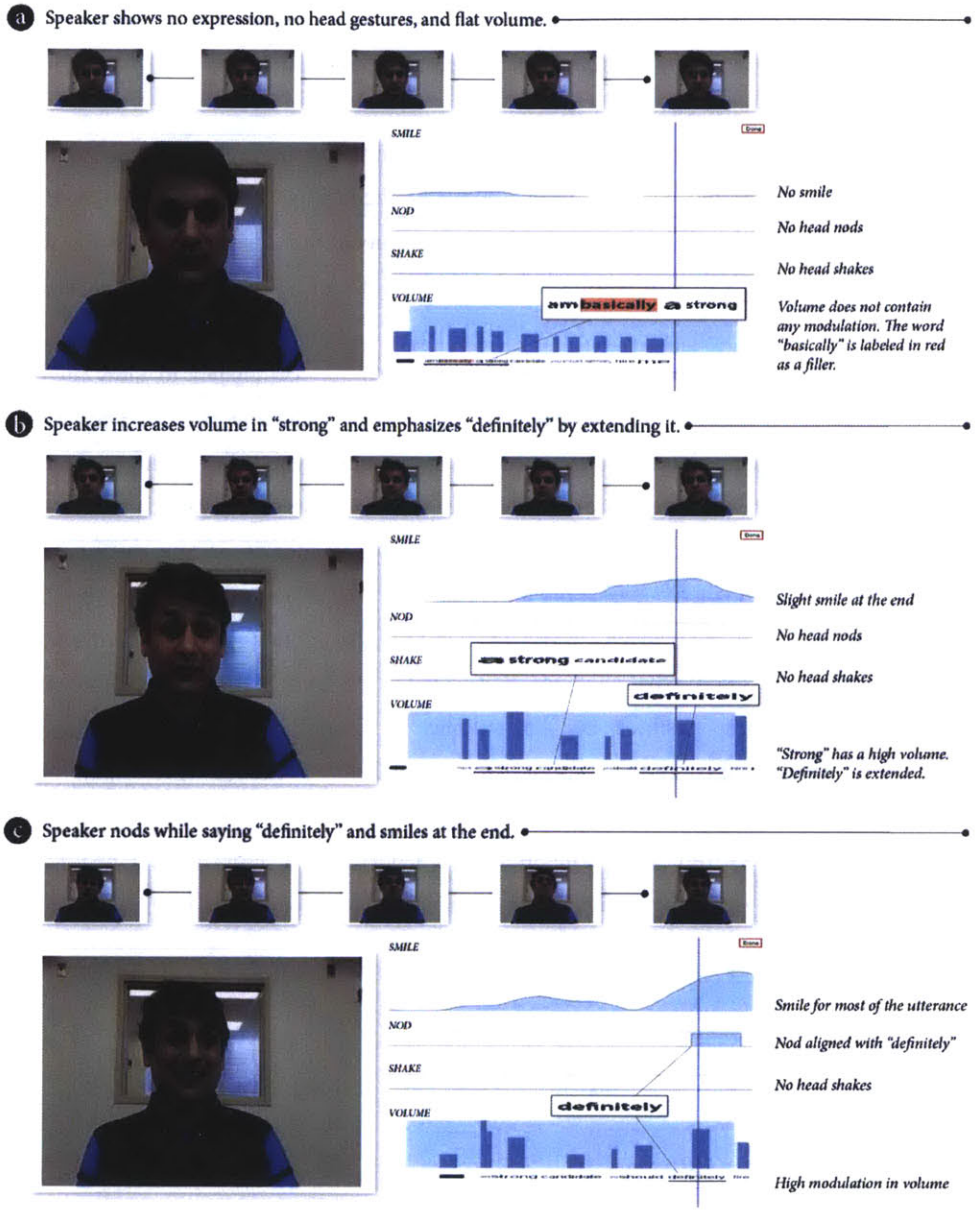


Figure 5-8. Three independent videos (a, b, c) of the same user. In each video, the user says, "I am a great candidate; you should definitely hire me" while varying certain properties of his nonverbal behaviors.

5.2 CHAPTER SUMMARY

This chapter presented the process of converting a multi-dimensional array of behavioral data into an interface that is intuitive, educational, and revealing.

The interface was designed to visualize the nonverbal data based on audio and video input. I started the process based on an intuition of adding shape, motion, and colors in text

data based on the prosodic interpretation of its speech. The user study revealed that the majority of the participants were unable to relate to and interpret the meaning of the graphs.

As part of the second iteration, I added facial features (e.g., smiles and head gestures) along with the prosodic interpretation of the speech signal. I also added colors consistent with traffic lights to indicate good or bad (e.g., red is bad, green is good, yellow is in the middle). Participants liked the color-coded information, but were unable to relate the data to actual occurrences during the interaction.

In the third iteration, I embedded the video data directly in the interface so that the participants are able to contextualize their nonverbal data. I also added a few legends and options so that the participants were able to select/deselect data based on the variables of their interest. Participants appreciated watching their video along with their behavioral data, but wished there were an easy way to get the gist of the interaction right away. Some of the participants felt overwhelmed with many colors, legends, and the speed in which the interface refreshed itself.

As part of the final iteration, I split the feedback into two groups: Summary and Focused Feedback. Summary feedback plotted the smile track for the entire interaction along with pause, weak language, speaking rate, and intonation information. Focused feedback allowed users to view their video along with smile, head gestures, and transcription of the conversation coupled with volume modulation and word emphasis data.

The next chapter provides details of how the nonverbal training interfaces become part of a larger system enabling practice and enhancement of nonverbal behavior.

Chapter 6: My Automated Conversation coacH (MACH)

Figure 6-1. MACH interviews a participant.

This chapter puts together the findings of the previous chapters towards creating a new interaction scenario with computers to enhance nonverbal skills. Let us motivate the rest of this chapter by providing the following scenario.

> *Mike, a technically gifted college junior, worries about not getting any internship offers this year and thinks that he needs to improve his interview skills. After a 15-minute session with a career counselor, he receives recommendations to maintain more eye contact with the interviewer, end the interview with a social smile to appear friendly, and use intonation and loudness effectively to express enthusiasm. Mike returns to his dorm room with an understanding of several behaviors that he can improve for his upcoming interviews. He wishes to practice with and get feedback from a counselor, but schedule conflicts and limited counselor availability make it difficult. He is also unwilling to ask his peers for help, as he fears social stigma.*

Is it possible to help Mike and others like him improve their social skills using an automated system that is available ubiquitously — where they want and when they want?

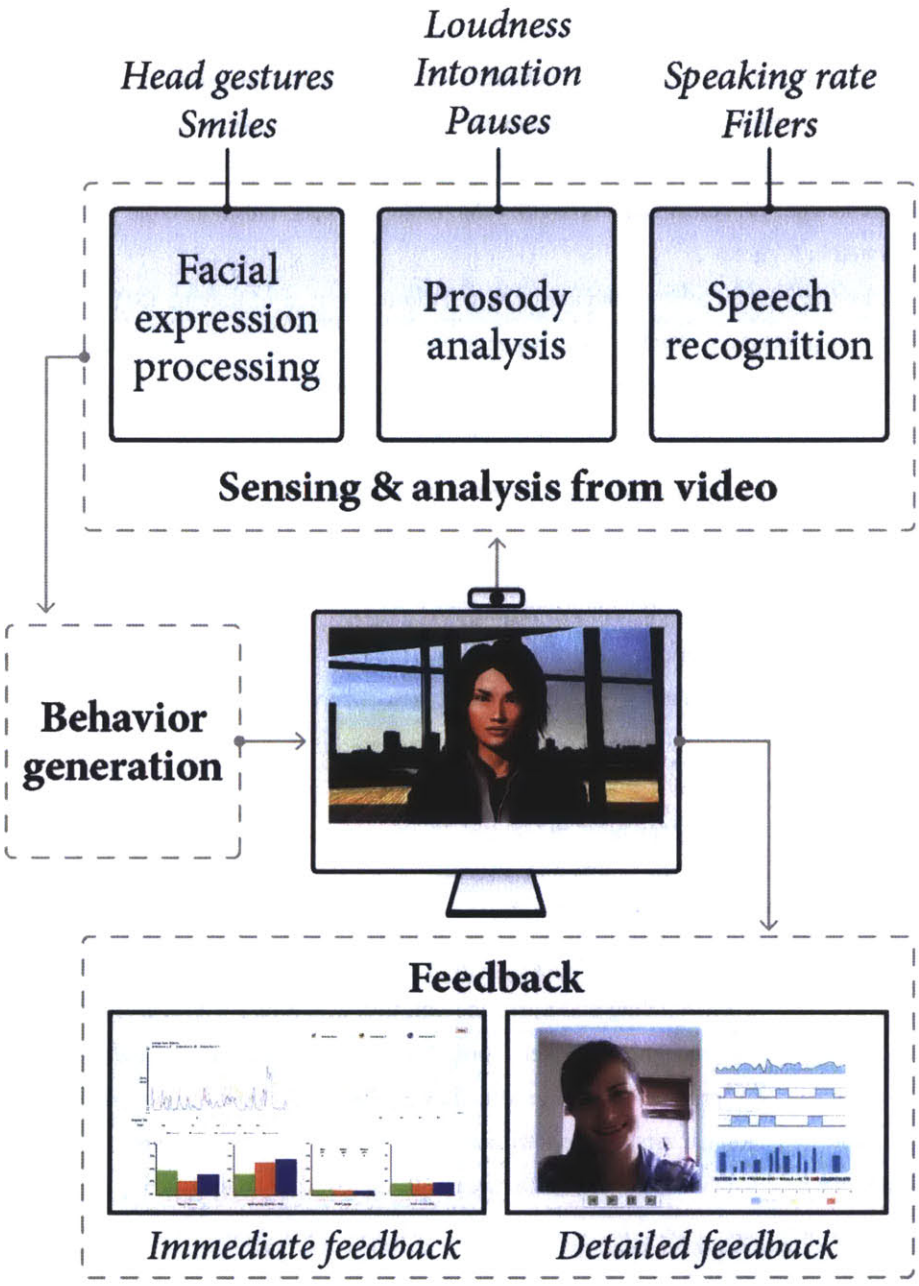


Figure 6-2. The MACH system works on a regular laptop, which processes the audio and video inputs in real-time. The processed data is used to generate the behaviors of the 3D character that interacts with and provides feedback to participants.

This chapter presents the design of MACH—My Automated Conversation coacH—a novel technology that automates elements of behavior analysis, embodied conversational

analysis, and data visualization to help people improve their conversational skills (Figure 6-1).

In this chapter, to create MACH, I integrate the components from the technology exploration described in the previous chapters including facial expressions processing, speech recognition, prosody analysis, conversational virtual agents, behavior synthesis engine, and training interfaces. MACH automatically processes elements of facial expressions and speech and generates conversational behaviors including speech and nonverbal behaviors, as illustrated in Figure 6-2.

In this thesis, I explore the use of MACH in the context of training for job interviews. During an interaction, MACH asks common interview questions used by human interview coaches, and after the interaction, it provides interviewees with personalized feedback to enable self-reflection on the success of the interview.

6.1 RESEARCH QUESTIONS

By incorporating the technical components towards developing MACH, I seek to address the following research questions:

1. How might computer technology use counseling as a metaphor, offering an individual the ability to practice social interaction and to receive useful feedback?
2. How would a system automatically sense, interpret, and represent conversational multimodal behavioral data in a format that is both intuitive and educational?
3. Will such technology elicit measurable improvements in social skills in participants?

6.2 INTERACTION DESIGN

The goal in designing MACH was to create an autonomous system that appears responsive, has backchanneling and mirroring abilities, acts aware with real-time processing of facial expressions, recognizes spoken words along with their intonation, seems life-like in size and resolution, provides real-time affective feedback, and positively impacts interview skills. The interaction must be non-intrusive with all the sensing conducted via a standard microphone-enabled webcam (i.e., no wearable headset or microphone or any other physiological sensors should be required). MACH should enable the following scenario for Mike and others like him.

> *Mike chooses to use the MACH system after class and over the weekend to improve his interview skills. He chooses one of the two counselors (shown in Figure 6-3), John, who appears on the screen, greets him, and starts the interview. When Mike speaks, John periodically nods his head to express acknowledgement, shares smiles, and mirrors*

Mike's head movements. After the interview, John asks Mike to review the feedback on the side of the screen. Mike sees his own smile track for the entire interaction and notices that he never smiled after the opening. He also gets measurements of his speaking rate, intonation, and duration of pauses. Mike also chooses to watch the video of himself during the interview, which helps him identify when his speaking volume was not loud enough for some segment of his answers and when his intonation went flat. He decides to continue practicing in his dorm room, at his parents' home over the weekend, and at the cafe where he usually studies, while the system keeps track of his performance across sessions, allowing for objective comparisons and self-reflection.

6.2.1 Design of the Automated Coach

Figure 6-3. The female and male coaches used in the MACH system.

In order to create a realistic interview experience, I focused on the following components motivated by the findings from Chapter 2.

Autonomous

MACH has been designed to be autonomous. Once started, it is able to execute its instructions on its own and conduct the interview session autonomously. Once the first round of interview ends, the participant is given the option to either terminate the session or go for another round of interview.

Embodied

As discussed in Chapter 2, the use of highly humanlike representations might elicit feelings of eeriness, an outcome often described as the "Uncanny Valley Effect" (Mori, 1970). However, in this work I take the position that interviews are often stressful and a visual and behavioral representation that supports the appearance of an intelligent and dominant

conversational human partner and elicits stress can help in creating a realistic interview experience (Justine Cassell, 2001) (Mutlu, 2011).

Speech Recognition

MACH enables the interaction with the participants in English and can follow simple speech commands. It saves the entire transcription of the interaction and displays it to the user for post review. However, performing semantic analysis on the recognized speech to determine the quality of the answers remains for future work.

Prosody Analysis

The prosodic model introduced in Chapter 4 was implemented in the MACH system. MACH was able to measure the perceptual loudness of each word and plot them along with the transcription of the interaction to indicate volume modulation. It could automatically calculate the pauses between each word and plot the *duration between words* consistently as *distances between words*. MACH could also understand word level prominence. For example, enunciation time of each word was used to vary the space that each word took to occupy in the interface. MACH also measured pitch variation during the interview in real-time.

Facial Expression Analysis

For the job interview scenario, I only enabled MACH to recognize participants' smiles, head gestures, and head tilts. MACH's ability to recognize more subtle smiles and provide differential feedback based on the type of smile remains for future work.

Training Scenario

MACH was designed through a contextual inquiry as a robust framework so that it could self-sustain as a job training tool. Participants could come in and interact with it. I do recommend that they watch a 2 minute introductory tutorial video. It was designed as an interview training platform that is consistent, repeatable, and measurable.

Affective Feedback

In the human interview study, explained in Chapter 3, career counselors maintained a neutral expression during the entire interview process with occasional nonverbal backchanneling behaviors (nodding of head, sharing smiles) and used more expressive language during the feedback session after the interview. Therefore, I designed MACH to display neutral acknowledgements in response to user behaviors and provide more detailed feedback at the end of the interview. In addition, I decided to design the Summary feedback at the end of the interview in the form of interactive visualizations in order to capture the finer aspects of the interviewee's behaviors, which MACH might not be able to effectively communicate using speech.

Hufcutt et al. argue that social skills are the most important construct during job interviews (A I Huffcutt et al., 2001). Motivated by that finding, MACH was designed to provide feedback on the nonverbal behavior of the interviewees as opposed to focusing on the quality of their answers.

Because what constitutes "good interview performance" is largely subjective (discussed in *Chapter 2.1.2:The Codebook of Nonverbal Behaviors*), and the development of an objective metric for interview performance is an open question, I chose to design visualizations that enabled users to engage in a process of guided self-exploration and learning in the interaction context. The design of these visualizations involved an iterative design process comprising several rounds of visual design, implementation, and formative evaluation. This process has been described in Chapter 5.

Automated Backchanneling

Cassell (J Cassell, 2001) states that a virtual character's behaviors make up a larger part of its affordances for interaction than does its appearance. Cassell and Tartaro (Justine Cassell & Tartaro, 2007) argued that along with embodiment and anthropomorphism, virtual characters should also follow the "behavioral" affordances of human communication. For example, nodding and smiling at appropriate times to acknowledge the interviewee's answers to questions might make MACH more credible than a virtual character that stares at the interviewee the entire interview. In addition, it has been shown in GrandChair (Smith, 2000) that a nodding embodied character led to longer responses than a tape-recorder asking the same questions. Thus, one of the design considerations for MACH was to make it appear responsive to and aware of the interviewee.

To automate the process of backchanneling, MACH mirrored head tilts and smiles with the participants (*Chapter 6.3.5: Timing and Synchronization*). It also nodded its head periodically (drawing upon the insights from *Chapter 3.2: Contextual inquiry*) to provide the illusion that it is paying attention to what the listener is saying.

Validation in Real Scenarios

MACH was designed to be validated with MIT undergraduate students in junior standing. The students were likely to be looking for internships and wanted a lot of practice with job interviews.

6.3 NONVERBAL BEHAVIOR SYNTHESIS

One of the important aspects of practicing nonverbal behavior is the conversation partner—in this case—MACH a humanoid 3D character who is able to converse with its users. I use an existing life-like 3D character platform called Multimodal Affective Reactive Characters (MARC) (Courgeon, Buisine, & Martin, 2009). In order to provide users with a

realistic interaction experience, MACH must appear and behave humanlike, adapting its behaviors to changes in the interaction. The implementation sought to achieve this level of realism by integrating the following four components into the animation of the virtual coach: arm and posture movements, facial expressions, gaze behavior, and lip synchronization.

6.3.1 Arm and Posture Animation

I designed a set of arm and postural animations to replicate behaviors that we observed in videos of our human interviewers, such as crossing arms, laying arms on the table, balance shift, and a number of hand gestures that accompanied speech.

Figure 6-4. Motion Capture Stage

The most efficient way to create life-like movements is using motion capture technology. For this application, we used the Optitrack® motion capture system to record postural animations. Our system uses 12 infra-red cameras and a set of 34 markers on the actor's body. As the virtual character is sited, we placed a chair in the center of the motion capture stage, Figure 6-4 and Figure 6-5, and recorded motion.

Using this setup, we recorded about 5 minutes of idle motions, and a few hand gestures. These motions were then exported as Bio Vision Hierarchy files (BVH) using Arena ("Arena," n.d.) and imported in MARC. Using this system, we were able to generate subtle and natural motions, such as balance shifting, that are difficult to create artificially. However, motions such as crossing arms occlude too many IR markers, resulting in the recorded motion being incorrect. Furthermore, it is challenging to use large props, such as a table, during the recordings. Therefore, body postures such as crossing the arms were manually created using MARC's dedicated editor Figure 6-6.

Figure 6-5. A participant is generating data using Optitrack motion capture system.

6.3.2 Lip Synchronization

In MARC, the lip synchronization was achieved using phonemes generated by Cereproc ("CereProc," n.d.) while generating the synthesized voice. Phonemes are converted to visemes, the geometry of the lips, and animated using curved interpolation.

6.3.3 Gaze Behavior

The implementation of the virtual agent's gaze behavior involved directing the agent's eyes and the head toward specific gaze targets such as the user's face, and simulating saccades (rapid movements of the eyes between and around fixation points) and blinks.

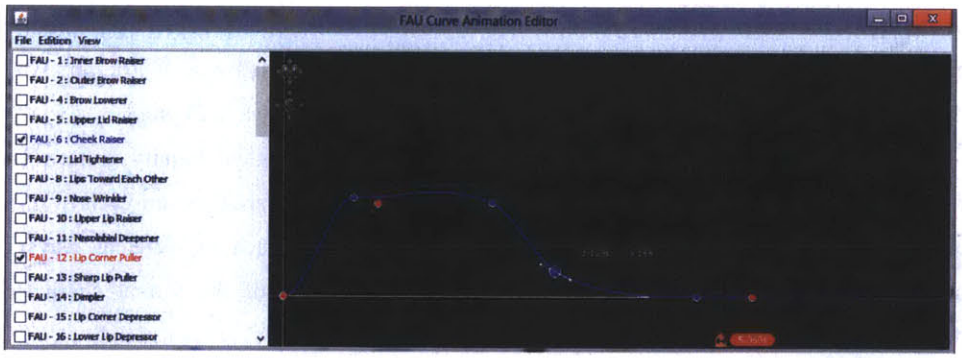


Figure 6-6. Facial Action Unit Curves Editor

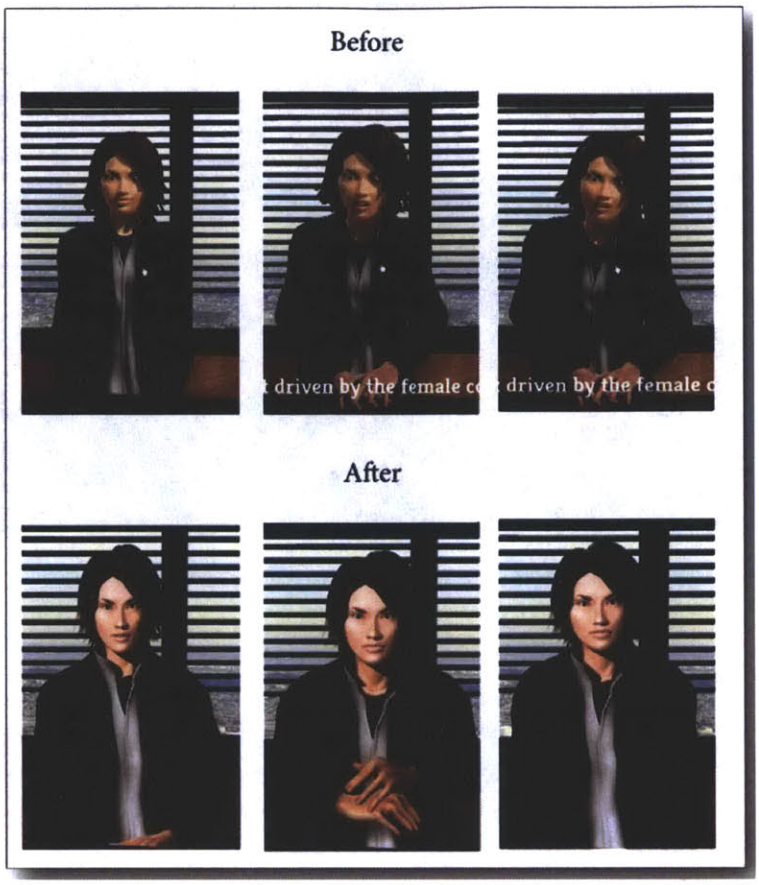


Figure 6-7. (before)The appearance of the virtual agent available through the MARC platform. (after) modification of the character in collaboration with LIMSI-France for the MACH platform.

6.3.4 Facial Animation

Facial behavior involves several communication channels, including facial expressions of emotions, movements of the eyes, and lip movements, among others. Facial expressions of emotion were created by controlling Facial Action Units (FAU) (Ekman & Friesen, 1978) through spline-based animations using a custom-built editor, as shown in Figure 6-6.

Common facial expressions that I observed in the contextual inquiry were social signals, such as polite smiles, and head nods. I designed several variations for each of these behaviors. Using several animations for a single expression, such as different ways of nodding, I sought to increase the dynamism and spontaneity of the virtual character's behaviors.

Head orientation and smiles were sensed by MACH and were mirrored exactly. However, the head orientation and smiles were extracted in every frame (30 times a second). It would appear unnatural to send commands to control the smiles and head movements of an agent that frequently. Therefore, I implemented a smoothing function and sent the averaged

parameters every second. This ensured that the agent wouldn't respond to sudden spikes in smiles or head movements. I also implemented functions in the MARC framework so that the agent wouldn't stop the mirroring behavior because of a sudden failure in tracking. If it loses data in the middle of mirroring an action, it follows an interpolation function to default to its neutral position. These functionalities ensured that the agent's behavior would appear gradual and in synch. Examples of the virtual agent mirroring smiles and head tilts are provided in Figure 6-8 and Figure 6-9 as part of successive frames. Head nods are, however, are not mirrored, but are controlled by a dedicated behavior manager.

6.3.5 Timing and Synchronization

I had to ensure that the animations had been designed were played at the appropriate times during the interaction. Because these interactions primarily involved MACH asking questions and listening to user responses, the virtual agent employed the majority of the behaviors during listening. In order to coordinate these behaviors in a realistic way, I implemented a *listening behavior module.* In the implementation, I leveraged the observations made by Cassell and Thorisson (Cassell & Thórisson, 1999) that nodding of the head following a regulatory function, and even independent of the speech content could be very useful in human-agent interfaces.

An analysis of the videos collected from the mock interviews revealed that in most interviews, counselors nod their heads to signal acknowledgment on average every 4.12 seconds. I utilized this observation by dynamically triggering and combining variations of head nods, arm and postural movements, and real-time mirroring of subtle smiles and head movements in the *listening behavior module*. The frequency at which the listening behaviors were exhibited was modulated by a randomly generated rate of frequency, as shown below. The value of x is randomly generated within the boundary of 0-1500 ms.

$$\textbf{\textit{Head nod frequency}} = 4.12\ s\ \pm x, \textbf{\textit{where}}\ 0 \leq x \leq 1500ms$$

Additionally, an animation was randomly selected from the set of nodding animations that I generated to prevent repetition and achieve spontaneous behavior. The listening module allowed smooth interruptions if the user ended his or her turn in the middle of an animation.

Mirroring Smiles

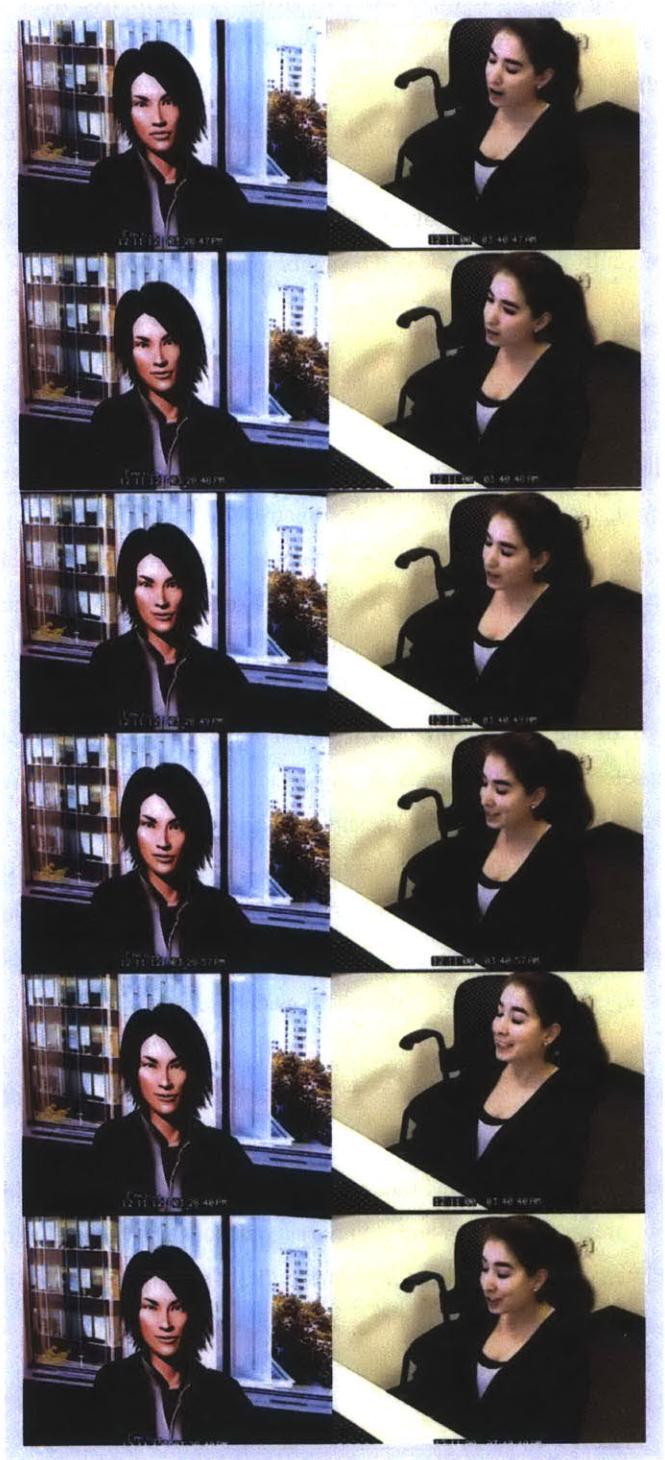


Figure 6-8. The virtual agent is mirroring participant's smiles in real-time. This figure displays successive frames of a video showing the progression of the mirroring.

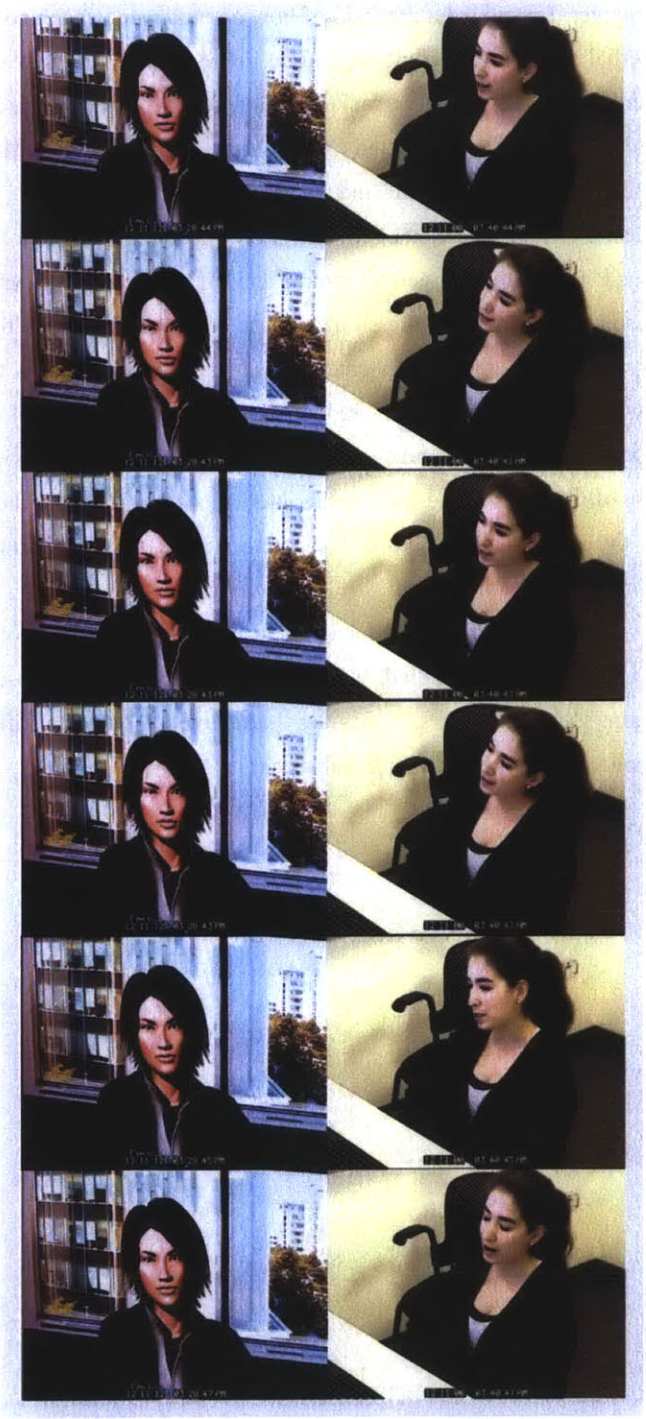


Figure 6-9. The virtual agent is mirroring the head tilt of the participant in real-time. This figure displays successive frames of a video, showing the progression of the mirroring.

6.4 CHAPTER SUMMARY

In this chapter, I provided a scenario of how an automated coach could be helpful in enhancing nonverbal behavior of people looking to improve their interview skills. I motivate the findings from the previous chapters, including Chapters 2, 3 and 4 towards developing an automated interview coach: My Automated Conversation coacH (MACH). MACH consists of a 3D virtual agent that can "see," "hear," and "respond" to its users.

The appearance and behavior of MACH were modeled using data from 28 practice interview sessions between MIT career counselors and MIT students. The system uses speech recognition, prosody, and facial expression analysis to perform the multimodal behavior sensing. The system allows the virtual character to use the sensed data to automatically synthesize behaviors including the arm and posture animation, lip synchronization, gaze behavior, facial animation, timing, and synchronization mimic certain behaviors (e.g., smiles, head movements) of the interviewees during the interaction. At the end of the interaction, the system displays the nonverbal nuances of the interaction as part of an interface using an educational and intuitive format.

MACH was designed to not only have, but also to integrate aspects of the following 9 characteristics: autonomous, embodied, understands users' speech, prosody, and facial expressions, trains human users, provides affective feedback, offers automated backchanneling, and can be validated with real scenarios.

Chapter 7: Evaluation of MACH in the Job Interview Scenario

The evaluation of MACH sought to answer the following three research questions:

1. How effective is MACH in helping users improve their interview skills?
2. Do users think MACH is paying attention to them?
3. Do users find MACH easy to use and helpful?

7.1 EXPERIMENTAL DESIGN

To answer the questions raised above, I designed a user study with three experimental groups and randomly assigned participants to one of these groups, as shown in Figure 7-1. Participants were recommended for the study by the MIT Career Services Office. In Group 1, the control group, participants watched educational videos on interviewing for jobs that were recommended by the MIT Career Services Office. Participants in Group 2 practiced interviews with MACH and watched themselves on video. In Group 3, participants practiced interviews with MACH, watched themselves on video, and received MACH's feedback on their behaviors, interacting with all the functionality available in MACH. This experimental design allowed testing the effectiveness and usability of the design of MACH against a baseline intervention of watching educational videos and against the use of MACH only for practice without feedback. All participants from the three groups were brought into the lab for a first interview with a professional career counselor before being assigned randomly to a group. Participants in the second and third groups were brought back into the lab for an hour-long intervention a few days after the initial interview. All participants were brought back into the lab a week after the initial interview, for the second time for participants in Group 1 and for the third time for those in Groups 2 and 3. They all went through one more interview with the same career counselor but with a different set of interview questions. The counselor was blind to the study conditions.

7.2 PARTICIPANTS

Ninety undergraduate students (53 females, 37 males) were recruited from the MIT campus. The distribution of participants across three conditions based on gender is shown in Table 7-1. It would have been ideal to have equal numbers of male and female participants in each group. However, it was not possible to recruit an equal number of male participants due to lack of participation from the male students at MIT. All of the participants were native English speakers, in junior standing, and likely to be looking for internships. Two

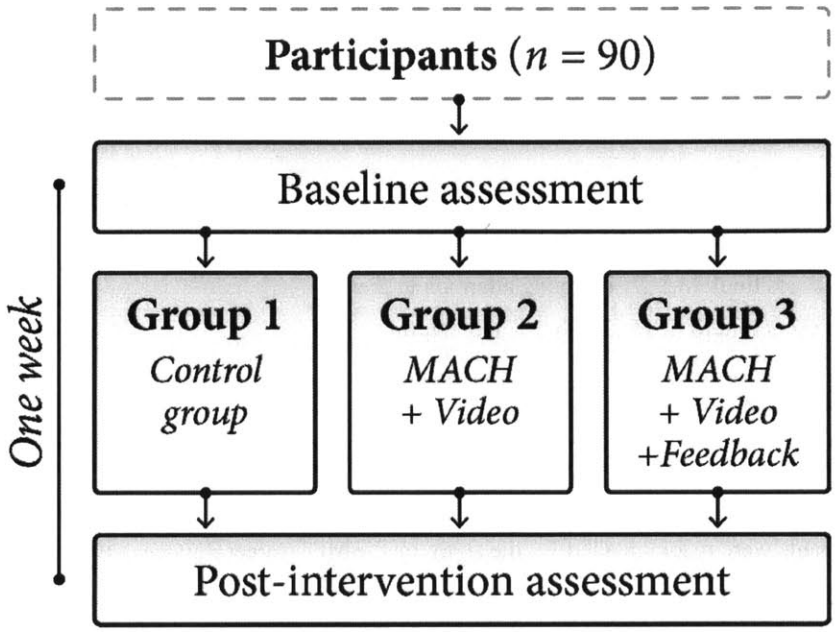


Figure 7-1 Study design and participant assignment to experimental groups.

professional career counselors (one male and one female) with several years of experience in conducting mock interviews and advising students on the interview process were hired. Participants were given a $50 Amazon gift card after the completion of the study, while the counselors were paid $50 per hour (standard rate recommended by the MIT Career Services) for conducting the interview sessions.

Table 7-1. Distribution of participants across conditions based on gender

Gender	Control	Video	Video + Feedback
Male	13	12	12
Female	17	17	19
Total	30	29	31

7.3 EXPERIMENT PROCEDURE

The participants were given a generic job description and were asked to pretend that they were being interviewed for a job position at their favorite company. The study setup of the interview was similar to Figure 3-2. To minimize gender-interaction variability (Nass et al., 1997), male students were paired with the male counselor, and female students with the female counselor. After the mock interview, the counselor rated the interviewee's interview performance, and the interviewee rated his or her own performance.

7.3.1 Intervention Preparation and Protocol

Students in the control group were asked to watch 20 minutes of educational videos (available at: http://tinyurl.com/MITCareerServices) on tips for successful interviews, before they came back for their final round of interview with the career counselor. All the students in the control group confirmed with the experimenter during the final day of the study that they had watched those videos. Students who were in Group 2 or 3 were invited to come back to the lab with an opportunity to interact with MACH. Before they could interact with MACH, participants in Group 2 were asked to watch a two-minute tutorial video (available at http://youtu.be/aO8-TUTllnI) which explained the interface and the interaction designed for their group. Similarly, participants in Group 3 watched a different video tutorial (available at: http://youtu.be/DHzA141L4F8) which explained the interface prepared for their group. The experimenter left the room during the interview with MACH and asked the participant to exit the room once the study was completed. Participants were told that they could practice as many times as they wished, but they had to practice using the system at least once. However, the session automatically terminated after the third practice (details of the possible user actions are demonstrated in Figure 7-3 and Figure 7-4). During the practice, MACH asked interview questions and provided feedback at the end of each interview session.

7.3.2 Apparatus

The experimental setup to interact with MACH is shown in Figure 7-2, in which MACH was displayed on a 46" Samsung Smart 3D TV. The MACH system ran on a Dell laptop (Precision M6700, Intel ® Core ™ i7-3820QM CPU @ 2.70 GHz). The system used a Logitech HD Pro Webcam C920 that was connected to the laptop and was put on top of the 46" display, as shown in Figure 7-2.

7.4 INTERVIEW QUESTIONS

The interview questions were determined with feedback from the MIT Career Services. One of the objectives was to ensure that the participants were not asked the same set of questions repeatedly across sessions. This was done to make sure that the participants did not get to memorize the answers. However, it is difficult to devise variations in questions such as, "Tell me about your background" or "Why do you think we should hire you" which, according MIT Career Services, are the most common and important interview questions. Also, it was desired to have the interview questions follow a consistent structure across sessions so that it would be possible to compare participants' performance before and after the intervention. I worked with the MIT Career Services to reword all the interview questions to prevent participants from memorizing the answers. The questions are described below.

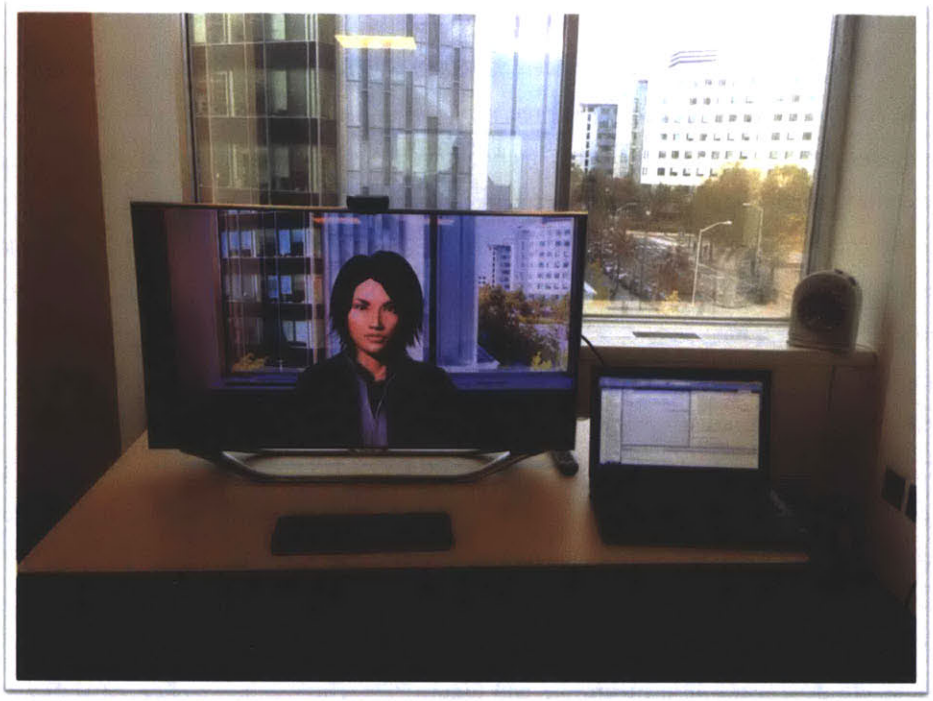


Figure 7-2. Experimental setup between MACH and the participants.

7.4.1 Pre Interview Questions Asked by the Career Counselor

Q1. So, please tell me about yourself?
Q2. Tell me about a time when you demonstrated leadership.
Q3. Tell me about a time when you were working on a team and faced with a challenge, how did you solve that problem?
Q4. Tell me about one of your weaknesses and what you are doing to overcome it?
Q5. Now why do you think we should hire you?

7.4.2 Post Interview Questions Asked by the Career Counselor

Q1. So, please tell me more about your background?
Q2. Do you think you are a leader? Tell me why.
Q3. Do you consider yourself a team player? Can you give me an example where there was a conflict in your team and you had to resolve it?
Q4. If you were given an option to change one thing about yourself, what would that be?
Q5. Why do you think you should get the job?

7.4.3 Interview Questions during Interaction with MACH

1st Round

Q1. Why don't you start with providing some details about your professional background?
Q2. Tell me about a leadership role that you had in the past. What did you find most challenging in that role? What was most rewarding?
Q3. Describe a time when a team member came to you for help. What was the situation and how did you respond?

Q4. What type of situation puts you under pressure? How do you handle it?
Q5. Lastly, what do you believe makes you a good match for this position?

2nd Round

Q1. Firstly, please provide some details about your academic background.
Q2. Could you give an example of when you had to adapt to an unfamiliar environment, and how you did so?
Q3. Describe a time when you saw a problem in a group you were involved with and took action to correct it.
Q4. What are your long-term goals to overcome some of your weaknesses?
Q5. Lastly, why do you want to work for our company?

3rd Round

Q1. Please, tell me a little bit about yourself.
Q2. What are some things you find unique about yourself?
Q3. Could you tell me about a time when you had to explain technical information to an audience with little expertise?
Q4. What is your main weakness and how do you plan to overcome it?
Q5. And finally why do you want this job?

7.5 USER INTERACTION

7.5.1 Video + Affective Feedback

In addition to measuring whether it is possible for individuals to improve their nonverbal skills by interacting with the system, it was also desired to understand how people would use the system. The system was designed such that users would have to interact with it and look through the summary and focused feedback at least once. Once they had completed a full cycle as demonstrated in Figure 7-3, users were given the option to either quit the session or interact one more time. The interaction terminated automatically after the third interaction. A state diagram demonstrating the possible options for users as they interact with the system is provided in Figure 7-3.

7.5.2 Video Feedback

One could argue that it may be possible for users to improve their nonverbal skills just by interacting with the system and then watching their video without the affective feedback. Therefore, I developed a variation of the system to allow for people to interact with it and simply watch their video. The sequence of actions available to the user for that version is shown in Figure 7-4. In the video version, a user interacts with MACH and then watches his or her video through the interface. After one round of interaction, the user could quit the interaction, or continue to do more rounds. The session automatically terminates after the 3rd interaction.

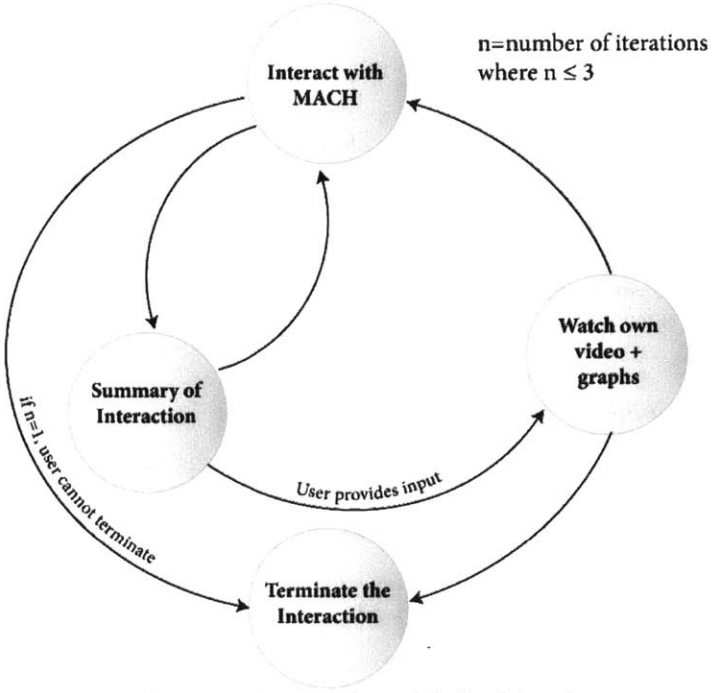


Figure 7-3. Sequence of actions for users trying the video and feedback interface

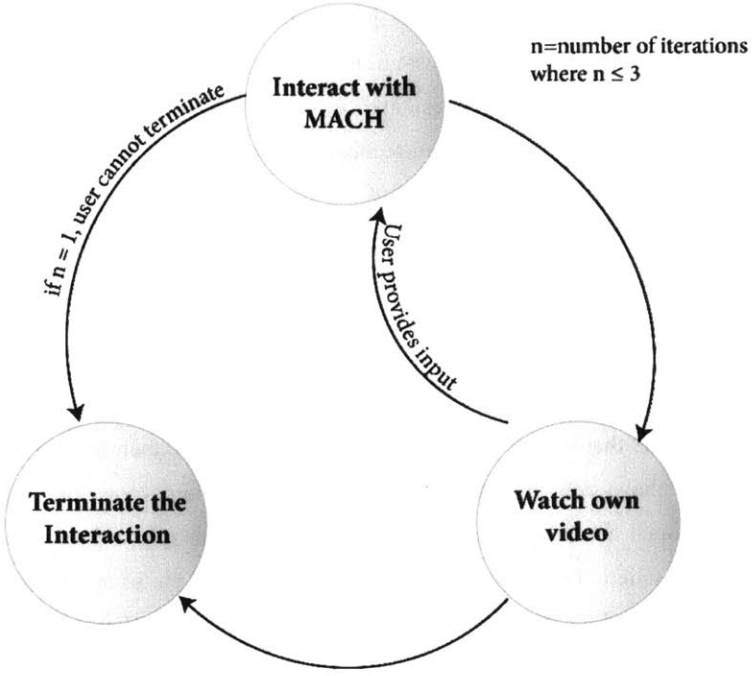


Figure 7-4. Sequence of actions for users trying the video only interface

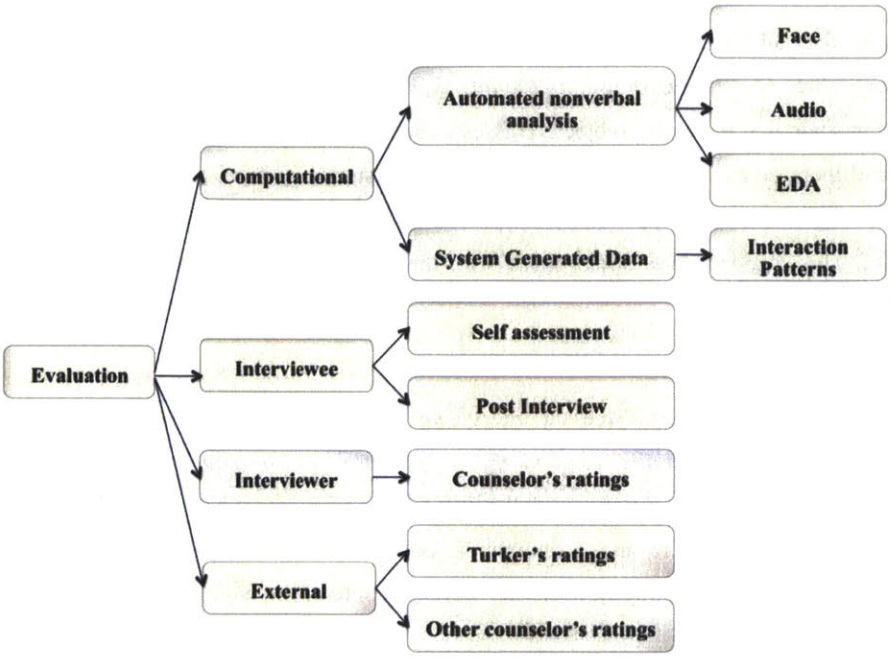


Figure 7-5. Evaluation framework of the intervention

7.6 EVALUATIONS

Figure 7-5 demonstrates the dimensions on which the intervention was evaluated. There were two main dimensions to the evaluations: 1) Human Judges; 2) Computational Assessments.

7.6.1 Human Judges

Interviewee's Self-assessment

Following the baseline and post-intervention interviews, the participant filled out a questionnaire that evaluated the participant's interview performance (Appendix B-1: Questionnaire used in the intervention for the participants.). Motivated by the initial contextual inquiry described in section 3.2.5, the questionnaire included items related to the participant's overall interview performance, presence, warmth, content, and competence, rated on a scale of 1 to 7.

Following the intervention of interacting with MACH, participants in Group 2 (video) and Group 3 (video + feedback) were asked to fill out questionnaires (available in *Appendix B-3: Questionnaire used in the intervention for the students on the MACH system (video group)* and *Appendix B-4: Questionnaire used in the intervention for the students on the*

MACH System (feedback group)) to evaluate the quality of the interaction with MACH. In addition, they all responded to the System Usability Scale (SUS) (Brooke, 1996) (available in *Appendix B-5: Questionnaire used in the intervention for the students on System Usability*), a ten-item scale that measures subjective assessments of usability. SUS scores range from 0 to 100 and increase as the perceived usability of the system increases. Finally, the participants were provided with the opportunity to provide open-ended verbal feedback after the study debrief. This feedback was audio–taped with the permission of the participants for further qualitative analysis.

Interviewer's Judgement

Along with the interviewee, the interviewer also filled out a questionnaire (available at *Appendix B-2: Questionnaire used in the intervention for the counselors/Turkers.*). This questionnaire, motivated by the initial contextual inquiry described in section 3.2.5, 69 included items related to the participant's overall interview performance (e.g., likely recommendation of being hired), presence, warmth, content, and competence, rated on a scale of 1 to 7.

Independent Judges – MIT Career Counselors

In addition to measuring the interviewer's and the interviewee's evaluations of the interview, two independent career counselors—one male and one female—from MIT's Career Services were recruited to rate the interview videos. It was expected that the ratings from these "independent counselors" would be more reliable, because (1) they were blind not only to the study conditions but also to the study phase, i.e., whether an interview was a baseline or post-intervention interview; (2) they did not interact with the participants and thus were less affected by biases that might have been introduced by interpersonal processes such as rapport; and (3) they could pause and replay the video, which might have enabled them to analyze the interviews more thoroughly. A custom-made interface was developed so that the counselors could rate the video using their computer browser, (A screenshot of the interface is provided in *Appendix G: Custom Interface Designed for Counselors and Turkers to Rate the Videos.*) In addition, a functionality was implemented to keep track of "play time" of the video as well as average time spent on assessing each participant to ensure that the counselors were spending a reasonable amount of time in analyzing in each participant.

Independent Judges – Mechanical Turkers

Along with having professional career counselors rate the videos, it was also desired to get more individuals to label the interviews using the Amazon Mechanical Turk interface. An interface was developed where the videos were put online for people to label (a screen shot is provided in *Appendix G: Custom Interface Designed for Counselors and Turkers to Rate the*

Videos). Recruiting reliable labelers in Mechanical Turk is a challenge as we are less likely to have any background information about them. The Turkers are usually motivated to get their work done in the shortest span of time to maximize the output for their time. Therefore, it was important to ensure that the Turkers were really watching the interviews entirely and spending a suitable amount of time going through the questionnaire. 4 interviews with examples of participants speaking very fast/slow, not making eye contact, or not providing relevant answers to the questions being asked were selected. The interface that was developed was re-used for the independent judges (MIT Career counselors) to keep track of the "play time" and "assessing the participant time." If the total time spent on the interface was not significantly higher than the duration of the video, the user entries were automatically rejected. Only four sample interviews were released to super Turkers with at least a 95% of hit rate. Based on the initial ratings of four videos, four individuals were pre-selected who consistently watched the entire videos and spent more than a minute to fill out the questionnaire. I have also used our subjective bias to analyze whether the Turkers' ratings were reliable (for example, picking up on lack of eye contact, or speaking rate for the obvious cases).

Once the Turkers' labels were produced, it was important to measure their agreement. We used Krippendorff's alpha coefficient (Krippendorff, 2004), a statistical measure of agreement when coding a set of units of analysis in terms of the values of a variable across two or more annotators. The formula for the alpha is the following:

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o, or observed disagreement, corresponds to the amount of pairwise disagreement observed between the annotators, and D_e, or expected disagreement, is the level of disagreement expected by chance as calculated from the data. The coefficient alpha ranges from -1 to 1, where α =1, corresponds to perfect reliability, α =0, indicates the absence of reliability, and α = -1 represents systematic disagreement.

7.6.2 Computational Assessment

In addition to the interviewee, interviewer, and external human raters, it was also desired to computationally analyze the differences of videos from PRE (baseline) and POST (after the intervention). The audio, video, and Electro-Dermal Activity (EDA) (Boucsein, 1992) of participants as well as the counselors were recorded. In addition, a logging function was implemented that recorded user clicks and average time spent on each interface as the users interacted with MACH. It allowed objective answering of the questions, "How much time does an average user spend looking at the Focused feedback as opposed to the Summary feedback?" and "Do people in the feedback option spend more time or more iterations with the system than people in the video option?"

7.7 RESULTS

The first step was to analyze the common properties of the interviews across three conditions for differences and inconsistencies. During the interview sessions with the MIT career counselors, there were two segments, interview and feedback. We measured the average duration of interviews and feedback across the three group types, as shown in Figure 7-6. For the average duration of interviews, there were no significant differences among the participants from the three groups. We also grouped the duration in terms of gender. The

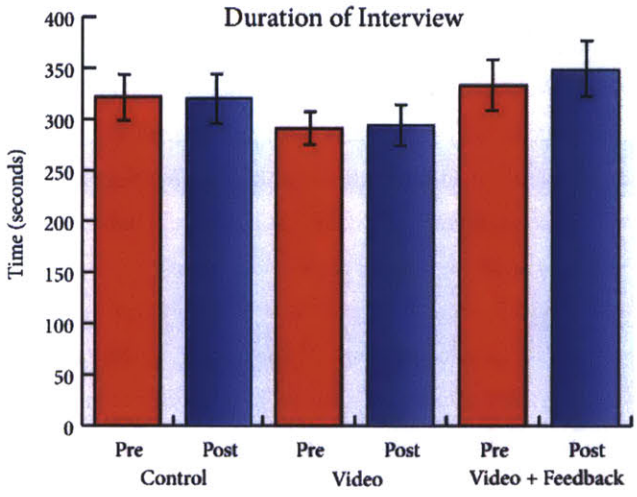


Figure 7-6. Average duration of interview sessions across three conditions.

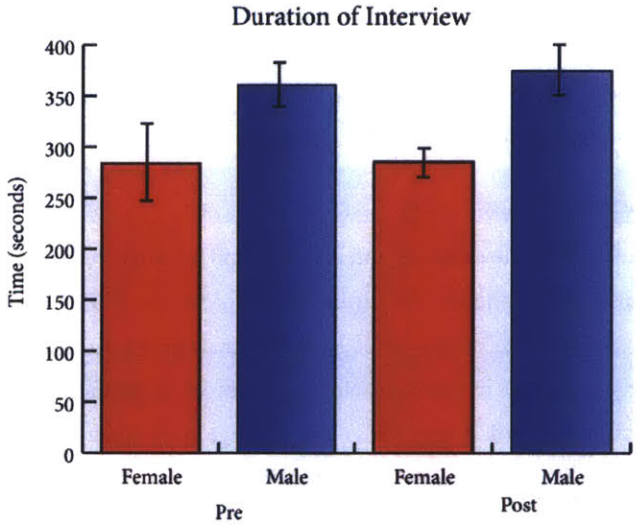


Figure 7-7. Average duration of interview across male and female participants

duration remained the same between PRE and POST interviews for both male and female participants, while male participants' interviews lasted slightly longer than the ones from female participants.

For the remaining sections, I report the statistical analysis of the ratings of the Interviewee's self-assessment, the interviewer's assessments, and the independent counselors and Mechanical Turkers' ratings. The analysis was conducted across several behavioral dimensions. Those behavioral dimensions were motivated by the contextual inquiry outlined in Chapter 3. Some of the behavioral dimensions were rephrased based on the feedback from the MIT career counselors. The dimensions were "overall performance," "recommended for hire," "would love to work with this person as a colleague," "engagement," "excited about the position," "eye contact," "smiled appropriately," "speaking rate," "usage of fillers," "appearing friendly," "paused when needed," "engaging tone of voice," "structured answers," "appeared calm," "appeared stressed," and "looked focused." The figures presented in this section were generated using the difference between the ratings after and before the intervention, i.e., improvement across the three intervention types. The effect of intervention type was analyzed using one-way analysis of variance (ANOVA), and the effects of intervention type and participant gender were analyzed using two-way ANOVA. Planned comparisons between the "feedback" and "control," and "feedback" and "video" interventions involved Scheffé's method.

In the following sections, I only provide statistical details and graphs for the significant results ($p < .05$) across intervention types. Full statistical analysis and graphs on every behavioral dimension across three intervention types are provided in the following appendices.

- *Appendix C-1: One-Way Anova Analysis on the Participant's Self-Ratings (Conditions)*
- *Appendix C-2: Two-Way Anova Analysis on the Participant's Self-Ratings (Conditions with Gender)*
- *Appendix C-3: Graphical Analysis Based on the Participant's Self-Ratings*
- *Appendix D-1: One-Way Anova Analysis Based on the Counselor's Ratings (Conditions)*
- *Appendix D-2: Two-Way Anova Analysis Based on the Counselor's Ratings (Conditions with Gender)*
- *Appendix D-3: Graphical Analysis Based on the Counselor's RatingsAppendix E-1: One-Way Anova Analysis Based on the Other counselor's Ratings (Conditions)*
- *Appendix E-1: One-Way Anova Analysis Based on the Other counselor's Ratings (Conditions)*
- *Appendix E-2: Two-Way Anova Analysis Based on the Other Counselor's Ratings (Conditions with Gender)*

- Appendix E-3: Graphical Analysis Based on the Other Counselor's Ratings
- *Appendix F-1: One-Way Anova Analysis Based on the Turker's Ratings (Conditions)*
- *Appendix F-2: Two-Way Anova Analysis Based on the turker's Ratings (Conditions with Gender)*
- *Appendix F-3: Graphical Analysis Based on the Turker's Ratings*

7.7.1 Interviewee's Self-Assessment

The first step was to measure whether the self-ratings of the participants were going up after the intervention. The differences of the rating (post-pre) were calculated and the percentage of participants whose ratings have gone up, gone down, or stayed the same, as shown in Table 7-2, were measured. The data presented in Table 7-2 demonstrate that most of the participants' self-ratings went up in dimensions such as "overall performance," "appear friendly," "good at pausing in between sentences," "engaging tone of voice," and "focused during the interview." The ratings went down on "I used a lot of fillers," "I appeared stressed during the interview," and "I appeared nervous during the interview." The next step was to determine whether the changes in ratings were statistically significant across the three conditions.

Table 7-2. Absolute changes in participants' ratings. The change is measured by subtracting pre ratings from the post ratings. ++ ratings went up; = ratings did not change; -- rating went down. The number of participants is expressed as a percentage of the overall participants.

	++ (%)	= (%)	-- (%)
Overall performance	44.83	42.53	12.64
Filler usage	29.89	34.48	35.63
Friendly	39.08	41.38	19.54
Pausing between sentences	49.43	25.29	25.29
Engaging tone of voice	32.18	50.57	17.24
Feel stressed	27.59	33.33	39.08
Appear focused	86.21	6.90	6.90
Found feedback useful	36.78	42.53	20.69
Felt nervous	29.89	34.48	35.63

The analysis using interviewees' self-ratings did not show any significant change across intervention types. For the full statistical analysis and the graphs, please see *Appendix C-1: One-Way Anova Analysis on the Participant's Self-Ratings (Conditions), Appendix C-2: Two-Way Anova Analysis on the Participant's Self-Ratings (Conditions with Gender), Appendix C-3: Graphical Analysis Based on the Participant's Self-Ratings.*

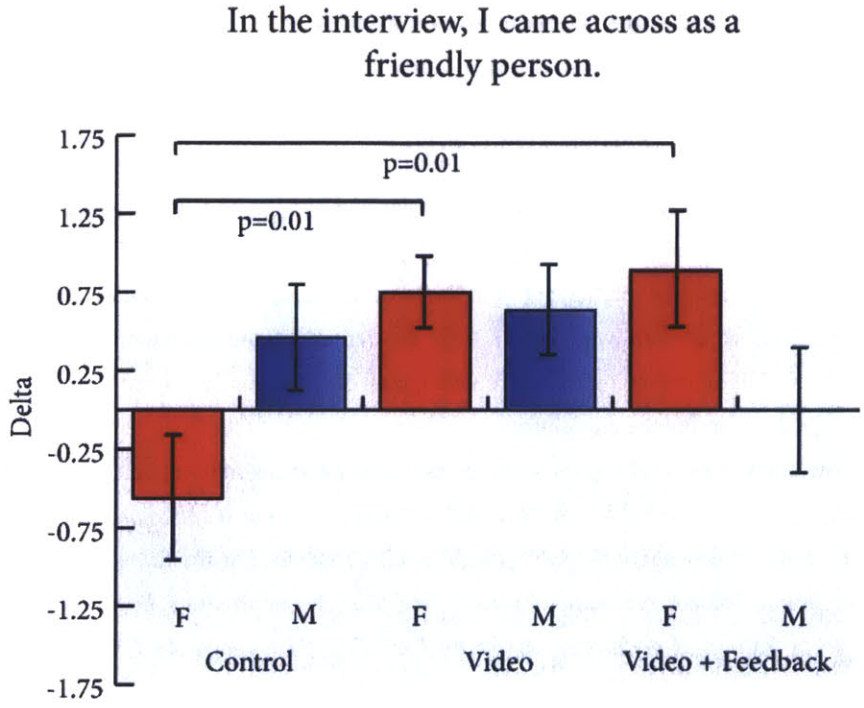


Figure 7-8. Improvement (post – pre) in participant's scores in item, "In the interview, I came across as a friendly person," across conditions, broken down to females (F) and males (M).

The analysis that considered participants' gender showed a significant interaction ($F[2,80]=4.13$, $p=0.02$) between intervention type and gender for the dimension of "friendliness." Comparisons showed that female participants' ratings of themselves in Group 1 (control), in terms of friendliness, were significantly lower than Group 2 (video) ($F[1,80]=6.49$, $p=0.01$) and Group 3 (video + feedback) ($F[1,80]=6.58$, $p=0.01$), as shown in Figure 7-8.

7.7.2 Interviewer's Assessment

In this section, the analysis is presented using ratings from the MIT career counselors who conducted the interviews and rated the interviewees at the end of the sessions. The analysis did not show any significant change across intervention types. For the full statistical analysis and the graphs, please see *Appendix D-1: One-Way Anova Analysis Based on the Counselor's Ratings (Conditions)*, *Appendix D-2: Two-Way Anova Analysis Based on the Counselor's Ratings (Conditions with Gender)*, *Appendix D-3: Graphical Analysis Based on the Counselor's Ratings.*

7.7.3 Independent Judges – MIT Career Counselors

In this section, the results are presented based on the ratings of the MIT career counselors who were not part of the intervention and did not know which viewed segments were PRE- or POST-intervention. The counselors watched the interview part of the videos of the participants (PRE and POST) in random order and rated them using a web browser. We implemented functionality to keep track of "play time" of the video as well as average time spent on assessing each participant. For both of the counselors, average time spent on assessing each participant was longer than the length of the corresponding video. For example, in addition to playing the entire video, both of the counselors spent at least a minute, on average, in assessing the participants.

The analysis showed significant changes across intervention types on the dimensions of "*overall improvement*" ($F[2,83]=4.89$, $p=0.01$), "*love to work as a colleague*" ($F(2,83)=6.67$, $p<0.01$), and "*excited about the job*" ($F[2,83]=3.24$, $p=0.04$). For the full statistical analysis and the graphs, please see *Appendix E-1: One-Way Anova Analysis Based on the Other counselor's Ratings (Conditions), Appendix E-2: Two-Way Anova Analysis Based on the Other Counselor's Ratings (Conditions with Gender), Appendix E-3: Graphical Analysis Based on the Other Counselor's Ratings.*

In the "overall dimension," the participants in the Group 3 (video + feedback) were rated to have demonstrated significant performance increase compared to the Group 2 (video) ($F[1,83]=6.91$, $p=0.01$) and the control ($F[1,83]=7.46$, $p=0.01$), as shown in Figure 7-9. The analysis that also considered participant gender showed a significant interaction between intervention type and gender ($F[2,80]=6.6701$, $p<0.01$).

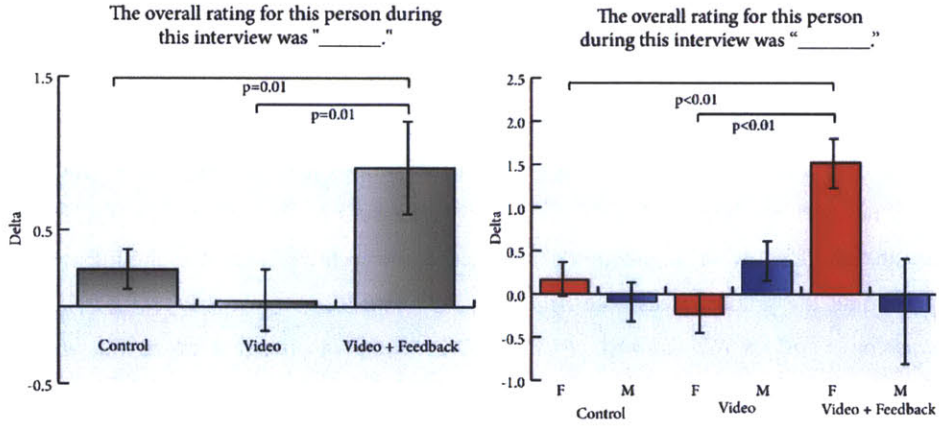


Figure 7-9. Improvement (post – pre) according to the independent counselors ratings in the item, "What was the overall performance during the interview," across conditions (left) and across conditions, broken down to females (F) and males (M) (right).

Comparisons showed that female participants in Group 3 (video + feedback) were rated higher in the "overall dimension" compared to the females in Group 2 (video) (F[1,80]=7.71, p<0.01) and the control (F[1,83]=7.46, p=0.01), as shown in Figure 7-9.

In the "would love to work with this person as a colleague" dimension, participants in Group 3 (video + feedback) group were rated higher than the participants in Group 2 (video) (F[1,83]=12.47, p< 0.01), and the control group (F[1,83] = 6.15, p=0.02). The analysis that also considered participant gender did not show any significant interaction between the intervention type and gender (F[2,80]=0.85, p=0.43). However, comparisons showed that females in Group 3 (video + feedback) were rated higher than the females in Group 2 (video) (F[1,80]=9.18, p<0.01) and the control (F[1,80]=7.38, p=0.01), as shown in Figure 7-10.

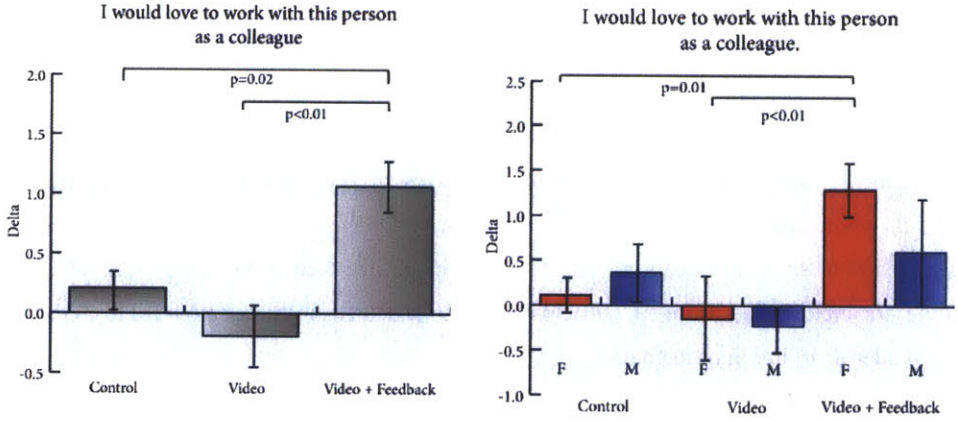


Figure 7-10. Improvement (post – pre) in independent counselor scores in item, "I would love to work with this person as a colleague," across conditions (left) and across conditions, broken down to females (F) and males (M) (right).

In the "*excited about the job*" dimension, participants in Group 3 (video + feedback) were rated to have demonstrated a significant performance increase compared to Group 2 (video) (F[1,83]=5.7371, p=0.02). The analysis that also considered participant gender showed a significant interaction between intervention type and gender (F[2,80]=3.38, p=0.04) for the "*excited about the job*" dimension. Comparisons showed that female participants in Group 3 (video + feedback) were rated higher in the "*excited about the job*" compared to the females in Group 2 (video) (F[1,80]=11.4594, p<0.01), as shown in Figure 7-11.

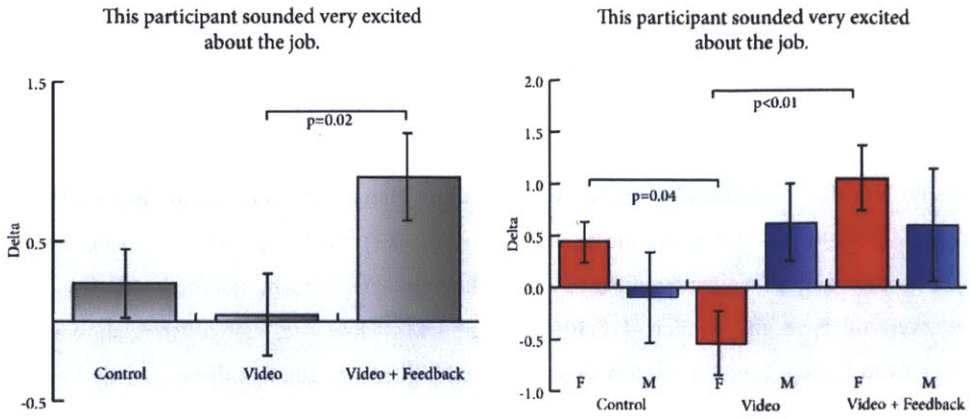


Figure 7-11. Improvement (post – pre) in independent counselor scores in item, "very excited about the job," across conditions (left) and across conditions, broken down to females (F) and males (M) (right).

7.7.4 Independent Judges – Mechanical Turkers

In this section, we report results gathered from the anonymous workers from Mechanical Turk. The Turkers watched interviews of the participants (pre and post) in random order and rated them using a web browser. Of the participants, 21 did not give us permission to share their data beyond the researchers involved in this study. Therefore, we could not include data from 21 participants in the Mechanical Turk study. Table 7-3 shows the gender distribution of the data.

Table 7-3. Distribution of Participants across three conditions that were shared with the Mechanical Turkers

Gender	Control	Video	Video + Feedback
Male	8	8	10
Female	14	14	15
Total	22	22	25

To get a sense of the consistency of the labels produced by the four Turkers, we perform Krippendorff's alpha on the labels from the four Turkers (labels are shown in *Appendix B-2: Questionnaire used in the intervention for the counselors/Turkers.*) The alpha was .82, which could be interpreted as acceptable agreement. We then take the average of the labels produced by the four Turkers and perform statistical analysis on the averaged ratings.

The analysis showed significant changes across intervention types on dimensions of "*overall improvement*" ($F[2,66]=4.01$, $p=0.02$) and "*sounds excited about the job*" ($F[2,66]=3.09$, $p=0.05$).

In "*overall improvement,*" the participants in Group 3 (video + feedback) were rated higher than the participants in Group 2 (video) (F[1,66]=7.07, p=0.01) and control (F[1,66]=4.42, p = 0.04), as shown in Figure 7-12. The analysis that also considered participant gender did not show any significant interaction between intervention type and gender.

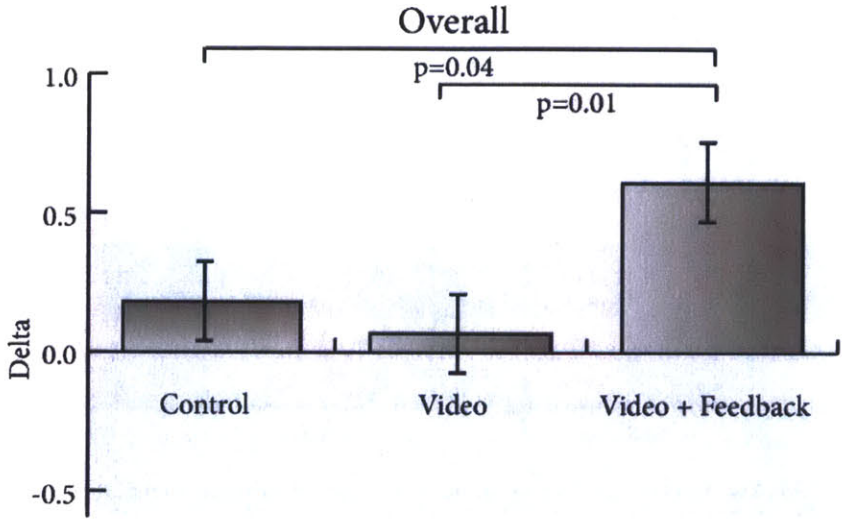


Figure 7-12. Improvement (post – pre) in Mechanical Turkers in item, "*Overall Improvement*" across conditions.

In "*excited about the job,*" the participants in Group 3 (video + feedback) were rated higher than the participants in Group 2 (video) (F[1,66]=5.12, p=0.02), as shown in Figure 7-13.

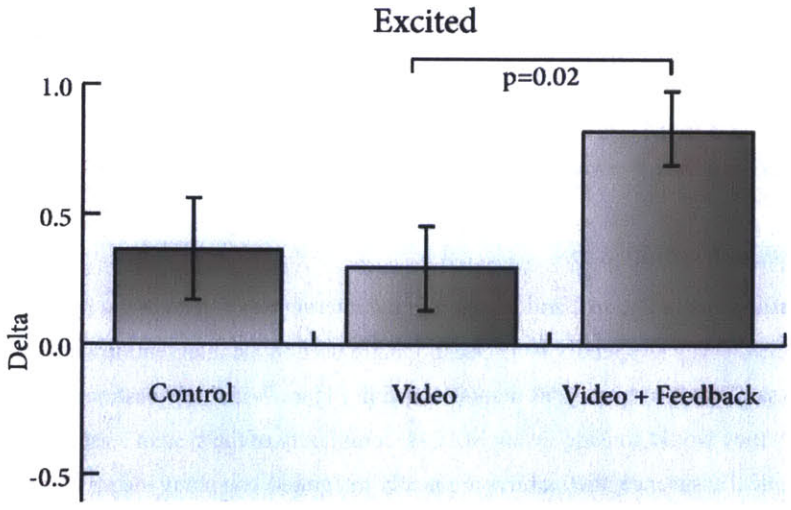


Figure 7-13. Improvement (post – pre) in Mechanical Turkers in item, "*excited about the job*" across conditions.

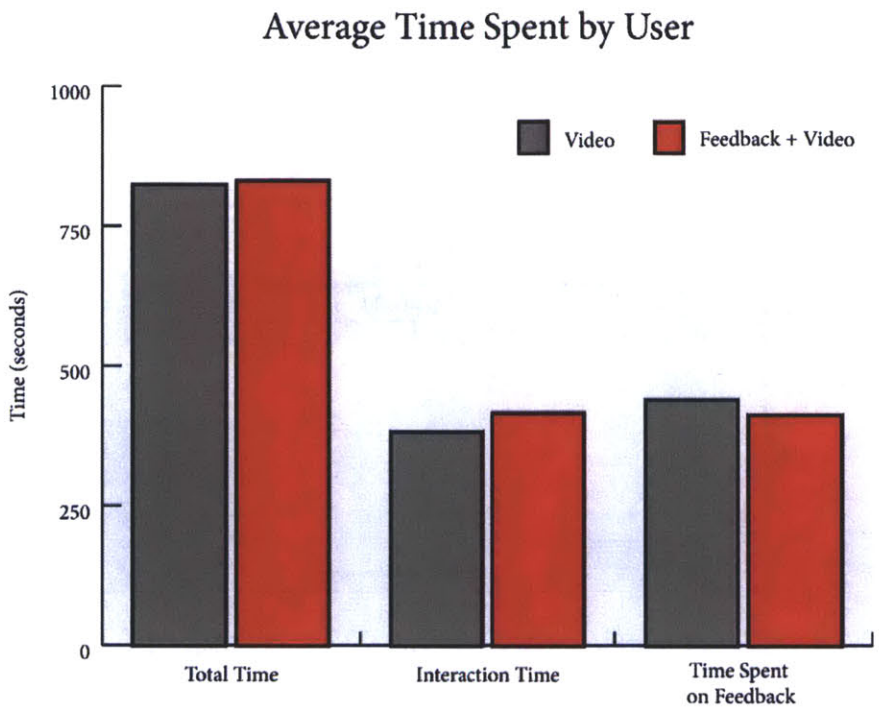


Figure 7-14. Average time spent by participants from Group 2 (video) and Group 3 (video + feedback) during the interaction with MACH, and time spent on looking through the feedback/video.

The analysis that also considered participant gender did not show any significant interaction between intervention type and gender.

7.7.5 Additional Assessments

The participants in Group 2 (video) and Group 3 (video + feedback) were given the option of practicing at least one round of interview with MACH, and could continue to practice up to 3 rounds. The analysis revealed that participants in both Group 2 and 3 did approximately 2 rounds of interviews, on average. The average time spent in interacting with MACH per session is also very similar across the two groups, as shown in Figure 7-14.

System Usability Score:

Participants in Group 2 and Group 3 in the intervention filled out the System Usability Score (SUS) survey (Appendix B-5: Questionnaire used in the intervention for the students on System Usability). The survey, as demonstrated in Figure 7-15, had questions relevant to how frequently they would consider using MACH, complexity of the system, ease of use, may need technical assistance, the features are neatly integrated, too many inconsistencies, quick to learn, feels very confident to use it, has a steep learning curve. The average of SUS

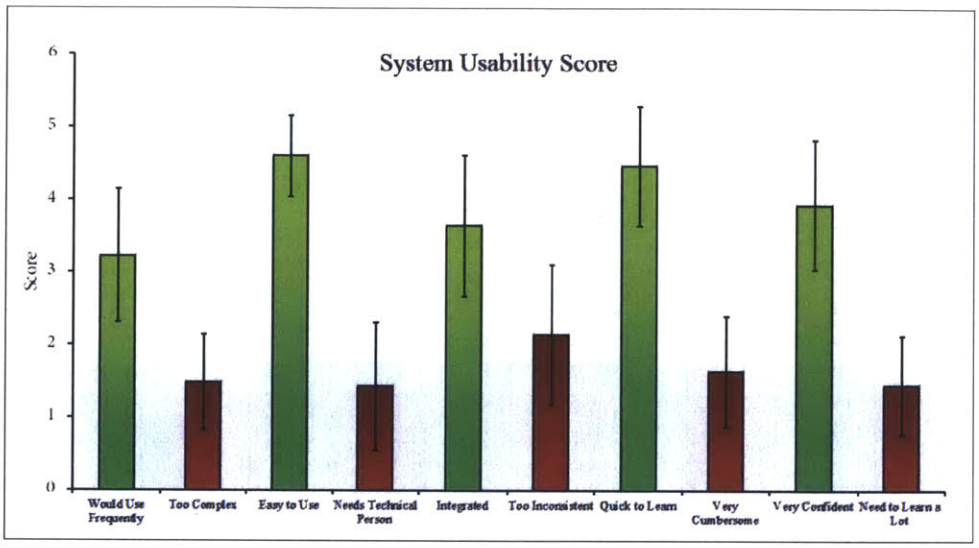


Figure 7-15. The average System Usability Score on MACH from the participants.

scores across all the participants was 80 (SD = 11). The evaluation of each individual dimension is demonstrated in Figure 7-15.

7.7.6 Post Interviews and Surveys with Participants

At the end of the interaction with MACH, participants were given a questionnaire with an opportunity to self-report their opinion on the feedback as well as the interaction. Figure 7-16 demonstrates the comparative usefulness of the nonverbal features that MACH provided to the users. This survey was filled out by the participants in Group 3 (video + feedback) only since they were the only ones who got to see that feedback.

Based on the feedback, feedback on speaking rate, weak language, and smiles were the most important, whereas transcription of the interview was not deemed very helpful.
Participants were specifically asked about the helpfulness of the interaction and feedback. For example, they were asked whether they enjoyed watching their video and whether they would like to use the system again in the future. In addition, the questionnaire also inquired about the appearance and behaviors of the agent. Participants' average scores on those questions are listed in Table 7-4.

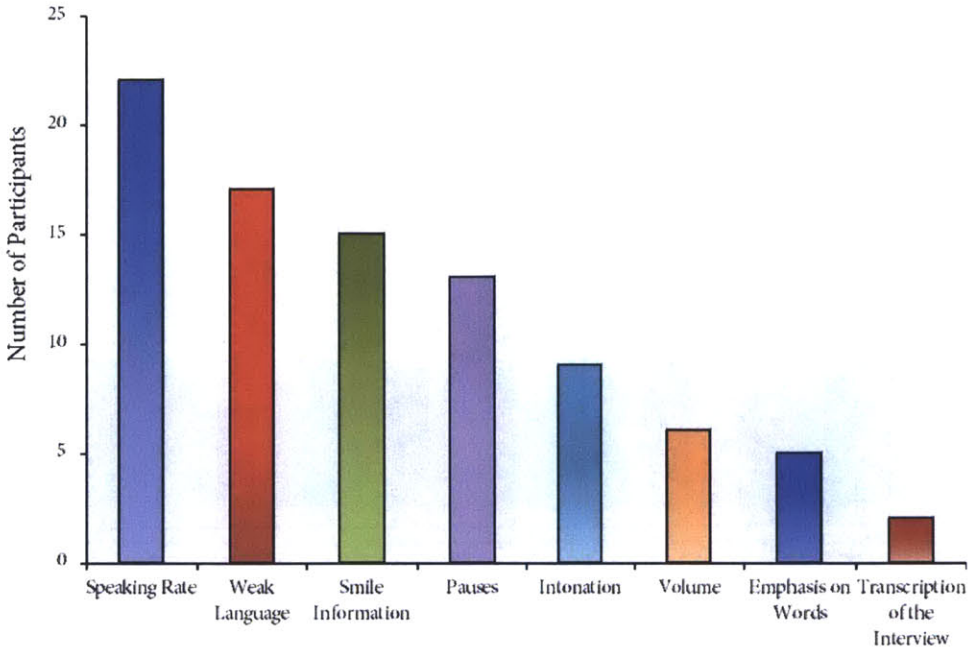


Figure 7-16 Comparative usefulness of feedback components based on the ratings of the participants from Group 3 (video + feedback).

Table 7-4. The summary of responses from the participants from the questionnaire. SD = Standard deviation. 1=Strongly disagree, 7=Strongly agree.

Questions	Average	SD
You found this interview experience useful.	5.76	1
Watching your own video useful.	6.32	.8
You explored new things about yourself through this practice.	5.2	1.2
You would prefer the interviewer to look like a human-like character.	4.2	.7
Interviewer was responding to your cues.	5.12	1.2

Post-Interviews: The paragraphs below describe findings from the participants' subjective evaluations of MACH and the open-ended feedback on their experience. The findings are grouped under themes in which qualitative and quantitative results overlap to represent user experience and system usability.

MACH as a Social Facilitator: The responsiveness of MACH's behavior was rated an average of 5.12 (*SD* = 1.4) by the participants in Groups 2 and 3. Most participants found the character's behavior to be natural.

"It has a lot of nonverbal stuff that you would want her to do. Some of the head tilt, acknowledging the speaking, nodding the head, act like it is listening to you and stuff..."

"I was surprised that it knew when to nod its head, especially when it seemed natural to nod."

"Just the way her eyes moved a little a bit... and after you responded... it seemed as if she was listening to you. I thought that made her kinda humanistic as opposed to a system."

People Accept as Humanlike: Overall, participants rated their preference toward a human-like character over a cartoon-like character at an average of 4.20 (SD = 0.70), suggesting a slight preference toward a human-like character. The excerpts below provide further insight into participant preferences:

"...being here talking to a machine, I felt quite comfortable, which I didn't think I would feel."

"I think the system is adding more value. I think if you were sitting across the table from me and you were recording, or taking notes, I would feel more intimidated. The fact that nobody is there is really helpful".

Self-reflective Feedback is Useful: Most of the participants disliked looking at their video during the intervention in Groups 2 and 3. However, all participants overwhelmingly agreed that watching their video was, while discomfiting, very useful (average rating of 6.3 out of 7, *SD* = 0.8). Additionally, participants rated whether they had learned something new about their behaviors at an average of 5.12 (*SD* = 1.20). This feeling was also reflected in the open-ended feedback from participants:

"I didn't like looking at my video, but I appreciated it."

"I think it is really helpful. You don't really know unless you record yourself. This provides more analysis. The pauses may not appear that long, but when you look at it, you see something else."

Speaking Rate was Most Useful: According to the participants' responses to the questionnaire, speaking rate, weak language (e.g., fillers), smile information, and pauses were the top four attributes of the visual feedback.

MACH is Easy to Use: The average of SUS ratings from participants in Groups 2 and 3 was 80 (*SD* = 11).

Figure 7-17 and Figure 7-18 provide visual illustrations of nonverbal properties of one male participant who went through the video + feedback intervention. Figure 7-17

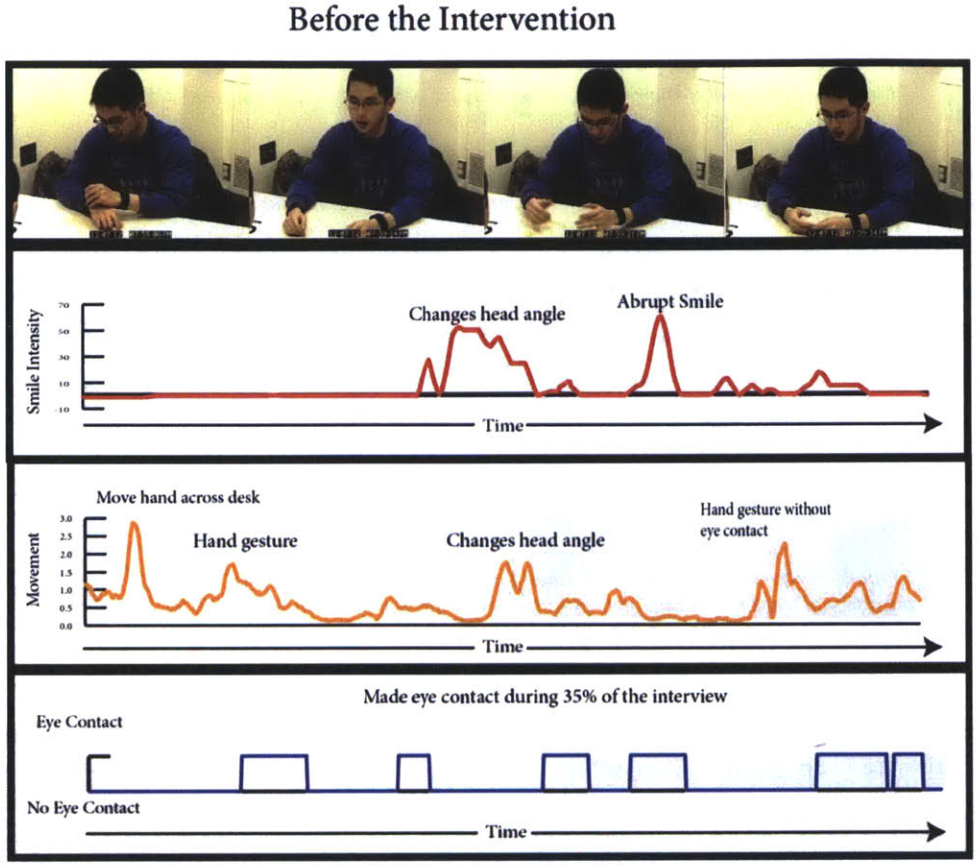


Figure 7-17. A 20 seconds video snippet of an interaction of a male participant before the intervention. Careful observation of the nonverbal data reveals that the participant demonstrates abrupt smile, less eye contact (35%), and lack of coordination between his head gestures and eye contact. Look at the Figure 7-18 to view the changes in his nonverbal behaviors after the intervention.

contains an approximately 20-second segment of the interview before the intervention, and Figure 7-18 contains ~20 seconds of the interview after the intervention. In both cases, the participant was asked the question, "So tell me about yourself."

Along with the thumbnail of the videos, three main nonverbal components (smiles, movement, eye contact) are visualized in Figure 7-17 and Figure 7-18. While the smile and movement were automatically extracted, the information on eye contact was manually coded in every frame for this example.

In Figure 7-17, the participant did not make much eye contact with the counselor. His smiles seemed to be short and abrupt. His head movements and hand gestures do not align well with the other modalities (e.g., eye contact). He made hand gestures while looking away from the counselor. This participant was randomly assigned to Group 3 (video +

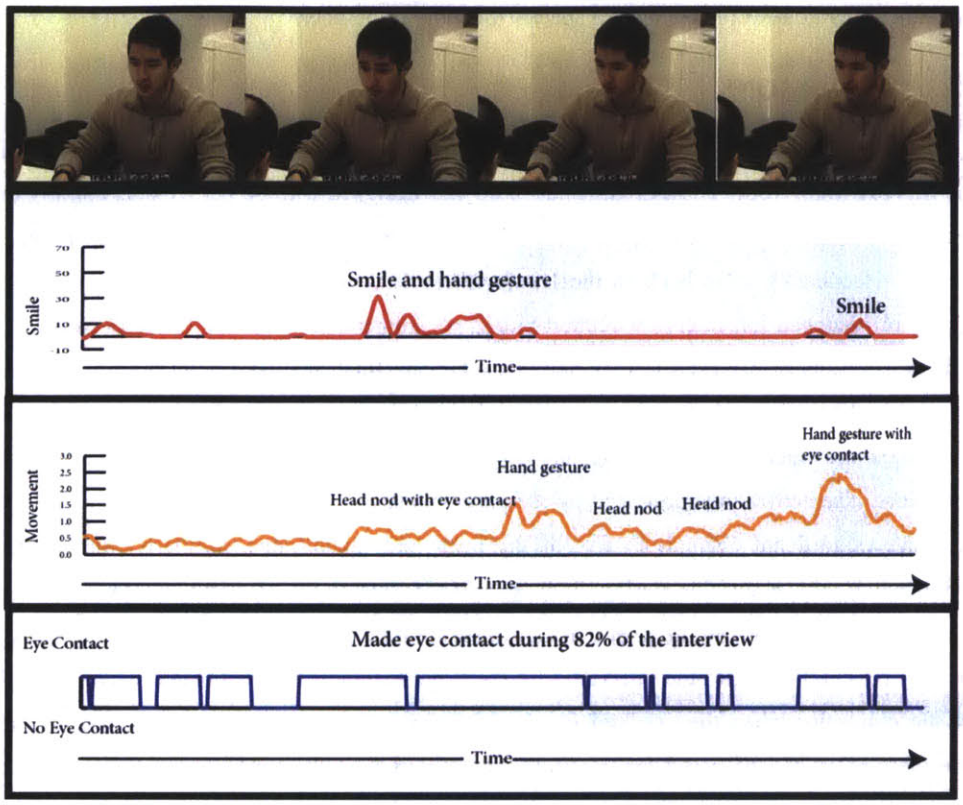


Figure 7-18. An approximately 20-second video snippet of an interaction of a male participant after the intervention. Careful observation reveals that that after the intervention the participant seemed to make much more eye contact with the interviewer. The participant also seemed to coordinate his nonverbal behaviors well. For example, the participant demonstrates a slight smile at the end with hand gesture while maintaining his eye contact.

feedback). He came back to the lab three days after his inter PRE-interview, and interacted with MACH for about an hour. He returned for his final round of interview with the same human counselor after one week. His ratings had significantly gone up during the post interview. Figure 7-18 provides a demonstration of his nonverbal cues from the first 20 seconds of his post-interview, when he was asked to tell more about his background.

In Figure 7-18, the first thing to notice is that his eye contact has significantly increased compared to his pre-interview experience. He did not seem to smile as much in the 20 seconds of segment when he talked about his background. However, he seemed to be effectively (i.e., while making eye contact) using his head movements during the interview. His movements also seemed smooth and in synch with other modalities (e.g., at the end, a slight smile with hand gesture and eye contact).

This example motivates the observation that there is no magic number for how much smiling or eye contact one should have. The most important thing is the coordination of modalities, without worrying about a specific quantity or number.

7.8 DISCUSSIONS

In this chapter, I presented the experimental setup, evaluation strategies, and results of the MACH framework. The experimental setup was designed with 90 participants coming to the lab for a mock interview with a career counselor. Participants in Group 2 (video) and 3 (video + feedback) came back to the lab for the intervention while participants in Group 1(control) watched job interview-related videos recommended by the MIT Career Services. All the participants came back after one week for one final interview with the same career counselor. We analyzed the ratings produced by the participants and the counselor who conducted the interview to observe any possible changes in each participant's nonverbal behaviors. The self-ratings produced by the participants within the span of one week did not yield any significant changes in any of the behavioral dimensions. Similarly, the ratings produced by the counselor during PRE and POST interviews also did not provide any statistical significance with intervention type. One possible explanation might be the design of the experimental setup for the interviews. In the experimental setup, participants were greeted by the counselor when they arrived for the interview. The counselor then led them to the interview room and prepared for the interview to start (i.e., the counselor had to assist the participants in putting the Affectiva Q sensor on). At the end of the interview, the counselor provided them with feedback on the interview performance. Then, participants left the interview room and moved to a different room where they filled out a survey on their performance. In the meantime, the counselor also rated the participants' performance. It is possible that the interaction between the participants and the counselor before and after the interview biased the judgment of participants' self-ratings as well as the ratings of the counselor for the participants. It is likely that the participants were influenced by the counselor's judgments/feedback when they self-rated themselves. Similarly, counselors' ratings might also have been biased by how well the participants had responded to the feedback. For example, some people engaged further with the counselors, provided explanations or interpretations for their performance, and sought further feedback. It is possible that the counselor incorporated the level of engagement and rapport during the feedback while rating the participants. A better design would have been having the participants and the counselors do the ratings right after the interview and then conduct the feedback later. However, initial greetings and pleasantries before the interview might still be a confounding factor.

Even though the overall effects were not significant in the counselors' and participants' ratings, there were significant gender effects. For example, female participants from Group 2 (video) and Group 3 (video + feedback) rated themselves higher in terms of "appearing friendly in the interview" after the intervention, compared to the participants in the control group. It is possible that interacting with the virtual agent and watching the feedback had encouraged them to appear friendlier in the interviews. This could also be attributed to placebo effects of going through the intervention. However, it was interesting to see that only the female students were impacted by this effect, whereas male participants' self-ratings remained unchanged.

Analyzing the ratings of the counselors who conducted the interviews did not reveal any statistically significant results on any possible behavioral dimensions across the intervention types. One thing to consider is that the two sets of interviews took place within a week. Counselors remembered most of the participants when they all came for the POST interview, and possibly were biased by participants' performance in the PRE interview. It is also possible that the counselors were partial towards the participants who were more receptive to the counselors' feedback, resulting in better rapport.

To introduce more objectivity into the labeling of participants' job interviews, two professional career counselors from MIT were requested to volunteer their time by rating the videos. These two career counselors were not part of the experiment, and were blind to the experiment and hypothesis conditions. They viewed the videos using a web browser, in which they were presented the PRE and POST interview videos of the participants in random order, blind to condition. They would still be rating two videos of each participant. But they would not know whether a video was from PRE or POST interview, resulting in uniform bias across conditions. The counselors also had the ability to pause and replay the videos, which might have enabled them to analyze the interviews more thoroughly. In this labeling process, the female MIT career counselor labeled the female participants, and the male MIT career counselor rated the male participants. We opted for that design so that we could compare their labels with the MIT career counselors who conducted and labeled the same set of interviews. Another practical limitation was the unavailability of the independent MIT career counselors. Despite their interest and commitment, they were unable to volunteer their time to label the entire 180 videos individually. Analysis on the independent MIT career counselors revealed that participants in Group 3 (video + feedback) demonstrated statistically significant improvement in dimensions such as "*overall performance*," "*would love to work as a colleague*," and "*excited about the job.*" The analysis considering gender exhibits that the female participants demonstrated statistically significant improvement in dimensions of "*overall improvement*" and "*excited about the job,*" whereas the ratings for male participants on those two dimensions did not change.

Additionally, we recruited four Mechanical Turkers, also blinded to the experiment and video conditions, so that we could do an analysis on their average ratings. Among the 90 participants, 21 participants (11 male and 10 females) did not give permission to share their data online. Therefore, only 69 participants' data were uploaded into the Mechanical Turk website. One could argue that Mechanical Turk is an unusual choice to rate job interview corpus as we don't know anything about the Turker's academic/professional background or any other relevant experience. Therefore, the ratings from the Turkers are more likely to be viewed as "general perception" as opposed to "expert opinion." Given the diverse nature of the background of the Turkers, it is also possible that their labels may contain discrepancies. However, the agreement among the Turkers was measured as .82 using Krippendorff's alpha. An alpha of .82 could be equated to a high level of agreement. Among all the behavioral dimensions, "*overall performance*" and "*excited about the job*" yielded statistically significant differences. Consistent with the ratings from the expert independent MIT career counselors, the Turkers also labeled the participants from Group 3 (video + feedback) to have demonstrated significant improvement compared to the participants from Group 2 (video) and control. Thus, both groups who were blind to the before-after condition associated significant improvement with Group 3 over Group 2 and control.

Analyzing the usage patterns of the participants as they interacted with MACH, we noticed that participants from Group 2 and Group 3 spent an equal amount of time interacting with MACH. Even though participants could have terminated the session after a single interaction, all the participants in Group 2 and 3 interacted with MACH on average 1.66 times during their sessions. Participants rated the human-like appearance of the virtual agent very highly. Most of the participants mentioned that if the virtual agent had been a cartoon; they might not have taken the interview experience very seriously. A majority of the participants felt that the virtual agent was responding to them as they spoke. Having a character that listened and paid attention was deemed useful by most of the participants. The system flow of MACH was very complicated, allowing participants to navigate in many possible ways during their interaction. However, most of the participants rated the usability of the system very highly (average System Usability Score = 80). Participants found feedback on most of the nonverbal variables useful. Speaking rate and smile information were rated to be the most useful features, whereas the transcription of the interview was rated to be the least helpful feedback.

In this intervention, it was not possible to verify whether the novelty effect of interacting with a new system was having an impact or not. However, during the debrief, participants mentioned that they were unlikely to use the system every day. They would be more inclined to use the system before an interview or a presentation as part of the preparation. It remains for future work to run a long-term study and monitor usage

information of users. With deployment of a web version of MACH, this could be tested in the future. Another limitation of this study was that it only lasted for one week and the evaluation was performed on a system level. An extension of this study would be to run a long-term intervention to better understand the changes in nonverbal properties across multiple practice and test sessions.

One interesting phenomenon was that there was a gender effect on the improvement of the behavior. Based on the MIT independent counselors' ratings, the female participants' improvement was statistically significant while the ratings for male participants did not change. One thing to keep in mind is that the number of male and female participants in the intervention was uneven (59% female and 41% male). Hence, it is possible that the group with the majority may dictate the statistical significance.

We uploaded 69 participants (26 males and 43 females) interview to Mechanical Turk, as 21 of the participants did not give us permission to share their data with others. There were 4 Turkers' (genders unknown) who labeled all the interviews in random order. Performing statistical analysis on their average ratings provided very similar findings to the ratings of the MIT Independent counselors. For example, based on the Turkers' ratings, participants in Group 3 (video + feedback) improved significantly in dimensions of "*overall performance*" and "*excited about the job*" compared to the participants in Group 2 and control. However, based on the Turkers' ratings, there were no significant gender effects. There could be two reasons for it. First, there were fewer participants in the Turkers' analysis compared to the analysis that considered MIT career counselors' ratings. Second, all the Turkers labeled both male and female interviewees. Therefore, if they had any gender bias, it would be distributed uniformly across their ratings on all the participants. With MIT career counselors, a female counselor rated the female participants and a male counselor rated the male participants. Give the rating arrangement, we are unable judge whether they had gender bias or not. Having more professional career counselors rate the interviews remains for future explorations.

The results presented on gender effects raises the question of whether there are any differences in how males and females learn and reflect on nonverbal behaviors. Do females learn more quickly from nonverbal behavior feedback than males or is this difference an artifact of other aspects of the study? How does our perception of effective nonverbal behavior get impacted by gender? What would have happened if we did not control for gender effects in this intervention? Would we have had similar results? The results presented in this thesis motivate the further exploration of these research questions.

7.9 CHAPTER SUMMARY

In this chapter, I presented the experimental setup to validate MACH in the context of job interviews. Ninety MIT undergraduate students were recruited and were randomly split into 3 groups: Group 1 (control), Group 2 (video), and Group 3 (video + feedback). All the participants in Groups 1, 2, and 3 interacted with a career counselor in the context of a job interview. Participants in Groups 2 and 3 were brought back to the lab to interact with MACH. All the participants in Groups 1, 2, and 3 came back after one week and did a POST interview with the same counselor. Analyses were presented, based on the ratings by four four groups: the interviewees, interviewers, independent MIT career counselors (who were not part of the experiment), and Mechanical Turkers. Based on the ratings by the independent MIT career counselors who were blind to the hypothesis, participants from Group 3 (video + feedback) demonstrated statistically significant improvement in dimensions such as "*overall performance,*" "*would love to work as a colleague,*" and "*excited about the job,*" whereas ratings for Group 2 (video) and control did not change. These findings were partially confirmed by the labels produced by four Mechanical Turkers where participants in Group 3 were rated to have significant improvement in "*overall performance*" and "*excited about the job*" dimensions. Post interviews with participants revealed that they found the interaction and feedback through MACH very useful, and questionnaire measures indicated that the system was easy to use.

The next chapter summarizes the major contributions of this thesis and provides a detailed discussion offering further explanations of the results reported in this thesis, along with highlighting some remaining challenges.

Chapter 8: Discussion, Conclusions and Future Directions

8.1 SUMMARY OF SIGNIFICANT CONTRIBUTIONS

8.1.1 Theoretical

In this thesis, I have presented computational evidence of humans demonstrating unique signatures of delight while going through negative stimuli, contradicting the popular belief of "true smiles." In addition, I provided new findings on morphological and dynamic properties of polite and amused smiles, confirming and expanding our current understanding of naturalistic expressions. I computationally demonstrated that it is possible for computer algorithms to outperform humans' ability to recognize smiles elicited under negative stimuli.

8.1.2 Methodological

I have designed a new interaction scenario with computers that could provide affective feedback to enhance human nonverbal skills. I demonstrated new ways to automatically visualize human nonverbal data for users to understand, and interpret their behaviors, as part of a computer interface.

8.1.3 Practical

I have implemented a system called "MACH – My Automated Conversation coacH" consisting of a virtual agent that is able to "see," "hear," and "respond" in real-time through a standard webcam. The system was contextualized to job interview scenarios and modeled through 28 mock interviews. Finally, MACH was given to 90 MIT undergraduate students as a means to improve their interview skills. Students who interacted with MACH and received nonverbal feedback were perceived to be better candidates than the students in the control group, judged by the independent MIT career counselors and Mechanical Turkers. The judges were not part of the experiment and were blind to the hypothesis.

8.2 SUMMARY OF KEY DESIGN ISSUES

Recall, from Chapter 6, that the following key questions formed the core evaluation aspects of this thesis.

- How might computer technology use counseling as a metaphor, offering an individual the ability to practice social interaction and to receive useful feedback?
- How would a system automatically sense, interpret, and represent conversational multimodal behavioral data in a format that is both intuitive and educational?

- Will such technology elicit measurable improvements in social skills in participants?

While addressing them, I reflected on a few design issues that are discussed in this section.

8.2.1 Readable Social Cues

MACH was designed to mirror certain aspects of the interaction. For example, it can mirror the participants' smiles and head movement modeled after the behaviors of human interviewers. The mirroring behavior was subtle, yet readable by the participants. The study demonstrated that the participants were less intimidated by an animated character as opposed to a real human, but they appreciated the character being responsive. MACH was programmed to demonstrate different variations of head nods periodically irrespective of the speech content and behavior of the participants. However, in debriefings, participants mentioned that they felt that the character was nodding at the right moments, and they inquired how the system was able to figure it out. Some of them felt that MACH would nod its head whenever they made a point; some felt that it would nod when their voice pitch went up, or they put emphasis on a certain point. In other words, participants were able to incorporate the semi-scripted head nods into their conversations and carry on. This phenomenon is not new. Dating back to 1964, the first AI chat bot, Eliza (Weizenbaum, 1966), was built to provide canned responses based on "keyword" spotting when chatting with a real human. Yet many users were convinced that Eliza actually understood them and asked for private sessions with it. Similar finding on nodding of the head leading to longer story sessions with an agent was also reported in GrandChair (Smith, 2000).

In this study, I have initially set up a mock interview study to better understand the nonverbal behavior of the interviewers. I noticed that interviewers were less expressive during the interview and nodded their heads periodically as part of providing acknowledgement. Therefore, when the virtual counselors were simulated, they were designed to be less expressive with their expressions and backchanneling behavior. However, if we were to try out this interaction in other scenarios (e.g., dating), it would have to be informed by relevant background literature, contextual inquiry and new technology development to model the unique characteristics of the interaction. Perhaps it may be possible to develop a computational theory that models the commonalities of each interaction, so that a new interaction scenario could be an instantiation of the theory. That possibility remains for future explorations.

8.2.2 The Real-time Performance

In face-to-face interactions we yield a lot of cues, often inadvertently, to our interlocutor, and inadvertently respond to them. All of that happens in milliseconds. Any

delay in processing and responding to those cues disrupts human communication. As we start to use realistic human-looking virtual characters, we run the risk of disrupting the interaction with even the slightest of delay in automatically processing and responding to cues. Therefore, one of our design considerations with MACH was to have it respond in real-time to reduce the processing lag as much as possible. With the availability of parallel processes, it was possible to coordinate sensing (e.g., audio and video) and synthesis (behavior, speech, lip synchronization) modules in a standard PC without introducing any significant delay. Even though MACH was less likely to be as fluid and fast as a real human, it was fast enough to self-execute and carry on a job interview.

8.2.3 Seeing is Believing: Coupling Objectivity with Subjectivity

Human interpretation of nonverbal behavior is subjective, while automated sensing of nonverbal behavior is objective. Modeling the human subjectivity with the objective data remains an open problem. In this thesis, I have introduced a design and experimental set-up that allowed the participants to apply human subjectivity to interpret quantitative objective nonverbal data.

Viewing the subtle nonverbal multidimensional behaviors in a graphical format is overwhelming, and very revealing. One of the design challenges was to abstract the high-dimensional numbers into formats so that people could understand and interpret their behavior without any additional instructions. The feedback would need to be objective, yet presented in a way such that each participant could apply his or her subjectivity to interpret it. Also, we allowed the participants to engage in practice interactions with a career counselor between and after the intervention. This enabled the participants to contextualize the automated feedback based on both the interaction experience and expert feedback from the career counselor.

Our post-interview with participants and their anonymous feedback after the session indicated that participants were able to interpret their behaviors during the interaction because it was instantaneous, consistent, and contextualized. The data shown after each session were objective. However, allowing the participants to interact multiple times and compare their data across sessions introduced a level of subjectivity.

8.3 REMAINING CHALLENGES

8.3.1 Sensing More Subtle Expressions

Smiling more during interactions (e.g., smiling at the beginning and ending of an interview) is a good thing. In the intervention developed in this thesis, a participant could view his or her smile track to get a sense of how much they smiled during the interaction. It was useful for people who forget to smile during interactions.

The next step would be to give feedback on the quality of the smile. As part of explorations in technology development, I have shown that people smile in many different scenarios, including even under frustration. I demonstrated that it is possible for computer algorithms to differentiate different types of smiles, in some cases, better than humans. It remains for future work to develop a working system driven by the theoretical framework developed in this thesis. Such a system could be capable of identifying the nuances of smiles, e.g., nervous smiles, embarrassed smiles, polite smiles, and amused smiles, so that users could better understand and reflect on their expressions.

I analyzed the properties of amused and polite smiles when they are shared vs. not shared. In the version of MACH presented in this thesis shared all the smiles. Perhaps, a smarter system of the future would learn to share only the kinds that people shared – probably mostly those of "joy".

Another computer vision challenge that remains unsolved is disambiguating the mouth movement during smiling and just talking. Opening the mouth to speak and opening the mouth to smile sometimes could have very similar patterns (e.g., firing of AU 12 - lip corner pull). It could be difficult for computer algorithms to reliably distinguish between "opening the mouth to speak" and "opening to mouth to smile." In the context of Human-Computer Interaction, when participants are expected to speak and emote at the same time, it is very crucial that the system knows the difference. One possibility is to go multimodal and analyze the acoustic properties of sound to look for signatures of delight. But for more subtle cases, this distinction remains a very challenging problem.

I collected Electrodermal Activity (EDA) data from the participants and the counselors in this study. The data could reveal insightful information about rapport by analyzing physiological signals of the interviewers and the interviewees and how that biases the judgment of the interviewees. This remains for future work.

8.3.2 Understanding the Speech Content

Analyzing the semantic content of the speech was not part of this thesis, and remains for future work. One practical limitation of not considering speech content analysis was the performance of the speech recognition engine. The speech recognition engines have improved significantly over the years, with their speaker-independent models and ability to work in noisy environments. However, they still have a long way to go to incorporate models that can compensate for regional dialects. For example, in the final study with MACH, participation was limited to native speakers of English only. But the speech recognizer still struggled to model the regional accent differences (e.g., the native Bostonian accent did not work very well). ☺

Here is an observation addressing the potential limit of the framework. I have spent numerous hours developing and evaluating the technology that drives MACH. I have a foreign English accent. So the speech recognition engine did not work well on me in the beginning. However, as I continued to test the system, it started to perform really well on me and at some point, it became perfect. Nuance Communications ("Nuance Communications," n.d.) claimed that the system learned my voice parameters and adjusted the recognition engines automatically. While I am gratified to see Nuance's speech recognition learning to adapt to my speaking style, I suspect that I might have also adjusted my speaking style to boost the speech recognition accuracy. It is a valid concern that users might change their natural speaking styles to make the system understand them. This could result in over-adaptation for long-term usages of this technology.

8.3.3 Posture Analysis

Body postures are very important indicators of nonverbal behaviors. Leaning forward or leaning backward, shoulder movement, and open postures contain much useful information about the interaction. With the Microsoft Kinect, it is possible to recognize body postures automatically. The limitation of using Microsoft Kinect is that it would have to be placed a few feet away from the participants for it to reliably track the body skeleton. The sensing becomes more problematic when the participant is seated, making only his or her upper body visible. In addition, it introduces a dependency on extra hardware that a user would have to buy. Given all of those considerations, I did not include posture analysis. However, Microsoft Kinect has come a long way in enabling near sensing, and the form factor could possibly change in the future, making it a ubiquitous platform.

8.3.4 Eye Tracking

In the version of the MACH prototype presented in this thesis, complete eye tracking was not incorporated. The system tried to infer eye movement from the head orientation. As part of future work, it may become possible to recognize where exactly the participant is gazing in pixel accuracy with the latest eye tracking technology. The current state-of-the-art eye trackers are expensive, bulky, and not scalable at this point (ASL sells them for $40,000 including software). But as they become affordable with a convenient form factor, it may become possible to give participants explicit information on how much eye contact they made in coordination with other behaviors. This will strengthen the existing computational approach to understanding the role of eye contact in face-to-face interactions and new ways to provide real-time feedback to people on their eye contact.

Another limitation of the study involving MACH was the position of the web-camera. It would have been ideal if the camera could have been placed at eye level and directly

looking at the participant. However, given the 46" large display, the camera was either placed at the bottom of the display or at the top. Therefore, the video that MACH processed was somewhat slanted, introducing a tilt effect in the sensing. The videos that the participants watched also exhibited the same tilt effect. As a result, it might have been difficult for the users to understand the subtle appropriateness of their eye contact. An analogous example would be the case of video conferencing. During video conference calls via Skype (or Google Hangout), if one wants to give the impression of making eye contact, one would have to look at the camera, not the person in the video. But people are more inclined to look at the video of the other person, which hampers eye contact during video conferences.

8.3.5 Novelty Effect and Long-term Intervention Efficacy

The intervention developed in this thesis was evaluated with the MIT undergraduate students in junior standing. At the time of the intervention, most of them were actively seeking internships for the next summer. Therefore, most of them were highly motivated and eager to participate in the study to improve their interview skills. They were given $50 Amazon gift certificates for their participation in the study, which added an extra level of motivation. In addition, given that MIT is one of the major technical universities, most of the students are excited to try out new technologies. The participants also appreciated and enjoyed watching the quantitative representation of their nonverbal behavior, given their technical background. They were able to interpret the meanings of graphs across sessions without any difficulty or additional explanation. However, challenges remain in how one would go about using other types of interfaces and intervention techniques with other populations or interaction scenarios. For example, to help an underprivileged population with their job interview skills, the intervention would have to be informed by a contextual inquiry and with modification and enhancement of the current technology. Will a different population be as motivated as the MIT students to try out this technology? What elements can we add to ensure motivation across populations? Would the interface need to be more abstract? Would training on how to use the technology be required? As we explore using the ideas and framework introduced in this thesis in different domains, we need to be very careful about how we adapt the technology.

In this thesis, the efficacy of the MACH framework was evaluated on a short-term basis. We were unable to have users use the technology for a long period of time to better understand the long-term effects across many test points. However, most of the participants did mention that for job interviews, they were less likely to use the system every day. They would be inclined to practice more before an actual interview.

The MACH system was displayed in a big 46" 3D TV kindly donated to us by Samsung. The resolution of the TV was very high making the interaction with the 3D

character in MACH appear real and compelling. The participants were left alone in the experiment room to allow privacy as they interacted with MACH. Many would argue that the life-like characteristics and high-resolution graphics on the TV were part of why participants enjoyed the interaction. As this technology scales and becomes available in personal computers, it remains to be tested how the standard resolution and display size affect the motivation and behavior modifications.

8.3.6 Judgement of Interviews

I made the argument in this thesis that by having MIT independent career counselors and Mechanical Turkers', who did not conduct the interviews, rate the videos in a random order, we removed the subjective bias from the labeling process. Therefore, the labels that they had produced were more reliable in judging the nonverbal behavior than the labels of the participants and the counselors who conducted the interviews. However, based on the literature of job interviews, I demonstrated that real interviews are not objective. Nonverbal behaviors play a big role in perceiving the competency of the candidate. Some aspects of those nonverbal behaviors are only present during face-to-face interactions. For example, a firm hand shake, occupying space, making eye contact, or sharing laughter are only possible in face-to-face interactions. It remains an empirical question whether the ratings of the independent labelers would have changed if they were able to incorporate the nonverbal behaviors only available in face-to-face interactions.

8.3.7 Automated Recommendation

While nonverbal behaviors are very objective, the human interpretation of those behaviors could be subjective and ambiguous. Therefore, in this thesis, I took the position of not making an absolute decision of the interview performance of the interviewees. Instead, I focused on displaying the nonverbal behavior to the participants in an intuitive format. As part of future work, it may be possible to give users feedback by comparing their nonverbal behavior with other user data. For example, instead of getting recommendations, users may get to compare themselves with other users who were rated strongly by judges. It may also be possible for users to share their data with other people through Mechanical Turk and receive human interpretation of computer generated objective feedback.

8.4 CONCLUSION

For the general users, MACH—the new social-emotional technology is not only a great coach, but also a fun and engaging experience. But to the scientific community, it holds much greater promise and meaning. Since a social coach can generate social cues and record fine grained continuous measurement autonomously, this coach can be deployed outside of the lab

or school, effectively increasing both the quantity and quality of data. The coach provides a repeatable, standardized stimulus and recording mechanism, removing subjective bias. Some aspects of recognition, production, and appearance of the social companion can be decomposed arbitrarily, turning off some modules (e.g., speech, eye-contact, head movements), while making others integral. This allows selectively probing responses to individual interaction variables, sometimes in combination that cannot be performed so easily by humans. The most exciting part of this technology is that it puts the users into the driving seat by allowing them to control the pace, difficulty, and quality of the interaction, which is very challenging to simulate with traditional coaching or therapies.

I believe that using MACH for job interviews is a precursor to many other social interaction scenario possibilities. MACH is only an autonomous character today that can train humans on nonverbal skills in a job interview through affective feedback. We spent two years designing, developing and evaluating the initial ideas. Now we are in a unique position to try out many other application areas including helping people with public speaking, social difficulties, language learning, customer service training, or even dating.

The immediate future work includes improving upon the ubiquitous nature of the use of MACH by extending the implementation to mobile platforms and settings as well as providing users with the ability to seamlessly distribute their repeated use of the system to different platforms over time. Additionally, I wish to provide users with the ability to compare their performance to their past performance through progress charts as well as to that of users with specific characteristics such as educational background, geographic region, and job experience. Perhaps, it is possible to create a sort of social network, based entirely on the face-to-face aspect of being social — a next-level LinkedIn — in which users can track their own communication progress and compare it against that of other users based on factors like education, work history, and location.

Social skills are very important aspects of our everyday life. Developing technology with the aim of providing ubiquitous access to social skills training not only introduces new scientific challenges, but also eventually contributes towards a better world.

Bibliography

Acoach, C. L., & Webb, L. M. (2004). The Influence of Language Brokering on Hispanic Teenagers' Acculturation, Academic Performance, and Nonverbal Decoding Skills: A Preliminary Study. *Howard Journal of Communications*, *15*(1), 1–19.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources And Evaluation*, *41*(3-4), 273–287. doi:10.1007/s10579-007-9061-5

Amalfitano, J. G., & Kalt, N. C. (1977). Effects of eye contact on the evaluation of job applicant. *Journal of Employment Counseling*, *14*, 46–48.

Ambadar, Z., Cohn, J. F., & Reed, L. I. (2009). All Smiles are Not Created Equal: Morphology and Timing of Smiles Perceived as Amused, Polite, and Embarrassed/Nervous. *Journal of Nonverbal Behavior*, *33*(1), 17–34. doi:10.1007/s10919-008-0059-5

Ananthakrishnan, S., & Narayanan, S. (2005). An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In *Proc. Int. Conf. Acoust. Speech Signal Process.* (pp. 268–272).

Ananthakrishnan, S., & Narayanan, S. S. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. , 16 Ieee Transactions On Audio Speech And Language Processing 216–228 (2008). NIH Public Access. doi:10.1109/TASL.2007.907570

Andersen, P. A., & Sull, K. K. (1985). Out of touch, out of reach: Tactile predispositions as predictors of interpersonal distance. *The Western Journal of Speech Communication*, *49*, 57–72.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions On Audio Speech And Language Processing*, *20*(2), 356–370. doi:10.1109/TASL.2011.2125954

Antonijevic, S. (2008). From Text to Gesture Online: A microethnographic analysis of nonverbal communication in the Second Life virtual environment. *Information Communication Society*, *11*(2), 221–238. doi:10.1080/13691180801937290

Arena. (n.d.). Retrieved July 23, 2013, from http://www.naturalpoint.com/optitrack/products/arena/)

Argyle, M. (1969). *Social Interaction* (p. 504). Transaction Publishers.

Attwood, T., & Lorna, W. (1998). *Asperger's Syndrome: A Guide for Parents and Professionals* (1st ed.). Jessica Kingsley Publishers.

Baron-Cohen, S. (2004). *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2247465/

Bartlett, M. S., Littlewort, G. C., Frank, M. G., Lainscsek, C., Fasel, I. R., & Movellan, J. R. (2006). Automatic Recognition of Facial Actions in Spontaneous Expressions. *Journal of Multimedia*, *1*(6), 22–35. doi:10.4304/jmm.1.6.22-35

Bass, J. D., & Mulick, J. A. (2007). Social play skill enhancement of children with autism using peers and siblings as therapists. *Psychology*, *44*(7), 727–736. doi:10.1002/pits

Beard, C. (2011). Cognitive bias modification for anxiety: current evidence and future directions. *Expert Review of Neurotherapeutics*, *11*(2), 299–311. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3092585&tool=pmcentrez&rendertype=abstract

Bellini, S., Akullian, J., & Hopf, A. (2007). Increasing Social Engagement in Young Children With Autism Spectrum Disorders Using Video Self-Modeling. *School Psychology Review*, *36*(1), 80–90. Retrieved from http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ788306

Bernard-Opitz, V., Sriram, N., & Nakhoda-Sapuan, S. (2001). Enhancing social problem solving in children with autism and normal children through computer-assisted instruction. *Journal of Autism and Developmental Disorders*, *31*(4), 377–384. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11569584

Billard, A., & Dautenhahn, K. (1999). Experiments in Learning by Imitation - Grounding and Use of Communication in Robotic Agents. *Adaptive Behavior*, *7*, 415–438.

Birdwhistell, R. (1970). *Kinetics in context*. Philadelphia: University of Pennsylvania Press.

Blackman, M. C. (2002). The Employment Interview via the Telephone: Are We Sacrificing Accurate Personality Judgments for Cost Efficiency? *Journal of Research in Personality*, *36*(3), 208–223. doi:10.1006/jrpe.2001.2347

Blanch, D. C., Hall, J. A., Roter, D. L., & Frankel, R. M. (2009). Is it good to express uncertainty to a patient? Correlates and consequences for medical students in a standardized patient visit. *Patient Education and Counseling*, *76*(3), 300–306. Retrieved from http://dx.doi.org/10.1016/j.pec.2009.06.002

Blue Planet Public Speaking. (n.d.). Retrieved November 06, 2013, from http://www.blueplanet.org/

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*, 97–110. Retrieved from http://xeds.eu/other/P_Boersma_Accurate_short-term_analysis_of_the_fundametnal_freq.pdf

Boersma, P., & Weenink, D. (n.d.). Praat: doing phonetics by computer [Computer program]. Retrieved June 21, 2013, from www.praat.org

Bolinger, D. (1989). *Intonation and Its Uses: Melody in Grammar and Discourse* (p. 470). Stanford University Press.

Botinis, A. (2001). Developments and paradigms in intonation research. *Speech Communication*, *33*(4), 263–296. doi:10.1016/S0167-6393(00)00060-1

Boucsein, W. (1992). *Electrodermal activity*. (I. Martin & P. H. Venables, Eds.)*Techniques in psychophysiology* (Vol. 3, pp. 3–67). Plenum Press. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Electrodermal+Activity#0

Brooke, J. (1996). SUS - A quick and dirty usability scale. (P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland, Eds.)*Usability evaluation in industry*, 189–194. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:SUS+-+A+quick+and+dirty+usability+scale#0

Brown, J. L., Krantz, P. J., McClannahan, L. E., & Poulson, C. L. (2008). Using script fading to promote natural environment stimulus control of verbal interactions among youths with autism. *Research in Autism Spectrum Disorders*, *2*, 480–497.

Buehlman, K. T., Gottman, J. M., & Katz, L. F. (1992). How a couple views their past predicts their future: Predicting divorce from an oral history interview. *Journal of Family Psychology*, *5*(3-4), 295–318. doi:10.1037/0893-3200.5.3-4.295

Buggey, T. (2005). Video Self-Modeling Applications With Students With Autism Spectrum Disorder in a Small Private School Setting. *Focus on Autism and Other Developmental Disabilities*, *20*(1), 52–63. Retrieved from http://foa.sagepub.com/cgi/doi/10.1177/10883576050200010501

Buhmann, M. D. (2001). A new class of radial basis functions with compact support. *Mathematics of Computation*, *70*(233), 307–318.

Burgoon, J. K. (1985). Nonverbal signals. In *Handbook of interpersonal communication* (pp. 344–390). Beverly Hills: Sage.

Burgoon, Judee K, & Buller, D. B. (1994). Interpersonal deception: III. Effects of deceit on perceived communication and nonverbal behavior dynamics. *Journal of Nonverbal Behavior*, *18*(2), 155–184. doi:10.1007/BF02170076

Buss, A. (1989). Personality as traits. *American Psychologist*, *44*, 1378–1388.

Butovskaya, M. L., Timentschik, V. M., & Burkova, V. N. (2007). Aggression, conflict resolution, popularity, and attitude to school in Russian adolescents. *Aggressive Behavior*, *33*(2), 170–183. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ab.20197/abstract

Campion, M. A., Campion, J. E., & Hudson, J. P. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology*, *79*(6), 998–1102. doi:10.1037/0021-9010.79.6.998

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel psychology*, *46*(4), 823–847.

Carney, D. R., Hall, J. A., & LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, *29*(2), 105–123. doi:10.1007/s10919-005-2743-z

Carrère, S., Buehlman, K. T., Gottman, J. M., Coan, J. A., & Ruckstuhl, L. (2000). Predicting marital stability and divorce in newlywed couples. *Journal of family psychology JFP journal of the Division of Family Psychology of the American Psychological Association Division 43*, *14*(1), 42–58. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10740681

Cartreine, J. A., Ahern, D. K., & Locke, S. E. (2010). A roadmap to computer-based psychotherapy in the United States. *Harvard Review of Psychiatry*, *18*(2), 80–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20235773

Cassell, J. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, *14*(1-2), 55–64. doi:10.1016/S0950-7051(00)00102-7

Cassell, J, Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). Embodiment in Conversational Interfaces: Rea. In *CHI 99 Proceedings of the SIGCHI conference on Human factors in computing systems* (Vol. 7, pp. 520–527). ACM Press. doi:10.1145/302979.303150

Cassell, Justine. (2000). Embodied Conversational Agents. (Justine Cassell, J. Sullivan, S. Prevost, & E. Churchill, Eds.)*Social Psychology*, *40*(1), 26–36. doi:10.1027/1864-9335.40.1.26

Cassell, Justine. (2001). Embodied conversational agents: representation and intelligence in user interfaces. *AI Magazine*, *22*(4), 67–84. doi:10.1027/1864-9335.40.1.26

Cassell, Justine, & Tartaro, A. (2007). Intersubjectivity in human–agent interaction. *Interaction Studies*, *3*, 391–410. Retrieved from http://www.ingentaconnect.com/content/jbp/is/2007/00000008/00000003/art00003

Cassell, Justine, & Thórisson, K. R. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, *13*(4-5), 519–538. doi:10.1080/088395199117360

CereProc. (n.d.). Retrieved June 28, 2013, from http://www.cereproc.com/

Chang, C., & Lin, C. (2001). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* , *2*(3), 1–30. doi:10.1145/1961189.1961199

Chen, K., Hasegawa-Johnson, M., & Cohen, A. (2004). An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. In *Proc. Int. Conf. Acoust. Speech Signal Process.* (pp. 509–512).

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., … Haehnel, D. The Mobile Sensing Platform: An Embedded Activity Recognition System. , 7 Ieee Pervasive Computing 32–41 (2008). IEEE Computer Society. doi:10.1109/MPRV.2008.39

Clark, D. M. (2001). A cognitive perspective on social phobia. In W. R. Crozier & L. E. Alden (Eds.), *International Handbook of Social Anxiety Concepts Research and Interventions Relating to the Self and Shyness* (Vol. 42, pp. 405–430). John Wiley & Sons Ltd. doi:10.1016/S0006-3223(97)87445-8

Cockburn, J., Bartlett, M., Tanaka, J., Movellan, J., Pierce, M., & Schultz, R. (2008). SmileMaze: A Tutoring System in Real-Time Facial Expression Perception and Production in Children with Autism Spectrum Disorder. In *Proceedings from the IEEE International Conference on Automatic Face & Gesture Recognition.*

Cogger, J. W. (1982). Are you a skilled interviewer? *The Personnel journal, 61*(11), 840–843.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding, 91*(1-2), 160–187. doi:10.1016/S1077-3142(03)00081-X

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets Multiresolution and Information Processing, 2*(2), 57–72.

Congreve, W., & Gifford, R. (2006). Personality and Nonverbal Behavior: A Complex Conundrum. In V. Manusov & M. L. Patterson (Eds.), *The SAGE Handbook of Nonverbal Communication* (pp. 159–181). Thousand Oaks, CA: SAGE Publications.

Consolvo, S., Mcdonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., ... Landay, J. A. (2008). Activity sensing in the wild: a field trial of ubifit garden. In *CHI '08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vol. 08, pp. 1797–1806). New York: ACM. doi:10.1145/1357054.1357335

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*(5), 565–579. doi:10.1037/0021-9010.80.5.565

Cortes, C., & Vapnik, V. (1995). Support-vector networks. (L. Saitta, Ed.)*Machine Learning, 20*(3), 273–297. doi:10.1007/BF00994018

Courgeon, M., Buisine, S., & Martin, J. (2009). Impact of Expressive Wrinkles on Perception of a Virtual Character's Facial Expressions of Emotions. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents* (pp. 201–214). Amsterdam, The Netherland: Springer Verlag. doi:10.1007/978-3-642-04380-2_24

Craske, M. G. (1999). *Anxiety Disorders: Psychological Approaches to Theory and Treatment* (p. 425). Boulder: Westview Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi:10.1007/BF02310555

Cuijpers, P., Marks, I. M., Van Straten, A., Cavanagh, K., Gega, L., & Andersson, G. (2009). Computer-aided psychotherapy for anxiety disorders: a meta-analytic review. *Cognitive Behaviour Therapy, 38*(2), 66–82. Retrieved from http://dx.doi.org/10.1080/16506070802694776

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: a literature review. *Language and Speech, 40 (Pt 2)*(2), 141–201. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9509577

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19363178

DeGroot, T., & Gooty, J. (2009). Can Nonverbal Cues be Used to Make Meaningful Personality Attributions in Employment Interviews? *Journal of Business and Psychology*, *24*(2), 179–192. doi:10.1007/s10869-009-9098-0

DiSalvo, C. A., & Oswald, D. P. (2002). Peer-Mediated Interventions to Increase the Social Interaction of Children With Autism: Consideration of Peer Expectancies. *Focus on Autism and Other Developmental Disabilities*, *17*(4), 198–207. Retrieved from http://foa.sagepub.com/cgi/doi/10.1177/10883576020170040201

Dixit, A. K., & Skeath, S. (2004). *Games of Strategy. Construction* (p. 665). W. W. Norton & Company. Retrieved from http://www.econ.tuwien.ac.at/hanappi/Lehre/GameTheory/2010/Dixit_1999a.pdf

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying facial actions. (D. Touretzky, M. Mozer, & M. Hasselmo, Eds.)*IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(10), 974–989. doi:10.1109/34.799905

Duncan Jr. Starkey. (1969). Nonverbal communication. *Psychological Bulletin*, *72*, 118–137.

Edinger, J. A., & Patterson, M. L. (1983). Nonverbal involvement and social control. *Psychological Bulletin*, *93*(1), 30–56. doi:10.1037//0033-2909.93.1.30

Ekman, P. (1982). *Emotion in the human face*. (P. Ekman, Ed.)*Emotion in the human face* (p. 464). Cambridge University Press.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press*. Palo Alto.

El Kaliouby, R. (2005). *Mind-reading machines: automated inference of complex mental states. Bell Laboratories memorandum* (p. 185). Retrieved from http://www.mendeley.com/research/mindreading-machines-automated-inference-of-complex-mental-states/

El Kaliouby, R., & Robinson, P. (2004). Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *In the IEEE International Workshop on Real Time Computer Vision for Human Computer Interaction, CVPR*. Ieee. doi:10.1109/CVPR.2004.427

FaceTracker. Facial Feature Tracking SDK. (2002). Neven Vision.

Feldman, R. S., Philippot, P., & Custrini, R. J. (1991a). Social competence and nonverbal behavior. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior Studies in emotion and social interaction* (pp. 329–350). Cambridge University Press.

Feldman, R. S., Philippot, P., & Custrini, R. J. (1991b). Social competence and nonverbal behavior. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior Studies in emotion and social interaction* (pp. 329–350). Cambridge University Press.

Ferris, G. L., Witt, L. A., & Hochwarter, W. A. (2001). Interaction of social skill and general mental ability on job performance and salary. *Journal of Applied Psychology*, *86*(6), 1075–1082. doi:10.1037//0021-9010.86.6.1075

Fogg, B. J. (2003). *Persuasive Technology: Using Computers to Change What We Think and Do*. (Y. Kort, W. IJsselsteijn, C. Midden, B. Eggen, & B. J. Fogg, Eds.)*Persuasive Technology Using Computers to Change What We Think and Do* (Vol. 5, p. 283). Morgan Kaufmann. doi:10.4017/gt.2006.05.01.009.00

Fridlund, A. J. (1991). Sociality of Solitary Smiling. *Potentiation by an Implicit Audience*, *60*(2), 229-240.

Friedland, G., Hung, H., & Yeo, C. Y. C. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. , IEEE International Conference on Acoustics Speech and Signal Processing (2009) 4069–4072 (2009). Ieee. doi:10.1109/ICASSP.2009.4960522

Froba, B., & Ernst, A. (2004). Face detection with the modified census transform. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 91–96). IEEE. doi:10.1109/AFGR.2004.1301514

Furui, S. (2005). 50 years of progress in speech and speaker recognition. *Journal of the Acoustical Society of America*, *116*(4), 2497–2498. Retrieved from http://aprodeus.narod.ru/Information/Rech/Furui_SPCOM05.pdf

Ganz, J. B., Kaylor, M., Bourgeois, B., & Hadden, K. (2008). The Impact of Social Scripts and Visual Cues on Verbal Communication in Three Children With Autism Spectrum Disorders. *Focus on Autism and Other Developmental Disabilities*, *23*, 79–94.

Gårding, E. (1979). Sentence intonation in Swedish. *Phonetica*, *36*(3), 207–215. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=523515

Gårding, E. (1982). Swedish prosody. Summary of a project. *Phonetica*, *39*(4-5), 288–301. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7156209

Gifford, R., Ng, C. F., & Wilkinson, M. (1985). Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology*, *70*(4), 729–736. doi:10.1037//0021-9010.70.4.729

Gottman, J. (2003). *The Mathematics of Marriage*. MIT Press.

Gottman, J., Markman, H., & Notarius, C. (1977). The Topography of Marital Conflict. *A Sequential Analysis of Verbal and Nonverbal Behavior*, *39*(3), 461–477. doi:10.2307/350902

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., Werf, R. J. Van Der, & Morency, L. (2006). Virtual Rapport. (J. Gratch, M. Young, R. Aylett, D. Ballin, & P. Olivier, Eds.)*Intelligent Virtual Agents*, *4133*, 14 – 27. doi:10.1007/11821830_2

Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating Rapport with Virtual Agents. (C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé, Eds.)*Intelligent Virtual Agents*, *4722*, 125–138. Retrieved from http://www.springerlink.com/index/X568357400058UM7.pdf

Grynszpan, O., Martin, & Nadel, J. (2005). Human Computer Interfaces for Autism : Assessing the Influence of Task Assignment and Output Modalities. *Education*, 1419–1422. doi:10.1145/1056808.1056931

Gunes, H., & Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions*, *1*(1), 68–99. doi:10.4018/jse.2010101605

Halberstadt, A. G., & Hall, J. A. (1980). Who's getting the message? Children's nonverbal skill and their evaluation by teachers. *Developmental Psychology*, *16*(564-573), 564–573. doi:10.1037/0012-1649.16.6.564

Hall, J. A., & Bernieri, F. J. (Eds.). (2001a). *Interpersonal Sensitivity: Theory and Measurement* (1st ed.). Lawrence Erlbaum Associates.

Hall, J. A., & Bernieri, F. J. (Eds.). (2001b). *Interpersonal Sensitivity: Theory and Measurement* (1st ed.). Lawrence Erlbaum Associates.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1).

Harper, R. G., Wiens, A. N., & Matarazzo, J. D. (1978). *Nonverbal Communication: The State of Art*. New York: John Wiley & Sons.

Harrison, R. P. (n.d.). Nonverbal communication. In *Handbook of communication* (pp. 93–115). Chicago: Rand McNally.

Hartigan, D. B. (2011). *Towards more accurate recognition of patient emotion cues: meta-analysis of training literature and development of an assessment tool and multicomponent intervention for clinicians*. Northeastern University.

Hastings, M. E., Tangney, J. P., & Stuewig, J. (2008). Psychopathy and identification of facial expressions of emotion. *Personality and Individual Differences*, *44*(7), 1474–1483. doi:10.1016/j.paid.2008.01.004

Hernandez, J., Hoque, E., Drevo, W., & Picard, R. W. (2012). Mood meter: counting smiles in the wild. In *UbiComp '12 Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 301 – 310). New York: ACM.

Hilton, M. F., Scuffham, P. A., Sheridan, J., Cleary, C. M., & Whiteford, H. A. (2008). Mental ill-health and the differential effect of employee type on absenteeism and presenteeism. *Journal of occupational and environmental medicine American College of Occupational and Environmental Medicine*, *50*(11), 1228–1243.

HireVue. (n.d.). Retrieved June 11, 2013, from http://new.hirevue.com/

Hirst, D., & Di Cristo, A. (1984). French Intonation: A Parametric Approach. *Die Neueren Sprachen*, *83*(5), 554–569.

Hirst, D., & Di Cristo, A. (1998). A Survey of Intonation Systems. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems A Survey of Twenty Languages* (pp. 1–44). Cambridge University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=LClvNiI4k0sC&pgis=1

Hoey, J. (2004). *Decision Theoretic Learning of Human Facial Displays and Gestures*. University of British Columbia.

Hollandsworth, J. G. . J., Kazelskis, R., Stevens, J., & Dressel, M. E. (1979). Relative Contributions of Verbal, Articulative, and Nonverbal Communication to Employment Decisions in the Job Interview Setting. *Personnel Psychology*, *32*(2), 359–67.

Hollandsworth, J. G., Glazeski, R. C., & Dressel, M. E. (1978). Use of social-skills training in the treatment of extreme anxiety and deficient verbal skills in the job-interview setting. *Journal of Applied Behavior Analysis*, *11*(2), 259–269.

Hoque, M. E., Courgeon, M., Martin, J., Mutlu, B., & Picard, R. W. (2013). MACH: My Automated Conversation coacH. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013)*. Zurich, CH.

Hoque, M. E., Lane, J. K., Kaliouby, R., Goodwin, M., & Picard, R. W. (2009). Exploring Speech Therapy Games with Children on the Autism Spectrum. *10th Annual Conference of the International Speech Communication Association, Interspeech*, 2–5.

Hoque, M. E., & Picard, R. W. (2011). Acted vs. natural frustration and delight: Many people smile in natural frustration. In *Face and Gesture 2011* (Vol. 17, pp. 354–359). IEEE. doi:10.1109/FG.2011.5771425

Hoque, M. E., Yeasin, M., & Louwerse, M. M. (2006). Robust Recognition of Emotion from Speech, 42–53.

Hoque, M., Mcduff, D. J., Morency, L., & Picard, R. W. (2011). Machine Learning for Affective Computing. (S. D'Mello, A. Graesser, B. Schuller, & J.-C. Martin, Eds.)*Media*, *6975*, 6975. Retrieved from http://dl.acm.org/citation.cfm?id=2062927

Howlin, P. (2000). Outcome in Adult Life for more Able Individuals with Autism or Asperger Syndrome. *Autism*, *4*(1), 63–83. doi:10.1177/1362361300004001005

Huang, L., Morency, L.-P., & Gratch, J. (2011). Virtual Rapport 2.0. In *IVA'11 Proceedings of the 10th international conference on Intelligent virtual agents* (pp. 68–79).

Huffcutt, A I, Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897–913. Retrieved from http://www.apa.org

Huffcutt, Allen I, Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *The Journal of applied psychology*, *86*(5), 897–913.

Huffcutt, Allen I, & Winfred, A. J. (1994). Hunter and Hunter (1984) revisited :Interview validity for entry level jobs. *Journal of Applied Psychology*, *79*(2), 184–190. doi:10.1037/0021-9010.79.2.184

I T, J. (2002). *Principal Component Analysis. Journal of the American Statistical Association* (2nd ed., Vol. 98, p. 487). Springer Series in Statistics. Retrieved from http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4

Ivan, L., & Duduciuc, A. (2011). Social skills, nonverbal sensitivity and academic success. The key role of centrality in student networks for higher grades achievement. *Revista de Cercetare si Interventie Sociala, 33*, 151–166.

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., ... Wrede, B. (2004). The ICSI meeting project: Resources and research. In *Proc. Int. Conf. Acoust. Speech Signal Process.*

Jiang, B., Valstar, M., Martinex, B., & Pantic, M. (2011). A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modelling. *Journal of Latex Class Files, 6*(1).

JobOn. (n.d.). Retrieved July 11, 2013, from jobon.com

Johnson, W. L., Vilhjalmsson, H., & Marsella, S. (2009). Serious Games for Language Learning : How Much Game , How Much AI ? (C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker, Eds.)*Information Sciences, 23*(4), 12–15. doi:10.1108/14777280910970738

Juslin, P., & Scherer, K. (2005). Vocal expression of affect. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65–135). Oxford: Oxford University Press. Retrieved from http://elib.fk.uwks.ac.id/asset/archieve/e-book/PSYCHIATRIC- ILMU PENYAKIT JIWA/The New Handbook of Methods in Nonverbal Behavior Research.pdf#page=76

Kanade, T., Cohn, J. F., & Tian, Y. T. Y. (2000). Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 46–53). IEEE Comput. Soc. doi:10.1109/AFGR.2000.840611

Kanner, L. (1943). Autistic disturbances of affective contact. (V. H. Winston, Ed.)*Nervous Child, 2*(2), 217–250. Retrieved from http://affect.media.mit.edu/Rgrads/Articles/pdfs/Kanner-1943-OrigPaper.pdf

Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., & Narayanan, S. S. (2011). SailAlign: Robust long speech-text alignment. In *Proc of Workshop on New Tools and Methods for VeryLarge Scale Phonetics Research.*

Kawato, S., & Ohya, J. (2000). Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes." In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 40–45). IEEE Comput. Soc. doi:10.1109/AFGR.2000.840610

Keltner, D., & Ekman, P. (2000). *Facial expression of emotion* (pp. 236–249). New York: Guilford Press.

Kenny, P., Parsons, T. D., Gratch, J., Leuski, A., & Rizzo, A. A. (2007). Virtual Patients for Clinical Therapist Skills Training. (C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé, Eds.)*IVA 07 Proceedings of the 7th international conference on Intelligent Virtual Agents, 4722*, 197–210. doi:10.1007/978-3-540-74997-4

Kim, J., Hill, R. W., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., ... Hart, J. (2009). BiLAT : A Game-Based Environment for Practicing Negotiation in a Cultural Context. *International Journal of Artificial Intelligence in Education*, *19*(3), 289–308. Retrieved from http://iospress.metapress.com/index/T4M154U714064246.pdf

Kim, K., Eckhardt, M., Bugg, N., & Picard, R. W. (2009). The Benefits of Synchronized Genuine Smiles in Face-to-Face Service Encounters. In *CSE '09 Proceedings of the 2009 International Conference on Computational Science and Engineering* (pp. 801–808). Washington, DC: IEEE Computer Society. Retrieved from dspace.mit.edu/openaccess-disseminate/1721.1/56007?

Kim, T., Chang, A., Holland, L., & Pentland, A. S. (2008). Meeting Mediator : Enhancing Group Collaboration with Sociometric Feedback. *Group Dynamics*, 3183–3188. doi:10.1145/1460563.1460636

Kipp, M., Neff, M., & Albrecht, I. (2008). An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources And Evaluation*, *41*(3-4), 325–339. doi:10.1007/s10579-007-9053-5

Kline, P. (2000). *The handbook of psychological testing. Filozofska Istrazivanja* (Vol. 25, pp. 493–506). Routledge. Retrieved from http://books.google.com.library.gcu.edu:2048/books?hl=en&lr=&id=lm2RxaKaok8C&oi=fnd&pg=PR8&dq=psychological+testing&ots=BM3FzdUjZr&sig=FesTlagNEuypCmUBZHApDu7QtOo

Kluft, E. S. (1981). *Nonverbal communication and marriage: An investigation of the movement aspects of nonverbal communication between marital partners*. Drexel University.

Knapp, M. L. (1972). *Nonverbal Communication in Human Interaction*. Trade Paper.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. (T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist, Eds.)*Intelligent Virtual Agents*, *3661*, 1–14. doi:10.1007/11550617_28

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin*, *60*(4), 408–420. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14035446

Krasny, L., Williams, B. J., Provencal, S., & Ozonoff, S. (2003). Social skills interventions for the autism spectrum: essential ingredients and a model curriculum. *Child And Adolescent Psychiatric Clinics Of North America*, *12*(1), 107–122. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12512401

Krieger, B., Kinébanian, A., Prodinger, B., & Heigl, F. (2012). Becoming a member of the work force: perceptions of adults with Asperger Syndrome. *Work Reading Mass*, *43*(2), 141–57. doi:10.3233/WOR-2012-1392

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology. Education* (Vol. 79, p. 440). Sage. doi:10.2307/2288384

Krumhuber, E., Manstead, A. S. R., & Kappas, A. (2007). Temporal Aspects of Facial Displays in Person and Expression Perception: The Effects of Smile Dynamics, Head-tilt, and Gender. *Journal of Nonverbal Behavior*, *31*(1), 39–56. doi:10.1007/s10919-006-0019-x

Kublbeck, C., Ruf, T., & Ernst, A. (2009). A Modular Framework to Detect and Analyze Faces for Audience Measurement Systems. In *GI Jahrestagung* (pp. 3941–3953).

Kueblbeck, C., & Ernst, A. (2006). Face detection and tracking in video sequences using the modified census transformation. *Journal on Image and Vision Computing*, *24*(6), 564–572. Retrieved from http://www.sciencedirect.com/science/article/pii/S0262885605001605

Kupersmidt, J. B., Coie, J. D., & Dodge, K. A. (1990). The role of poor peer relationships in the development of disorder. (S. R. Asher & J. D. Coie, Eds.)*Peer rejection in childhood*, 274–305. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1990-97775-009&site=ehost-live

Ladd, G. W. (1990). Having friends, keeping friends, making friends, and being liked by peers in the classroom: Predictors of children's early school adjustment? *Child Development*, *61*(4), 1081–1100. doi:10.2307/1130877

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley & A. P. Danyluk (Eds.), *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning* (Vol. pages, pp. 282–289). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. doi:10.1038/nprot.2006.61

LaFrance, M. (2011). *Lip Service*. W. W. Norton & Company.

Lahm, E. A. (1996). Software that engages young children with disabilities: A study of design features. *Focus on Autism and Other Developmental Disabilities*, *11*(2), 115–124. doi:10.1177/108835769601100207

Levine, D. (2000). Virtual Attraction: What Rocks Your Boat. *CyberPsychology Behavior*, *3*(4), 565–573. doi:10.1089/109493100420179

Levitan, R., Gravano, A., & Hirschberg, J. (2011). Entrainment in Speech Preceding Backchannels. *Proc of ACL 2011*, 113–117. Retrieved from http://www.aclweb.org/anthology/P/P11/P11-2020.pdf

Lien, J. J., Kanade, T., Cohn, J. F., & Li, C.-C. L. C.-C. (1998). Automated facial expression recognition based on FACS action units. *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 390–395. doi:10.1109/AFGR.1998.670980

Lindström, M., Ståhl, A., Höök, K., Sundström, P., Laaksolathi, J., Combetto, M., … Bresin, R. (2006). Affective diary: designing for bodily expressiveness and self-reflection. In *CHI EA '06 CHI '06 Extended Abstracts on Human Factors in Computing Systems* (Vol. Montr\'{e}, pp. 1037–1042). New York: ACM Press. doi:10.1145/1125451.1125649

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In *IEEE International Conference on Automatic Face Gesture Recognition and Workshops FG* (pp. 298–305). Ieee. doi:10.1109/AFGR.2008.4813406

Lord, R. G., & Maher, K. J. (1993). Leadership and Information Processing: Linking Perceptions and Performance. *Academy of Management Review*, *18*(1), 153–156. doi:10.2307/258827

Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. In *Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 200–205). IEEE Comput. Soc. doi:10.1109/AFGR.1998.670949

MacDonald, R., Clark, M., Garrigan, E., & Vangala, M. (2005). Using video modeling to teach pretend play to children with autism. *Behavioral Interventions*, *20*(4), 225–238. doi:10.1002/bin.197

Madsen, M., El Kaliouby, R., Eckhardt, M., Hoque, M. E., Goodwin, M. S., & Picard, R. (2009). Lessons from participatory design with adolescents on the autism spectrum. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems CHI EA 09*, 3835–3840. doi:10.1145/1520340.1520580

Madsen, M., el Kaliouby, R., Goodwin, M., & Picard, R. (2008). Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder. *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility - Assets '08*, 19. doi:10.1145/1414471.1414477

Marsh, A. A., & Blair, R. J. R. (2008). Deficits in facial affect recognition among antisocial populations: a meta-analysis. *Neuroscience & Biobehavioral Reviews*, *32*(3), 454–465. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17915324

Martin, C. D. The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places [Book Review]. , 34 Ieee Spectrum 305 (1997). Cambridge University Press. doi:10.1109/MSPEC.1997.576013

Mast, M. S., & Hall, J. A. (2004). Who is the boss and who is not? Accuracy of judging status . *Journal of Nonverbal Behavior*, *28*, 145–165.

Mayer, R. E., & Chandler, P. (2001). When Learning Is Just a Click Away: Does Simple User Interaction Foster Deeper Understanding. *Journal of Educational Psychology, Vol. 93 Is*, 390 ST – When Learning Is Just a Click Away: Does. Retrieved from http://www.acu.edu:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=5442786&site=ehost-live&scope=site

McClure, E. ., & Nowicki, S. J. (n.d.). Associations between socialanxiety and nonverbal processing skill in preadolescent boys and girls. *Journal of Nonverbal Behavior*, *25*, 3–19.

McConnell, S. R. (2002). Interventions to facilitate social interaction for young children with autism: review of available research and recommendations for educational intervention and future research. *Journal of Autism and Developmental Disorders*, *32*(5), 351–372. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12463515

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599–616. doi:10.1037/0021-9010.79.4.599

Mcduff, D., Kaliouby, R. El, Senechal, T., Amr, M., Cohn, J. F., & Picard, R. (2013). Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild. In *The 2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'10)*. Portland.

McEvoy, M. A., Odom, S. L., & McConnell, S. R. (1992). Peer social competence intervention for young children with disabilitie. In S. L. Odom, S. R. McConnell, & M. A. McEvoy (Eds.), *Social competence of young children with disabilities: Issues and strategies for intervention*. Baltimore: Paul Brooks.

McGovern, T. V., & Tinsley, H. E. . (1978). Interviewer evaluations of interviewee nonverbal behavior . *Journal of Vocational Behavior*, *13*(2), 163–171.

McKeown, G., Valstar, M. F., Cowie, R., & Pantic, M. (2010). The SEMAINE corpus of emotionally coloured character interactions. *Multimedia and Expo ICME 2010 IEEE International Conference on*, 1079–1084. doi:10.1109/ICME.2010.5583006

Mehrabian, A. (1968). Communication without words. *Psychology Today*, 52–55.

Mehrabian, Albert. (2008). Communication without words. *Communication theory 2nd ed.* Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc6&NEWS=N&AN=2008-08778-013

Messinger, D. S., Fogel, A., & Dickson, K. L. (1999). What's in a smile? *Developmental Psychology*, *35*(3), 701–708. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10380861

Mignault, A., & Chaudhuri, A. (2003). The Many Faces of a Neutral Face: Head Tilt and Perception of Dominance and Emotion. *Journal of Nonverbal Behavior*, *27*(2), 111–132. doi:10.1023/A:1023914509763

Mitchell, P., Parsons, S., & Leonard, A. (2007). Mitchell, Parsons & Leonard (2007) VR learning_social_skills ASD. *Journal of Autism and Developmental Disorders*. Springer. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2008-00370-018&site=ehost-live

Morency, L., Kok, I. De, & Gratch, J. (2008). Predicting Listener Backchannels : A Probabilistic Multimodal Approach. (H. Prendinger, J. C. Lester, & M. Ishizuka, Eds.)*Intelligent Virtual Agents*, *5208/2008*(1), 1–14. doi:10.1007/978-3-540-85483-8

Morency, L. P. Modeling Human Communication Dynamics [Social Sciences]. , 27 IEEE Signal Processing Magazine 112–116 (2010). doi:10.1109/MSP.2010.937500

Morgan, N., & Fosler-Lussier, E. Combining multiple estimators of speaking rate. , 2 Proceedings of the 1998 IEEE International Conference on Acoustics Speech and Signal

Processing ICASSP 98 Cat No98CH36181 729–732 (1998). Ieee. doi:10.1109/ICASSP.1998.675368

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology*, *58*(3), 583–611.

Mori, M. (1970). The Uncanny Valley. *Energy*, *7*(4), 33–35. doi:10.1162/pres.16.4.337

Muller, E., Schuler, A., Burton, B. A., & Yates, G. B. (2003). Meeting the vocational support needs of individuals with Asperger Syndrome and other autism spectrum disabilities. *Journal of Vocational Rehabilitation*, *18*(3), 163–175.

Murphy, K. P. (2001). The Bayes Net Toolbox for Matlab. *Computing Science and Statistics*, *33*(2), 20. Retrieved from http://citeseer.ist.psu.edu/murphy01bayes.html

Murray, D. (1997). Autism and information technology:therapy with computers. In *Autism and Learning: a guide to good practice*. London: David Fulton Publishers.

Mutlu, B. (2011). Designing Embodied Cues for Dialog with Robots. *AI Magazine*, *32*(4), 17–30. doi:10.1609/aimag.v32i4.2376

Nanjo, H., & Kawahara, T. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. , 1 2002 IEEE International Conference on Acoustics Speech and Signal Processing I–725–I–728 (2002). IEEE. doi:10.1109/ICASSP.2002.5743820

Narayanan, S. A. B., & Potamianos, A. A. C. (2002). Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, *10*(2), 65–78. doi:10.1109/89.985544

Nass, C., Moon, Y., & Green, N. (1997). Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology*, *27*(10), 864–876. doi:10.1111/j.1559-1816.1997.tb00275.x

Nooteboom, S. (1997). The prosody of speech: melody and rhythm. In *The handbook of phonetic sciences* (pp. 640–673). Blackwell, Oxford: W.J. Hardcastle, J. Laver (eds.).

Nowicki Jr., S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: the diagnostic analysis of nonverbal accuracy scale. *Journal of nonverbal behavior*, *18*, 9–35.

Nowicki, S., & Carton, E. (1997). The relation of nonverbal processing ability of faces and voices and children's feelings of depression and competence. *The Journal of genetic psychology*, *158*(3), 357–363. Retrieved from http://heldref-publications.metapress.com/index/h18377666144g566.pdf

Nowicki, S., & Mitchell, J. (1998). Accuracy in identifying affect in child and adult faces and voices and social competence in preschool children. *Genetic social and general psychology monographs*, *124*(1), 39–59. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9495028

Nuance Communications. (n.d.). Retrieved June 28, 2013, from http://www.nuance.com/for-developers/dragon/index.htm

Ochs, M., Niewiadomski, R., & Pelachaud, C. (2010). How a virtual agent should smile?: morphological and dynamic characteristics of virtual agent's smiles. In *IVA'10 Proceedings of the 10th international conference on Intelligent virtual agents* (pp. 427–440). Springer. Retrieved from http://www.springerlink.com/index/59U7338607Q843G5.pdf

Olguin Olguin, D., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible organizations: technology and methodology for automatically measuring organizational behavior. *IEEE transactions on systems man and cybernetics Part B Cybernetics a publication of the IEEE Systems Man and Cybernetics Society*, *39*(1), 43–55. doi:10.1109/TSMCB.2008.2006638

Pantic, M, Valstar, M., Rademaker, R., & Maat, L. (2005). Web-Based Database for Facial Expression Analysis. *2005 IEEE International Conference on Multimedia and Expo*, 317–321. doi:10.1109/ICME.2005.1521424

Pantic, Maja. (2009). Machine Analysis of Facial Behaviour : Naturalistic & Dynamic Behaviour 2 . The Process of Automatic Facial Behaviour Analysis. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *364*(1535), 3505–13. doi:10.1098/rstb.2009.0135

Pantic, Maja, Rothkrantz, L. J. M., & Koppelaar, H. (1998). Automation of nonverbal communication of facial expressions. *EUROMEDIA 98, SCS International*, 86–93. Retrieved from citeseer.ist.psu.edu/487993.html

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL : A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. (E. Nzewi, G. Reddy, S. Luster-Teasley, V. Kabadi, S.-Y. Chang, K. Schimmel, & G. Uzochukwu, Eds.)*Journal of Retailing*, *64*(1), 12–40. doi:10.1016/S0148-2963(99)00084-3

Pardas, M., Bonafonte, A., & Landabaso, J. L. (2002). Emotion recognition based on MPEG-4 Facial Animation Parameters. *2002 IEEE International Conference on Acoustics Speech and Signal Processing*, *4*, IV–3624–IV–3627. doi:10.1109/ICASSP.2002.5745440

Parke, R. D., O'Neil, R., Spitzer, S., Isley, S., Welsh, M., Wang, S., ... Cupp, R. (1997). A longitudinal assessment of sociometric stability and the behavioral correlates of children's social acceptance. *MerrillPalmer Quarterly*, *43*(4), 635–662.

Parsons, C. K., & Liden, R. C. (1984). Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology*, *69*(4), 557–568. doi:10.1037/0021-9010.69.4.557

Patel, R. A., Hartzler, A., Pratt, W., & Back, A. (2013). Visual Feedback on Nonverbal Communication : A Design Exploration with Healthcare Professionals. In *Pervasive Health*.

Patrick, H. (1997). Social self-regulation : Exploring the relations between children' s social relationships , academic self-regulation, and school performance. *Educational Psychologist*, *32*(4), 209–220. doi:10.1207/s15326985ep3204_2

Pierrehumbert, J. (1980). *The Phonetics and Phonology of English Intonation*. (Anonymous, Ed.)*dspacemitedu*. MIT. Retrieved from

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=11122867722242645121

Pollard, N. L. (n.d.). Development of social interaction skills in preschool children with autism: A review of the literature. In *Child & Family Behavior Therapy* (pp. 1–16). Binghamton: Haworth Press.

Posthuma, R. A., Morgeson, F. P., & Campion, M. A. (2002). Beyond Employment Interview Validity: a Comprehensive Narrative Review of Recent Research and Trends Over Time. *Personnel Psychology*, *55*(1), 1–81. doi:10.1111/j.1744-6570.2002.tb00103.x

Pulakos, E. D., & Schmitt, N. (1995). Experience-Based and Situational Interview Questions: Studies of Validity. *Personnel Psychology*, *48*(2), 289–308. doi:10.1111/j.1744-6570.1995.tb01758.x

Quattoni, A., Wang, S., Morency, L.-P., Collins, M., & Darell, T. (2007). Hidden Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, *29*(10), 1848–1852.

Rabiner, L. R. (1990). A tutorial on hidden Markov models and selected applications in speech recognition. In A. Waibel & K.-F. Lee (Eds.), *Readings in speech recognition* (Vol. 77, pp. 267–296). San Francisco: Morgan Kaufmann Publishers Inc. doi:10.1109/5.18626

Ray, C. E., & Elliott, S. N. (2006). Social adjustment and academic achievement: A predictive model for students with diverse academic and behavior competencies. *School Psychology Review*. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-33750564263&partnerID=40&md5=3dd4d4ee1dc89a83969d46a6568a2131

Rickel, J., & Johnson, L. W. (2000). *Task-oriented collaboration with embodied agents in virtual worlds* (pp. 95–122). Cambridge: MIT Press.

Ring, L., Barry, B., Totzke, K., & Bickmore, T. (2013). Addressing Loneliness and Isolation in Older Adults: Proactive Affective Agents Provide Better Support. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*.

Rogers, S. J., & Bennetto, L. (2000). Intersubjectivity in autism: The roles of imitation and executive function. (A. Wetherby & B. M. Prizant, Eds.)*Autism Spectrum Disorders A transactional developmental perspective*, 79–107. Retrieved from http://www.library.gatech.edu:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2001-00134-004

Rosenberger, T. (1998). *Prosodic Font: the Space between the Spoken and the Written. Visual Studies*. Massachusetts Institute of Technology. Retrieved from http://alumni.media.mit.edu/~tara/ProsodicFont.pdf

Rosenthal, R., Hall, J., Matteg, A., Rogers, M. R. ., & Archer, P. L. (1979). *Sensitivity to nonverbal communication: The PONS Test*. Baltimore: Johns Hopkins University Press.

Ross, K., & Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Comput. Speech Lang.*, *10*, 155 –185.

Ruscio, A. M., Brown, T. A., Chiu, W. T., Sareen, J., Stein, M. B., & Kessler, R. C. (2008). Social fears and social phobia in the USA: results from the National Comorbidity Survey Replication. *Psychological Medicine*, *38*(1), 15–28. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2262178&tool=pmcentrez&rendertype=abstract

Sacks, O. (1995). *An Anthropologist on Mars: Seven Paradoxical Tales*.

Sambeth, A., Ruohio, K., Alku, P., Fellman, V., & Huotilainen, M. (2008). Sleeping newborns extract prosody from continuous speech. *Clin Neurophysiol*, *119*(2), 332–341. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18069059

Schlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations. Monterey BrooksCole*. Brooks/Cole. Retrieved from http://www.getcited.org/pub/102061888

Schmidt, F. L., & Rader, M. (1999). Exploring the boundary conditions for interview validity: Meta-analytic validity findings for a new interview type. *Personnel Psychology*, *52*(2), 445–464. doi:10.1111/j.1744-6570.1999.tb00169.x

Schneider, K., & Josephs, I. (1991). The expressive and communicative functions of preschool children's smiles in an achievement-situation. *Journal of Nonverbal behaviour*, *15*(3), 185–198. doi:10.1007/BF01672220

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ... Martin, W. (2011). Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing*, *3*(2), 1–20. doi:10.1109/T-AFFC.2011.34

Shlenker, B. R. (1980). *Impression management: The self-concept, social identity, and interpersonal relations* (p. 344). Monterey, CA: Brooks/Cole Pub. Co.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., & Tur, G. (2000). Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, *32*(1-2), 127–154. doi:10.1016/S0167-6393(00)00028-5

Siegler, R. S. (1995). How Does Change Occur: A Microgenetic Study of Number Conservation. *Cognitive Psychology*, *28*(3), 225–273. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0029320238&partnerID=40&md5=2a3da05898beb4fe1259c317fc67346f

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., ... Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. In *2nd Internatioanl Conference on Spoken Language Processing ICSLP 92* (pp. 867–870).

Smith, J. (2000). *GrandChair : conversational collection of grandparents' stories*. Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/62350

Sridhar, V. K. R., Bangalore, S., & Narayanan, S. S. Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. , 16 Ieee

Transactions On Audio Speech And Language Processing 797–811 (2008). Association for Computational Linguistics. doi:10.1109/TASL.2008.917071

Ståhl, A., Höök, K., Svensson, M., Taylor, A. S., & Combetto, M. (2008). Experiencing the Affective Diary. *Personal and Ubiquitous Computing*, *13*(5), 365–378. doi:10.1007/s00779-008-0202-7

Strickland, D., Marcus, L. M., Mesibov, G. B., & Hogan, K. (1996). Brief report: two case studies using virtual reality as a learning tool for autistic children. *Journal of Autism and Developmental Disorders*. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=8986851

Sturm, J., Herwijnen, O. H., Eyck, A., & Terken, J. (2007). Influencing social dynamics in meetings through a peripheral display. *Proceedings of the ninth international conference on Multimodal interfaces ICMI 07*, 263. doi:10.1145/1322192.1322238

Tamburini, F., & Caini, C. (2005). An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech. *International Journal of Speech Technology*, *8*(1), 33–44. doi:10.1007/s10772-005-4760-z

Tartaro, A., & Cassell, J. (2008). Playing with Virtual Peers : Bootstrapping Contingent Discourse in Children with Autism. In *ICLS'08 Proceedings of the 8th international conference on International conference for the learning sciences* (Vol. 2, pp. 382–389). International Society of the Learning Sciences. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=5027430324945433305related:2TJmez4FxUUJ

Teeters, A., Kaliouby, R. El, & Picard, R. (2006). Self-Cam: feedback from what would be your social partner. In *ACM SIGGRAPH 2006 Research posters on SIGGRAPH 06* (Vol. MI, p. 138). ACM Press. doi:10.1145/1179622.1179782

Thorsen, N. G. (1980). A study of perception of sentence intonation--evidence from Danish. *Journal of the Acoustical Society of America*, *67*(3), 1014–1030. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7358909

Tian, Y. I., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 97–115. doi:10.1109/34.908962

Tickle-Degnen, L., & Rosenthal, R. (1990). The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, *1*(4), 285–293. doi:10.1207/s15327965pli0104_1

Vaish, A., & Striano, T. (2004). Is visual reference necessary? Contributions of facial versus vocal cues in 12-month-olds' social referencing behavior. *Developmental Science*, *7*(3), 261–269. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=15595366

Valstar, M. F., Gunes, H., & Pantic, M. (2007). How to Distinguish Posed from Spontaneous Smiles using Geometric Features. In *ICMI '07 Proceedings of the 9th international conference on Multimodal interfaces* (pp. 38–45). New York: ACM. Retrieved from http://www.cs.nott.ac.uk/~mfv/Documents/ICMI07_155_ValstarEtAl_final.pdf

Wagner, J., Lingenfelser, F., & Andre, E. (2011). The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognitions.

Wagner, Johannes, Lingenfelser, F., & Andre, E. (2011). The Social Signal Interpretation Framework (SSI) for Real Time Signal Processing and Recognition. In *Proceedings of Interspeech*. Florence.

Wallach, H. S., Safir, M. P., & Bar-Zvi, M. (2009). Virtual reality cognitive behavior therapy for public speaking anxiety: a randomized clinical trial. *Behavior Modification*, *33*(3), 314–338. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19321811

Wang, D., & Narayanan, S. An Acoustic Measure for Word Prominence in Spontaneous Speech. , 15 Ieee Transactions On Audio Speech And Language Processing 690–701 (2007). doi:10.1109/TASL.2006.881703

Wang, D. W. D., & Narayanan, S. S. Robust Speech Rate Estimation for Spontaneous Speech. , 15 Ieee Transactions On Audio Speech And Language Processing 2190–2201 (2007). doi:10.1109/TASL.2007.905178

Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. doi:10.1145/365153.365168

Welsh, Parke, R. D., Widaman, K., & O'Neil, R. (2001). Linkages between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology*, *39*(6), 463–482.

Wentzel, K. R. (1991a). Social Competence at School: Relation Between Social Responsibility and Academic Achievement. *Review of Educational Research*, *61*(1), 1–24. doi:10.3102/00346543061001001

Wentzel, K. R. (1991b). Relations between Social Competence and Academic Achievement in Early Adolescence. *Child Development*, *62*(5), 1066–1078. doi:10.2307/1131152

Wiesnerf, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the lemployment interview. *Journal of Occupational Psychology*, *61*(4), 275–290. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8325.1988.tb00467.x/abstract

Wightman, C. W., & Ostendorf, M. Automatic labeling of prosodic patterns. , 2 Ieee Transactions On Speech And Audio Processing 469–481 (1994). IEEE. doi:10.1109/89.326607

Wilczynski, S. M., Menousek, K., Hunter, M., & Mudgal, D. (2007). Individualized education programs for youth with Autism Spectrum Disorders. *Psychology in the Schools*, *44*(7), 653. doi:10.1002/pits

Wood, J. A. (2006). NLP revisited: nonverbal communications and signals of trustworthiness. *Journal of Personal Selling and Sales Management*, *XXVI*, 197–204.

Wooters, C., & Huijbregts, M. (2008). The ICSI RT07s Speaker Diarization System. *Multimodal Technologies for Perception of Humans*, *4625*, 509–519. doi:10.1007/978-3-540-68585-2_48

Wright, P. M., Lichtenfels, P. A., & Pursell, E. D. (1989). The structured interview: Additional studies and a metaanalysis. *Journal of Occupational Psychology, 62*, 191–199.

Yacoob, Y., & Davis, L. S. (1996). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(6), 636–642. doi:10.1109/34.506414

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics* (Vol. 123, pp. 5687–5690). doi:10.1121/1.2935783

Appendices

APPENDIX A: WEAK LANGUAGE DICTIONARY

- Anyways
- Whatever
- Things
- Guess
- Well
- Like
- Basically
- Totally
- Really
- Dude
- Ok
- My bad
- Um
- That is SO not true
- To tell ya the truth
- My bad, duh…
- LOL, GTG, BRB, OK?
- Are you really going to go there?
- Dude
- I'm not going to lie to you
- So, anyways
- I go, like yeah, and she says..
- Know what I mean?
- Totally!
- That is SO not going to happen
- Let me make a point
- Let me start by
- And this is just my opinion
- First things first
- make no mistake
- It takes all kinds, I must admit
- What I mean to say
- I'm going to tell you that
- Basically, my point is
- What I'm trying to say is
- In other words, let me just add
- At the end of the day, it is what it is
- Like I said,
- I don't want to repeat myself
- When all is said and done
- At any rate, that's the thing
- Look, I could go on and on

APPENDIX B-1: QUESTIONNAIRE USED IN THE INTERVENTION FOR THE PARTICIPANTS.

PART II: Questionnaire for Participants

* Required

Participant number *

Name?

Major

My performance in this interview was '_____' *

1 2 3 4 5 6 7

really bad ◯ ◯ ◯ ◯ ◯ ◯ ◯ really good

I found the feedback from the counselor useful. *

1 2 3 4 5 6 7

strongly agree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly disagree

I felt nervous during the interview. *

1 2 3 4 5 6 7

strong disagree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly agree

I spoke really slow during the interview. *

1 2 3 4 5 6 7

strongly agree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly disagree

I use a lot of words such as "umm", "like", "basically" etc. when I talk. *

1 2 3 4 5 6 7

strongly disagree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly agree

In the interview, I came across as a friendly person. *

1 2 3 4 5 6 7

strongly disagree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly agree

I am good at pausing in between my statements. *

1 2 3 4 5 6 7

strongly agree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly disagree

I think I have an engaging tone of voice. *

1 2 3 4 5 6 7

strongly disagree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly agree

I was stressed out during the interview. *

1 2 3 4 5 6 7

strong agree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly disagree

I was focused during the interview. *

1 2 3 4 5 6 7

strong agree ◯ ◯ ◯ ◯ ◯ ◯ ◯ strongly disagree

Submit

Never submit passwords through Google Forms.

APPENDIX B-2: QUESTIONNAIRE USED IN THE INTERVENTION FOR THE COUNSELORS/TURKERS.

Questionnaire for Counselors

* Required

Participant no *

Participant name *

The overall rating for this person during this interview was "_____." *

	1	2	3	4	5	6	7	
very bad	◯	◯	◯	◯	◯	◯	◯	very good

I would definitely recommend hiring this person. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

I would love to work with this person as a colleague. *

	1	2	3	4	5	6	7	
strongly agree	◯	◯	◯	◯	◯	◯	◯	strongly disagree

This participant appeared very engaged during the interview. *

	1	2	3	4	5	6	7	
strongly agree	◯	◯	◯	◯	◯	◯	◯	strongly disagree

This participant sounded very excited about the job. *

	1	2	3	4	5	6	7	
strongly disgree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant maintained eye contact during the interview. *

	1	2	3	4	5	6	7	
strongly agree	◯	◯	◯	◯	◯	◯	◯	strongly disagree

This participant smiled when appropriate during the interview. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant's speaking rate was "_____." *

	1	2	3	4	5	6	7	
really slow	◯	◯	◯	◯	◯	◯	◯	really fast

This participant used a lot of fillers (e.g., "umm" "ahh" "like" etc.). *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant appeared friendly during the interview. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant paused when needed during the interview. *

	1	2	3	4	5	6	7	
strongly agree	◯	◯	◯	◯	◯	◯	◯	strongly disagree

This participant demonstrated engaging tone of voice during the interview. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant's answers were very structured. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant did not seem calm during the interview. *

	1	2	3	4	5	6	7	
strongly agree	◯	◯	◯	◯	◯	◯	◯	strongly disagree

This participant understood my feedback on what s/he needs to work on. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant looked stressed during the interview. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

This participant looked focused on the interview. *

	1	2	3	4	5	6	7	
strongly disagree	◯	◯	◯	◯	◯	◯	◯	strongly agree

Submit

Never submit passwords through Google Forms.

APPENDIX B-3: QUESTIONNAIRE USED IN THE INTERVENTION FOR THE STUDENTS ON THE MACH SYSTEM (VIDEO GROUP)

MACH: Video Questionnaire

* Required

Participant ID *

How many rounds of interview did you practice? *

Interviewer was responding to your behaviors. *

	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

You would prefer the interviewer to look like a cartoon, not a human. *

	1	2	3	4	5	6	7	
Strongly agree	○	○	○	○	○	○	○	Strongly disagree

You explored new things about your own behavior through this practice. *

	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

The life-like characteristics of the interviewer stressed you out. *

	1	2	3	4	5	6	7	
Strongly agree	○	○	○	○	○	○	○	Strongly disagree

Being asked the interview questions stressed you out. *

	1	2	3	4	5	6	7	
Strongly agree	○	○	○	○	○	○	○	Strongly disagree

You found this interview experience useful. *

	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

Watching your own video was useful. *

	1	2	3	4	5	6	7

Strongly disagree ○ ○ ○ ○ ○ ○ ○ Strongly agree

The overall interview experience was very stressful for you. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You would have preferred to receive additional feedback apart from watching your video. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You were bored watching your own video. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You would consider using this software in the future for self-reflection. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You would recommend your friends to use this software. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

Submit

Never submit passwords through Google Forms.

APPENDIX B-4: QUESTIONNAIRE USED IN THE INTERVENTION FOR THE STUDENTS ON THE MACH SYSTEM (FEEDBACK GROUP)

MACH: Feedback Questionnaire

* Required

Participant ID *

How many rounds of interview did you practice? *

Interviewer was responding to your behaviors. *

1 2 3 4 5 6 7

Strongly disagree ○ ○ ○ ○ ○ ○ ○ Strongly agree

You would prefer the interviewer to look like a cartoon, not a human. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You explored new things about your own behavior through this practice. *

1 2 3 4 5 6 7

Strongly disagree ○ ○ ○ ○ ○ ○ ○ Strongly agree

The life-like characteristics of the interviewer stressed you out. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

Being asked the interview questions stressed you out. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You found this interview experience useful. *

1 2 3 4 5 6 7

Strongly disagree ○ ○ ○ ○ ○ ○ ○ Strongly agree

Watching your own video was useful. *

1 2 3 4 5 6 7

Strongly disagree ○ ○ ○ ○ ○ ○ ○ Strongly agree

The overall interview experience was very stressful for you? *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

Please rate how much you liked with the immediate feedback (without your video) *

1 2 3 4 5 6 7

not useful at all ○ ○ ○ ○ ○ ○ ○ very useful

Please rate how much you liked with the time line feedback (containing your video) *

1 2 3 4 5 6 7

not useful at all ○ ○ ○ ○ ○ ○ ○ very useful

What aspects of the feedback did you find most useful (check all that applies. Check NA if none applies) *

- ☐ Smile information
- ☐ Pauses
- ☐ Speaking rate
- ☐ Weak language
- ☐ Intonation
- ☐ Head nod/shake
- ☐ Volume
- ☐ Transcription of the interview
- ☐ Emphasis on words
- ☐ NA

You would consider using this software in the future for self-reflection. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

You would recommend your friends to use this tool. *

1 2 3 4 5 6 7

Strongly agree ○ ○ ○ ○ ○ ○ ○ Strongly disagree

Submit

Never submit passwords through Google Forms.

APPENDIX B-5: QUESTIONNAIRE USED IN THE INTERVENTION FOR THE STUDENTS ON SYSTEM USABILITY

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very cumbersome to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

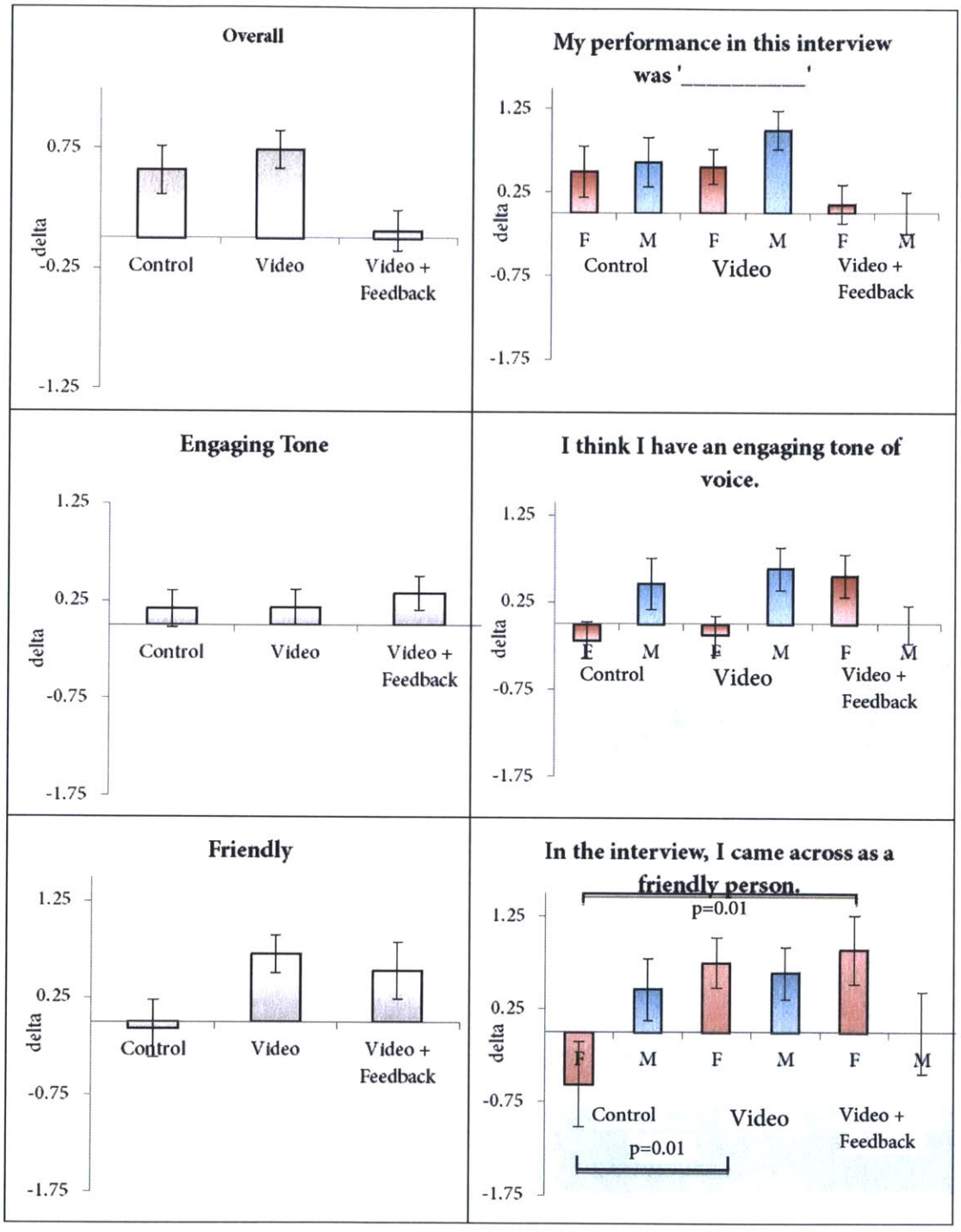
APPENDIX C-1: ONE-WAY ANOVA ANALYSIS ON THE PARTICIPANT'S SELF-RATINGS (CONDITIONS)

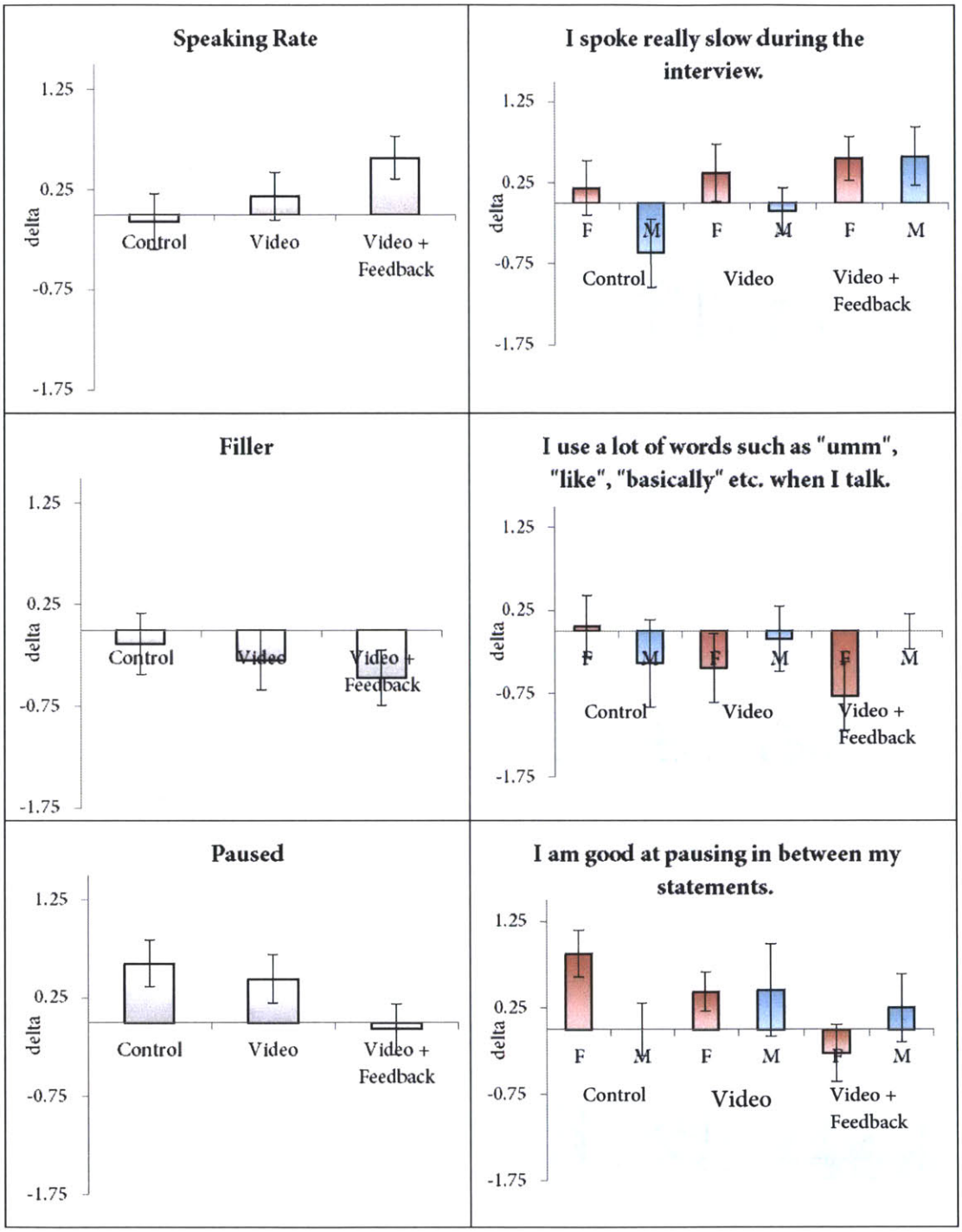
	Effect Test	Feedback vs. Video	Feedback vs. Control	Control vs. Video
Overall	F(2,83)=0.0070, p=0.99	F(1,83)=0.0018, p=0.96	F(1,83)=0.0058, p=0.94	F(1,83)=0.0134, p=0.91
Fillers	F(2, 83)=0.6574, p=0.52	F(1,83)=1.1289, p=0.29	F(1,83)=0.0182, p=0.89	F(1,83)=0.8510, p=0.36
Speaking rate	F(2, 83)=1.2779, p=0.28	F(1,83)=1.5290, p=0.22	F(1,83)=2.2027, p=0.14	F(1,83)=0.0478, p=0.83
Friendly	F(2,83)=1.2724, p=0.29	F(1,83)=0.3950, p=0.53	F(1,83)=0.9623, p=0.33	F(1,83)=2.4921, p=0.12
Pausing	F(2,83)=0.2192, p=0.80	F(1,83)=0.0925, p=0.76	F(1,83)=0.1356, p=0.71	F(1,83)=0.4361, p=0.51
Engaging Tone	F(2,83)=0.5460, p=0.58	F(1,83)=0.2292, p=0.63	F(1,83)=0.3393, p=0.56	F(1,83)=1.0861, p=0.30
Feedback Useful	F(2,83)=1.0608, p=0.35	F(1,83)=0.0002, p=0.99	F(1,83)=1.6439, p=0.20	F(1,83)=1.5261, p=0.22
Stressed	F(2,83)=2.4993, p=0.09	F(1,83)=1.9567, p=0.17	F(1,83)=0.7270, p=0.40	**F(1,83)=4.9187, p=0.03**
Nervous	F(2,83)=0.7733, p=0.46	F(1,83)=0.0246, p=0.88	F(1,83)=1.0096, p=0.32	F(1,83)=1.2863, p=0.26
Focused	F(2,83)=2.3548, p=0.10	F(1,83)=0.1454, p=0.70	**F(1,83)=4.2240, p=0.04**	F(1,83)=2.6342, p=0.11

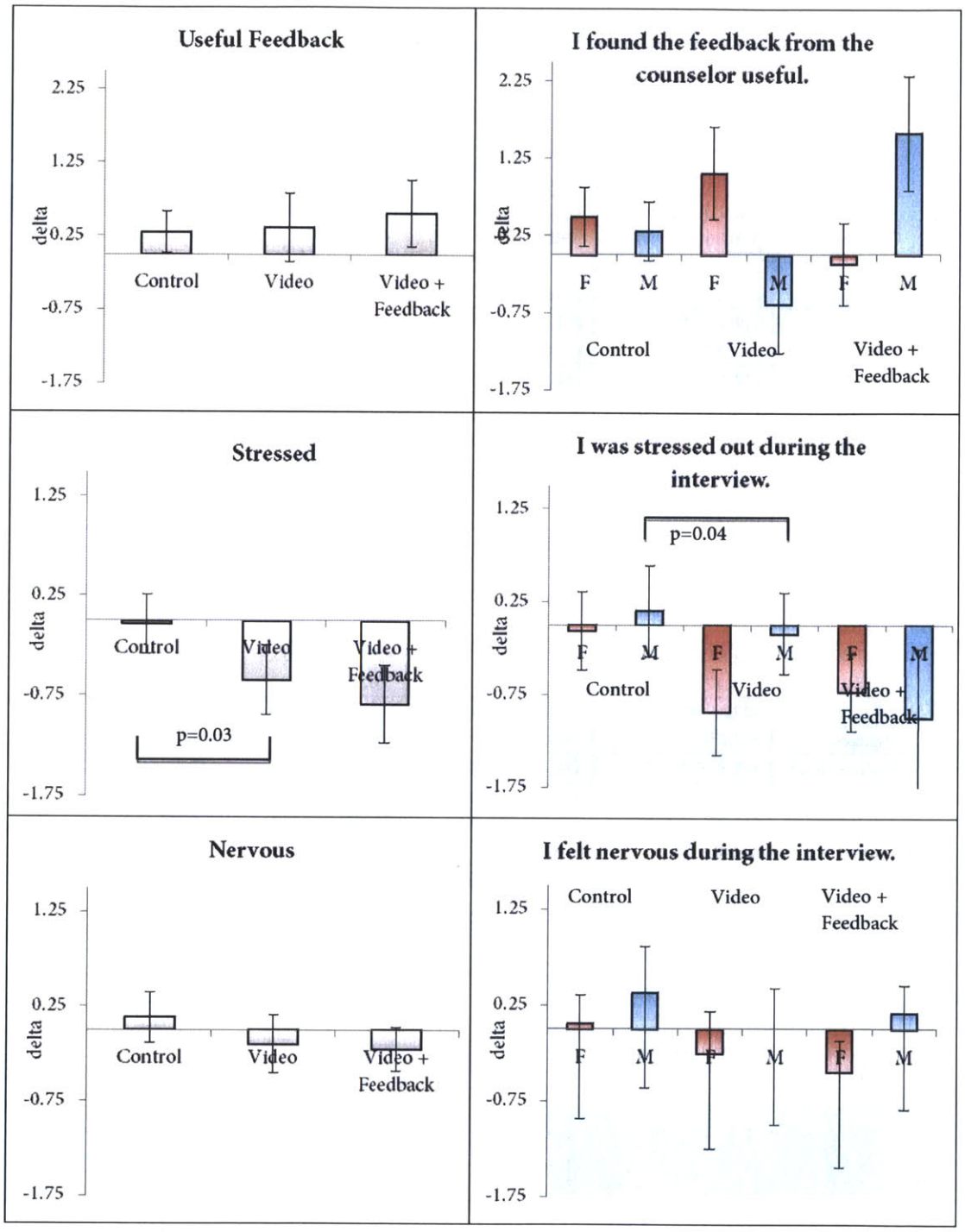
APPENDIX C-2: TWO-WAY ANOVA ANALYSIS ON THE PARTICIPANT'S SELF-RATINGS (CONDITIONS WITH GENDER)

	Effect Test	Female Feedback vs. Video	Female Feedback vs. Control	Female Control vs. Video	Male Feedback vs. Video	Male Feedback vs. Control	Male Control vs. Video
Fillers	F(2, 80)= 0.0823, p=0.92	F(1,80)=0. 4025, p=0.53	F(1,80)=0.0 203, p=0.89	F(1,80)=0. 5838, p=0.45	F(1,80)=0. 7808, p=0.38	F(1,80)=0. 1539, p=0.70	F(1,80)=0 .2500, p=0.62
Speaking Rate	F(2,80)= 0.1102, p=0.90	F(1,80)= 1.3893, p=0.24	F(1,80)=1.9 547, p=0.17	F(1,80)=0. 0379, p=0.85	F(1,60)=0. 2326, p=0.63	F(1,80)= 0.0115, p=0.91	F(1,80)=0 .3633, p=0.55
Friendly	**F(2,80) =4.131, p=0.02**	F(1,80)=0. 0035, p=0.95	**F(1,80)=6.5 805, p=0.01**	**F(1,80)=6. 4962, p=0.01**	F(1.80)=0. 9286, p=0.34	F(1,80)=2. 2042, p=0.14	F(1,80)=0 .2385, p=0.63
pausing	F(2,80)= 0.4413, p=0.64	F(1,80)=0. 0184, p=0.89	F(1,80)=0.7 096, p=0.40	F(1,80)=0. 9056, p=0.34	F(1,80)=0. 1053, p=0.75	F(1,80)=0. 2082, p=65	F(1,80)=0 .0148, p=0.90
Engaging Tone	F(2,80)= 0.5827, p=0.56	F(1,80)=0. 0109, p=0.92	F(1,80)=0.0 526, p=0.82	F(1,80)=0. 0144, p=0.90	F(1.80)=0. 7419, p=0.39	F(1,80)=0. 3992, p=0.53	F(1,80)=2 .1883, p=0.14
Overall	F(2,80)= 0.6649, p=0.52	F(1,80)=0. 0414, p=0.83	F(1,80)=0.3 636, p=0.55	F(1,80)=0. 6181, p=0.43	F(1,80)= 0.0270, p=0.87	F(1,80)=0. 3799, p=0.54	F(1,80)=0 .5887, p=0.45
Feedback Useful	F(2,80)= 0.4562, p=0.64	F(1,80)=0. 1240, p=0.73	F(1,80)=0.4 910, p=0.49	F(1,80)=0. 1109, p=0.74	F(1,80)=0. 1545, p=0.70	F(1,80)=1. 3304, p=0.25	F(1,80)=2 .3140, p=0.13
Stressed	F(2,80)= 0.4945, p=0.61	F(1,80)=1. 2318, p=0.27	F(1,80)=0.0 093, p=0.92	F(1,80)=1. 4125, p=0.24	F(1,80)=0. 6981, p=0.41	F(1,80)=1. 5511, p=0.22	**F(1,80)=4 .2173, p=0.04**
Nervous	F(2,80)= 0.0762, p=0.93	F(1,80)=0. 0944, p=0.76	F(1,80)=0.6 635, p=0.42	F(1,80)=1. 1967, p=0.28	F(1,80)=0. 0157, p=0.90	F(1,80)=0. 3183, p=0.57	F(1,80)=0 .1818, p=0.67
Focused	F(2,80)= 0.0315, p=0.97	F(1,80)=0. 1952, p=0.66	F(1,80)=2.8 150, p=0.10	F(1,80)=1. 4238, p=0.27	F(1,80)=0. 0032, p=0.95	F(1,80)=1. 3166, p=0.25	F(1,80)=1 .1354, p=0.29

APPENDIX C-3: GRAPHICAL ANALYSIS BASED ON THE PARTICIPANT'S SELF-RATINGS




Speaking Rate
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
I spoke really slow during the interview.
1.25
0.25
-0.75
-1.75
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Filler
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
I use a lot of words such as "umm", "like", "basically" etc. when I talk.
1.25
0.25
-0.75
-1.75
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Paused
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
I am good at pausing in between my statements.
1.25
0.25
-0.75
-1.75
delta
F
M
F
M
F
M
Control
Video
Video + Feedback


Useful Feedback
2.25
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
I found the feedback from the counselor useful.
2.25
1.25
0.25
-0.75
-1.75
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Stressed
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
p=0.03
I was stressed out during the interview.
1.25
0.25
-0.75
-1.75
delta
p=0.04
F
M
F
M
F
M
Control
Video
Video + Feedback
Nervous
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
I felt nervous during the interview.
1.25
0.25
-0.75
-1.75
delta
Control
Video
Video + Feedback
F
M
F
M
F
M

APPENDIX D-1: ONE-WAY ANOVA ANALYSIS BASED ON THE COUNSELOR'S RATINGS (CONDITIONS)

	Effect Test	Feedback vs. Video	Feedback vs. Control	Control vs. Video
Overall	F(2,85)=2.9289, p=0.06	**F(1,85)=5.6353, p=0.02**	F(1,85)=2.4028, p=0.12	F(1,85)=0.7381, p=0.39
Speaking Rate	F(2,85)=0.1596, p=0.85	F(1,85)=0.0121, p=0.91	F(1,85)=0.2908, p=0.59	F(1,85)=0.1694, p=0.68
Filler	F(2,85)=1.1515, p=0.32	F(1,85)=2.2883, p=0.13	F(1,85)=0.6894, p=0.41	F(1,85)=0.4892, p=0.49
Friendly	F(2,85)=0.2779, p=0.76	F(1,85)=0.4099, p=0.52	F(1,85)=0.4119, p=0.52	F(1,85)=0.0002, p=0.99
Paused	F(2,85)=0.4022, p=0.67	F(1,85)=0.0896, p=0.76	F(1,85)=0.3591, p=0.55	F(1,85)=0.7666, p=0.38
Engaging Tone	F(2,85)=1.4982, p=0.23	F(1,85)=2.9940, p=0.09	F(1,85)=0.7591, p=0.39	F(1,85)=0.7671, p=0.38
Hiring	F(2,85)=0.8595, p=0.43	F(1,85)=1.6293, p=0.21	F(1,85)=0.7636, p=0.38	F(1,85)=0.1789, p=0.67
Feedback	F(2,85)=0.4668, p=0.63	F(1,85)=0.3637, p=0.55	F(1,85)=0.9053, p=0.34	F(1,85)=0.1025, p=0.75
Stressed	F(2,85)=0.2661, p=0.77	F(1,85)=0.2492, p=0.62	F(1,85)=0.4985, p=0.48	F(1,85)=0.0347, p=0.85
Engaged	F(2,85)=0.5295, p=0.59	F(1,85)=0.9791, p=0.33	F(1,85)=0.5210, p=0.47	F(1,85)=0.0813, p=0.78
Excited about job	F(2,85)=1.1427, p=0.32	F(1,85)=0.1940, p=0.66	F(1,85)=2.1923, p=0.14	F(1,85)=0.9848, p=0.32
Eye Contact	F(2,85)=0.1296, p=0.88	F(1,85)=0.0188, p=0.89	F(1,85)=0.1332, p=0.72	F(1.85)=0.2386, p=0.63
Colleague	F(2,85)=0.2366, p=0.79	F(1,85)=0.3001, p=0.59	F(1,85)=0.3937, p=0.53	F(1.85)=0.0039, p=0.95
Smile	F(2,85)=0.9047, p=0.41	F(1,85)=0.5997, p=0.44	F(1,85)=1.7858, p=0.19	F(1,85)=0.2722, p=0.60
Not Calm	F(2,85)=1.0092, p=0.37	F(1,85)=1.9062, p=0.17	F(1,85)=0.9115, p=0.34	F(1,85)=0.2010, p=0.66
Focused	F(2,85)=0.2679, p=0.77	F(1,85)=0.1795, p=0.67	F(1,85)=0.1033, p=0.75	F(1,85)=0.1795, p=0.67

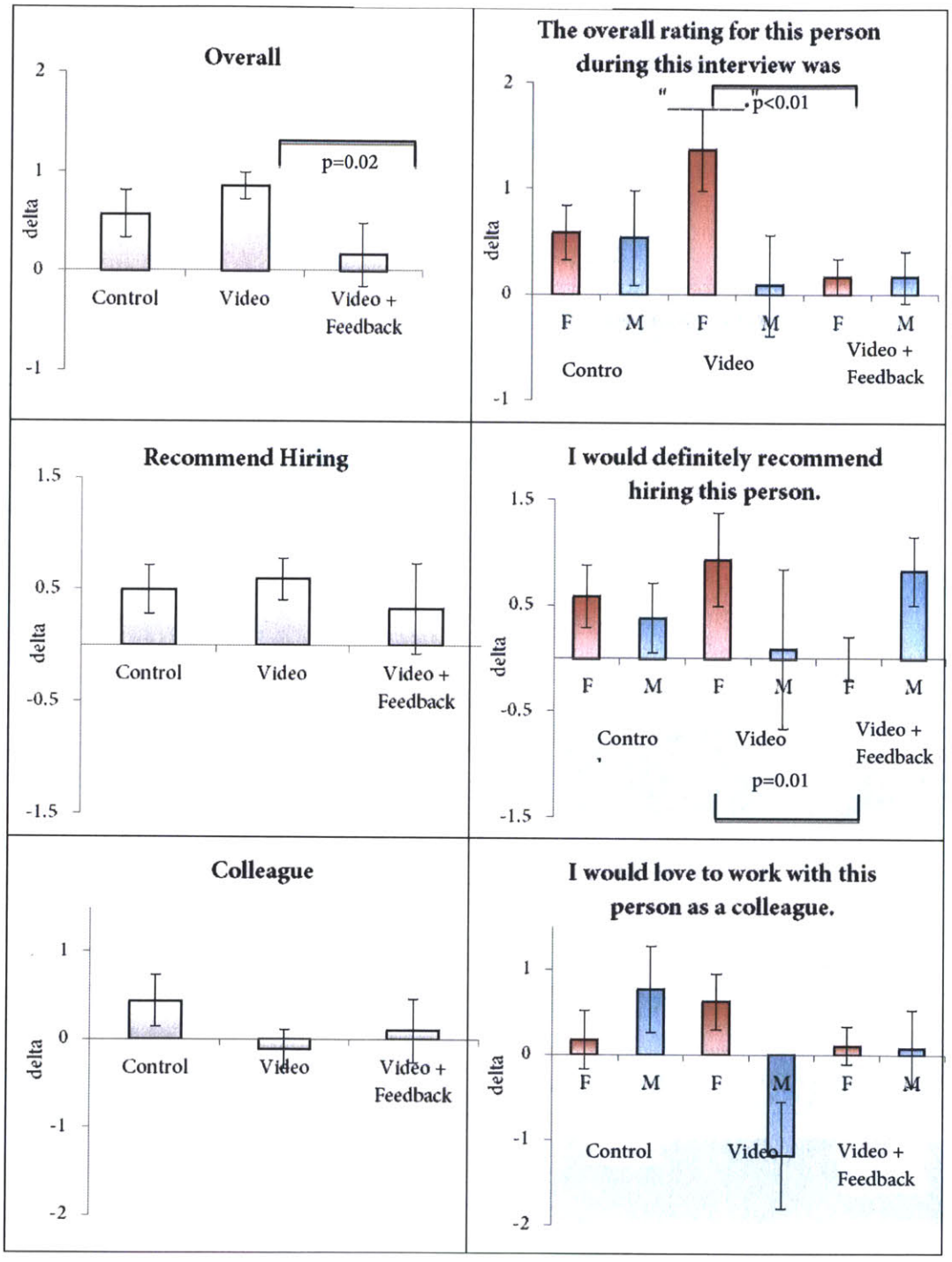
APPENDIX D-2: TWO-WAY ANOVA ANALYSIS BASED ON THE COUNSELOR'S RATINGS (CONDITIONS WITH GENDER)

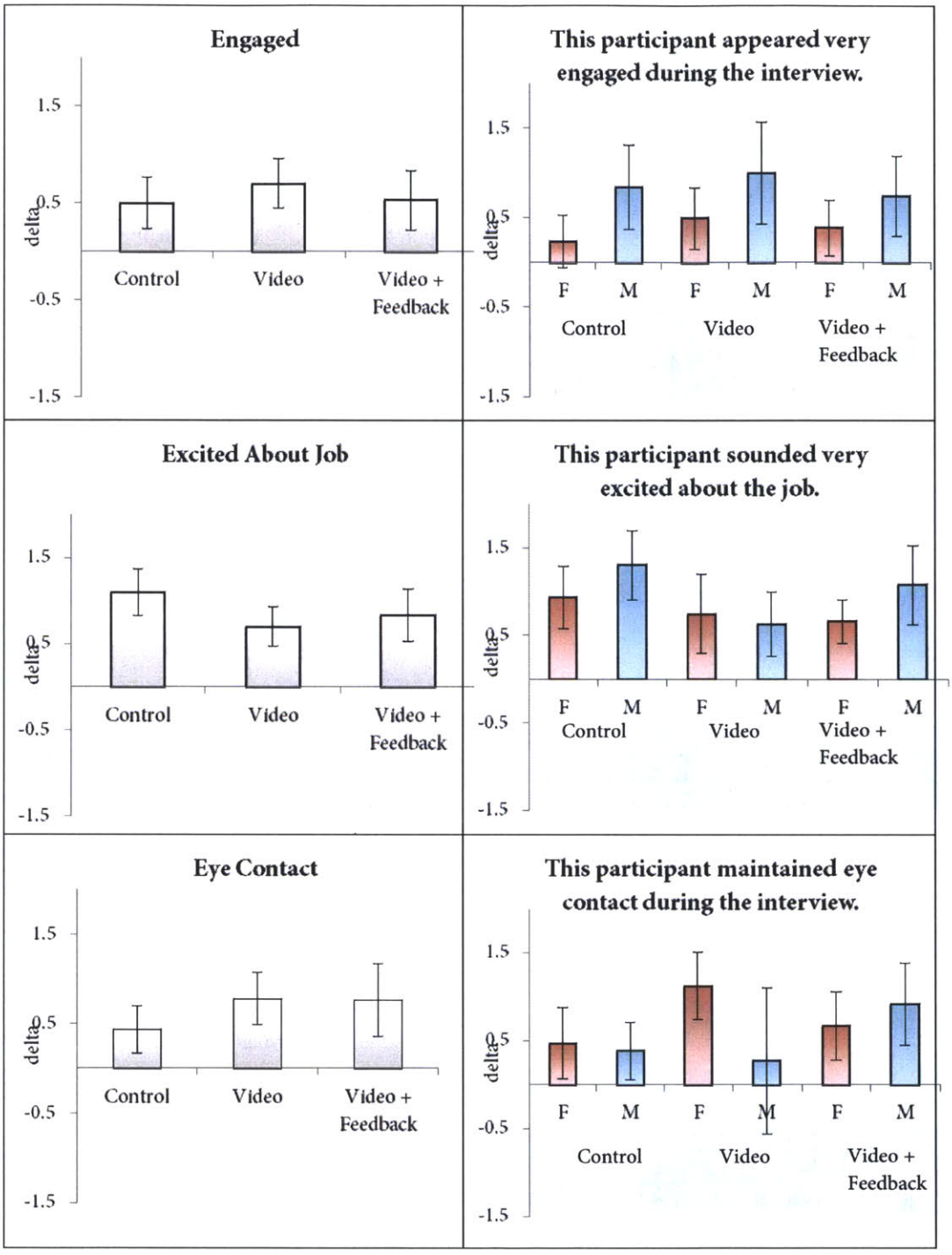
	Effect Test	Female Feedback vs. Video	Female Feedback vs. Control	Female Control vs. Video	Male Feedback vs. Video	Male Feedback vs. Control	Male Control vs. Video
Speaking Rate	F(2,82)= 0.3896, p=0.68	F(1,82)=0.0 443, p=0.81	F(1,82)- 0.0004, p=0.98	F(1,82) =0.0450, p=0.83	F(1,82)=0. 0209, p=0.89	F(1,82)=0. 6885, p=0.41	F(1,82)=0 .8815, p=0.35
Filler	F(2,82)= 1.4288, p=0.25	F(1,82)=0.0 330, p=0.86	F(1,82)=0. 3516, p=0.55	F(1,82) =0.1572, p=0.69	F(1,82)=4. 5798, p=0.04	F(1,82)=0. 3875, p=0.5354	F(1,82)=2 .3842, p=0.13
Friendly	F(2,82)= 1.1033, p=0.33	F(1,82)=0.0 381, p=0.85	F(1,82)=1. 2787, p=0.26	F(1,82) =0.8198, p=0.37	F(1,82)=0. 6635, p=0.42	F(1,82)=0. 1117, p=0.74	F(1,82)=1 .2872, p=0.26
Paused	F(2,82)= 1.5920, p=0.21	F(1,82)=0.3 143, p=0.58	F(1,82)=0. 4812, p=0.49	F(1,82) =0.0145, p=0.90	F(1,82)=0. 0380, p=0.85	F(1,82)=2. 9124, p=0.09	F(1,82)=2 .0706, p=0.15
Engaging Tone	F(2,82)= 0.0854, p=0.92	F(1,82)=2.1 230, p=0.15	F(1,82)=0. 2929, p=0.59	F(1,82) =0.8314, p=0.36	F(1,82)=0. 8192, p=0.37	F(1,82)=0. 4915, p=0.49	F(1,82)=0 .0547, p=0.82
Overall	F(2,82)= 2.6925, p=0.07	F(1,82)=9.7 962, p=0.01	F(1,82)=1. 3083, p=0.26	F(1,82) =3.9079, p=0.061	F(1,82)=0. 0008, p=0.98	F(1,82)=1. 2389, p=0.27	F(1,82)=1 .0774, p=0.20
Hiring	F(2,82)= 2.8169, p=0.07	F(1,82)=6.3 230, p=0.01	F(1,82)=2. 2292, p=0.14	F(1,82) =1.0625, p=0.30	F(1,82)=0. 9642, p=0.33	F(1,82)=0. 1549, p=0.69	F(1,82)=0 .3663, p=0.55
Feedback	F(2,82)= 0.7314, p=0.48	F(1,82)=0.3 712, p=0.54	F(1,82)=0. 0136, p=0.91	F(1,82) =0.2380, p=0.63	F(1,82)=0. 0335, p=0.86	F(1,82)=1. 7261, p=0.19	F(1,82)=1 .1554, p=0.29
Stressed	F(2,82)= 0.4722, p=0.63	F(1,82)=0.9 415, p=0.33	F(1,82)=0. 4192, p=0.52	F(1,82) =0.1079, p=0.74	F(1,82)=0. 1570, p=0.69	F(1,82)=0. 0963, p=0.76	F(1,82)=0 .5808, p=0.49
Engaged	F(2,82)- 0.0367, p=0.96	F(1,82)=0.7 850, p=0.38	F(1,82)=0. 2309, p=0.63	F(1,82) =0.1660, p=0.68	F(1,82)=0. 2415, p=0.62	F(1,82)=0. 2634, p=0.61	F(1,82)=0 .0000, p=1.00
Excited about Job	F(2,82)= 0.6764, p=0.51	F(1,82)=0.7 475, p=0.39	F(1,82)=0. 7846, p=0.38	F(1,82) =0.0001, p=0.99	F(1,82)=0. 1209, p=0.73	F(1,82)=1. 4364, p=0.23	F(1,82)=2 .2356, p=0.14
Eye Contact	F(2,82)= 0.5959, p=0.553 4	F(1,82)=0.4 909, p=0.49	F(1,82)=0. 0926, p=0.76	F(1,82) =0.9728, p=0.33	F(1,82)=0. 4048, p=0.53	F(1,82)=0. 0424, p=0.84	F(1,82)=0 .1928, p=0.66
Colleague	F(2,82)= 0.5974, p=0.55	F(1,82)=0.7 370, p=0.39	F(1,82)=0. 0405, p=0.84	F(1,82) =0.4243, p=0.52	F(1,82)=0. 0367, p=0.85	F(1,82)=0. 5003, p=0.48	F(1,82)=0 .7547, p=0.39
Smile	F(2,82)=	F(1,82)=0.1	F(1,82)=0.	F(1,82)	F(1,82)=0.	F(1,82)=1.	F(1,82)=0

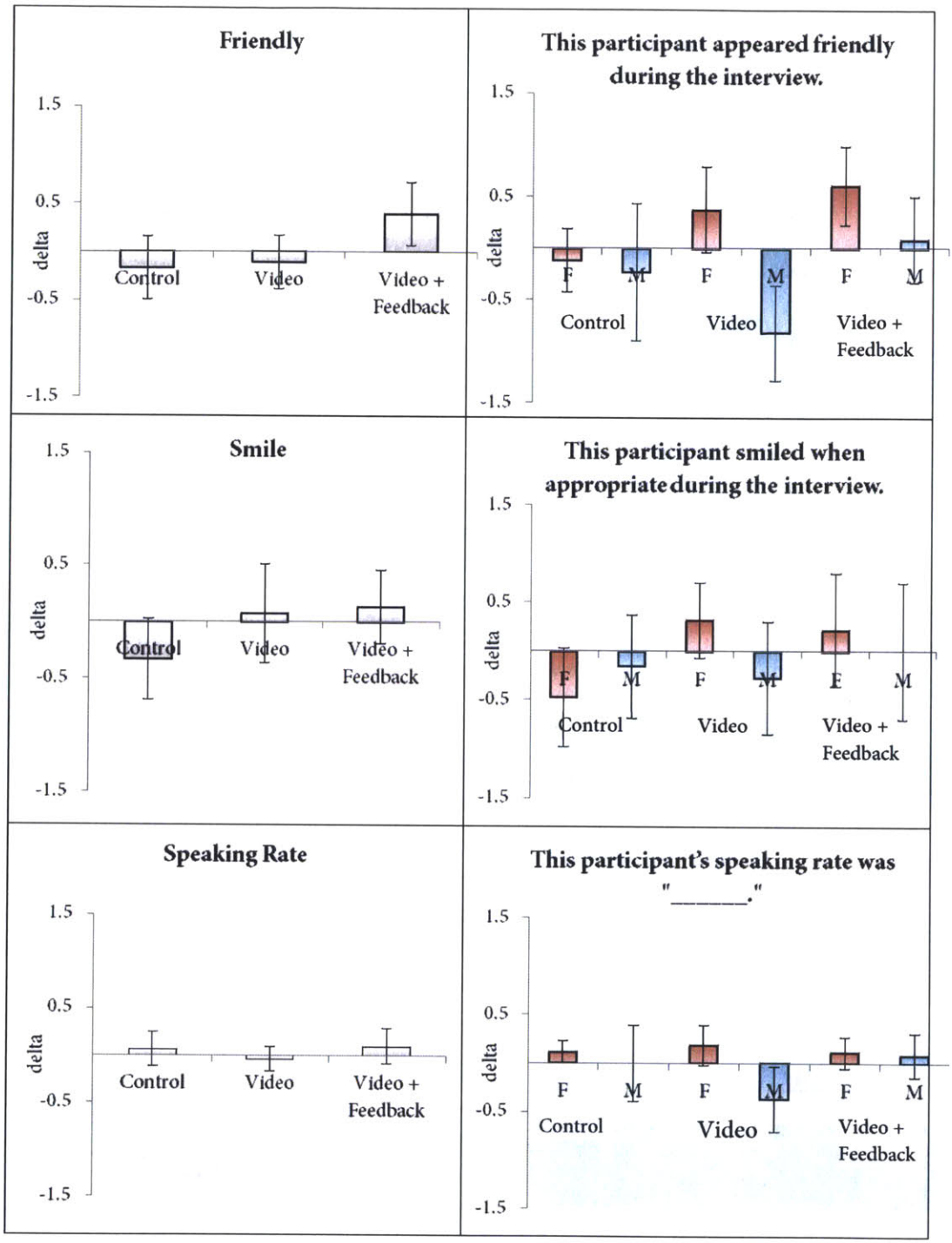
	0.2493, p=0.78	556, p=0.69	2870, p=0.59	=0.0172, p=0.90	5065, p=0.48	9372, p=0.17	.3855, p=0.54
Not Calm	F(2,82)= 1.6242, p=0.20	F(1,82)=4.5674, p =0.04	F(1,82)=2.6479, p=0.11	F(1,82) =0.2790, p=0.60	F(1,82)=0.1550, p=0.69	F(1,82)=0.1821, p=0.67	F(1,82)=0.0002, p=0.99
Focused	F(2.82)= 0.2370, p=0.79	F(1.82)=0.0269, p=0.87	F(1,82)=0.4818, p=0.49	F(1,82) =0.6983, p=0.41	F(1,82)=0.2129, p=0.65	F(1,82)=0.0918, p=0.76	F(1,82)=0.0294, p=0.86

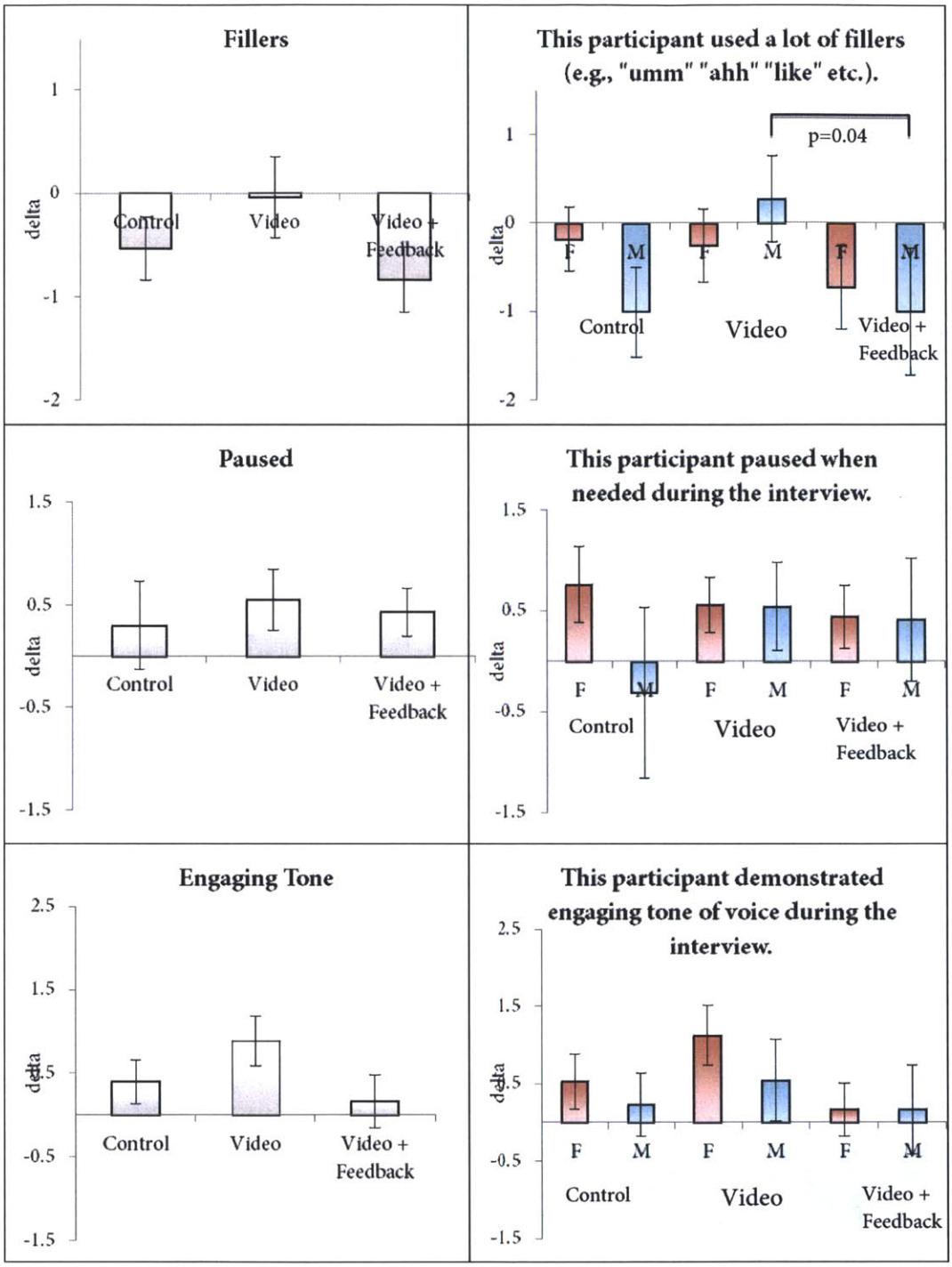
APPENDIX D-3: GRAPHICAL ANALYSIS BASED ON THE COUNSELOR'S RATINGS

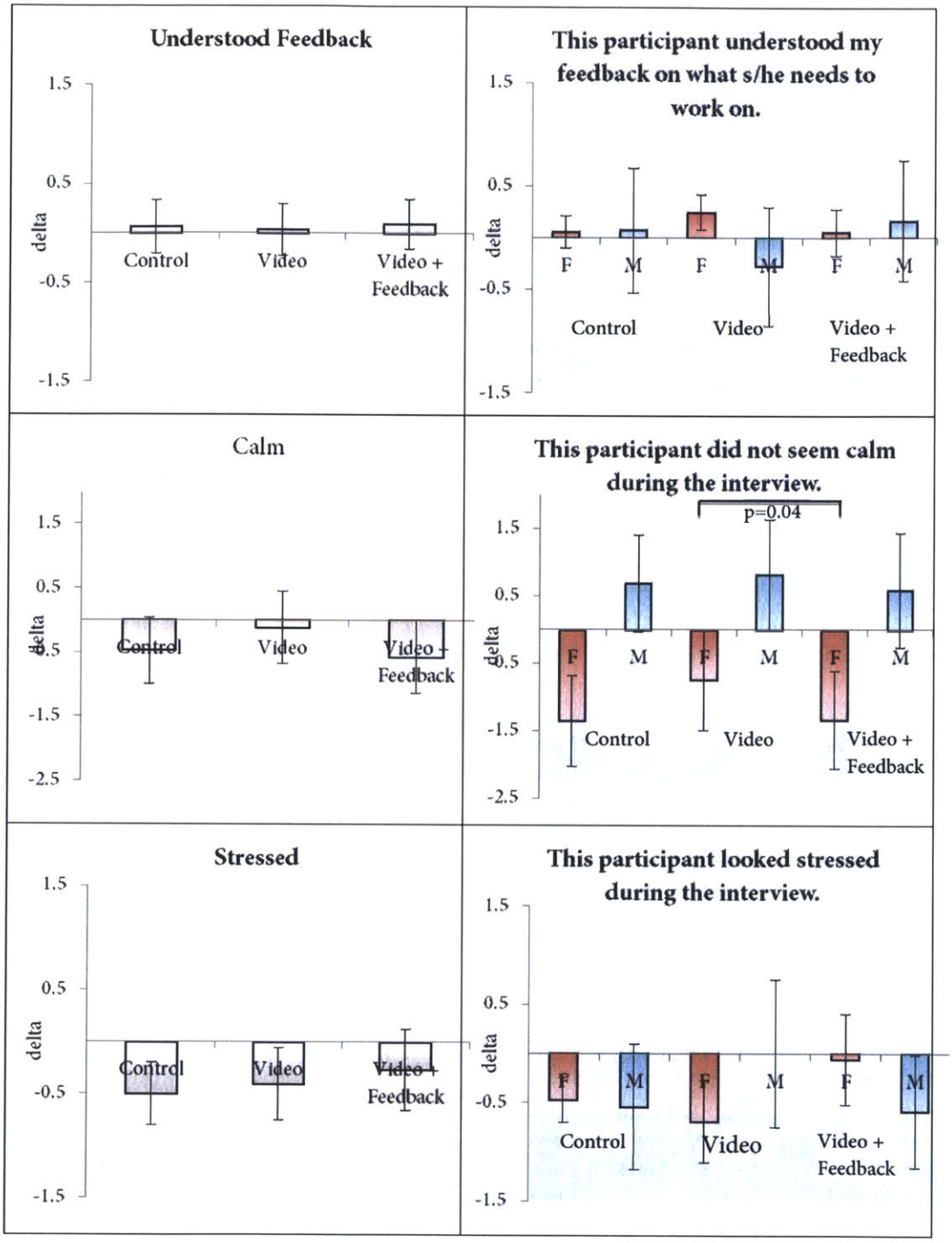
Original Counselor




Engaged
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant appeared very engaged during the interview.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Excited About Job
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant sounded very excited about the job.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Eye Contact
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant maintained eye contact during the interview.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback


Friendly
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant appeared friendly during the interview.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Smile
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant smiled when appropriate during the interview.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Speaking Rate
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant's speaking rate was "_____."
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback


Fillers
delta
1
0
-1
-2
Control
Video
Video + Feedback
This participant used a lot of fillers (e.g., "umm" "ahh" "like" etc.).
p=0.04
F
M
Control
Video
Video + Feedback
Paused
1.5
0.5
-0.5
-1.5
This participant paused when needed during the interview.
Engaging Tone
2.5
1.5
This participant demonstrated engaging tone of voice during the interview.


Understood Feedback
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant understood my feedback on what s/he needs to work on.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Calm
1.5
0.5
-0.5
-1.5
-2.5
delta
Control
Video
Video + Feedback
This participant did not seem calm during the interview.
p=0.04
1.5
0.5
-0.5
-1.5
-2.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
Stressed
1.5
0.5
-0.5
-1.5
delta
Control
Video
Video + Feedback
This participant looked stressed during the interview.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback

APPENDIX E-1: ONE-WAY ANOVA ANALYSIS BASED ON THE OTHER COUNSELOR'S RATINGS (CONDITIONS)

OTHER COUNSELORS (conditions)

	Effect Test	Feedback vs. Video	Feedback vs. Control	Control vs. Video
Overall	**F(2,83)=4.8940, p=0.01**	**F(1,83)=6.9156, p=0.01**	**F(1,83)=7.4593, p=0.01**	F(1,83)=0.0005, p=0.98
Recommend Hiring	F(2,83)=2.9876, p=0.06	**F(1,83)=5.8699, p=0.02**	F(1,83)=2.0283, p=0.16	F(1,83)=1.0471, p=0.31
Colleague	**F(2,83)=6.6688, p<0.01**	**F(1,83)=12.4701,p < 0.01**	**F(1,83)= 6.1526, p=0.02**	F(1,83)=1.2197, p=0.27
Engagement	F(2,83)=0.1417, p=0.87	F(1,83)=0.2599, p=0.61	F(1,83)=0.1404, p=0.71	F(1,83)=0.0206, p=0.87
Excitement	**F(2,83)=3.2414, p=0.04**	**F(1,83)=5.7371, p=0.02**	F(1,83)=0.3063, p=0.58	F(1,83)=3.5609, p=0.06
Eye Contact	F(2,83)=1.5832, p=0.21	F(1,83)=0.1316, p=0.72	F(1,83)=1.8707, p=0.18	F(1,83)=2.7739, p=0.10
Smile	F(2,83)=1.7226, p=0.18	F(1,83)=3.3257, p=0.07	F(1,83)=0.3038, p=0.58	F(1,83)=1.6089, p=0.21
Speaking rate	F(2,83)=0.7568, p=0.47	F(1,83)=1.5068, p=0.22	F(1,83)=0.2535, p=0.62	F(1,83)=0.5286, p=0.47
Filler	F(2,83)=0.1840, p=0.83	F(1,83)=0.0021, p=0.96	F(1,83)=0.3086, p=0.58	F(1,83)=0.2366, p=0.63
Friendly	F(2,83)=0.1222, p=0.89	F(1,83)=0.0335, p=0.86	F(1,83)=0.1021, p=0.75	F(1,83)=0.2360, p=0.63
Pauses	F(2,83)=1.9776, p=0.14	F(1,83)=1.6224, p=0.21	F(1,83)=3.7892, p=0.06	F(1,83)=0.3696, p=0.55
Engaging Tone	F(2,83)=0.0843, p=0.92	F(1,83)=0.0322, p=0.86	F(1.83)=0.0585, p=0.81	F(1,83)=0.1665, p=0.68
Structured	F(2,83)=0.7796, p=0.46	F(1,83)=0.1771, p=0.68	F(1,83)=0.7062, p=0.40	F(1,83)=1.4841, p=0.23
Not calm	F(2,83)=0.0630, p=0.94	F(1,83)=0.0759, p=0.78	F(1,83)=0.1069, p=0.74	F(1,83)=0.0017, p=0.97
Stressed	F(2,83)=0.0510, p=0.95	F(1,83)=0.1007, p=0.75	F(1,83)=0.0137, p=0.91	F(1,83)=0.0401, p=0.84
Focused	F(2,83)=0.1813, p=0.83	F(1,83)=0.0856, p=0.77	F(1,83)=0.3625, p=0.55	F(1,83)=0.0828, p=0.77
Authentic	F(2,83)=0.2380, p=0.79	F(1,83)=0.4197, p=0.52	F(1,83)=0.2639, p=0.61	F(1,83)=0.0215, p=0.88
Awkward	F(2,83)=0.5989, p=0.55	F(1,83)=0.4245, p=0.52	F(1,83)=1.1715, p=0.28	F(1,83)=0.1551, p=0.69

APPENDIX E-2: TWO-WAY ANOVA ANALYSIS BASED ON THE OTHER COUNSELOR'S RATINGS (CONDITIONS WITH GENDER)

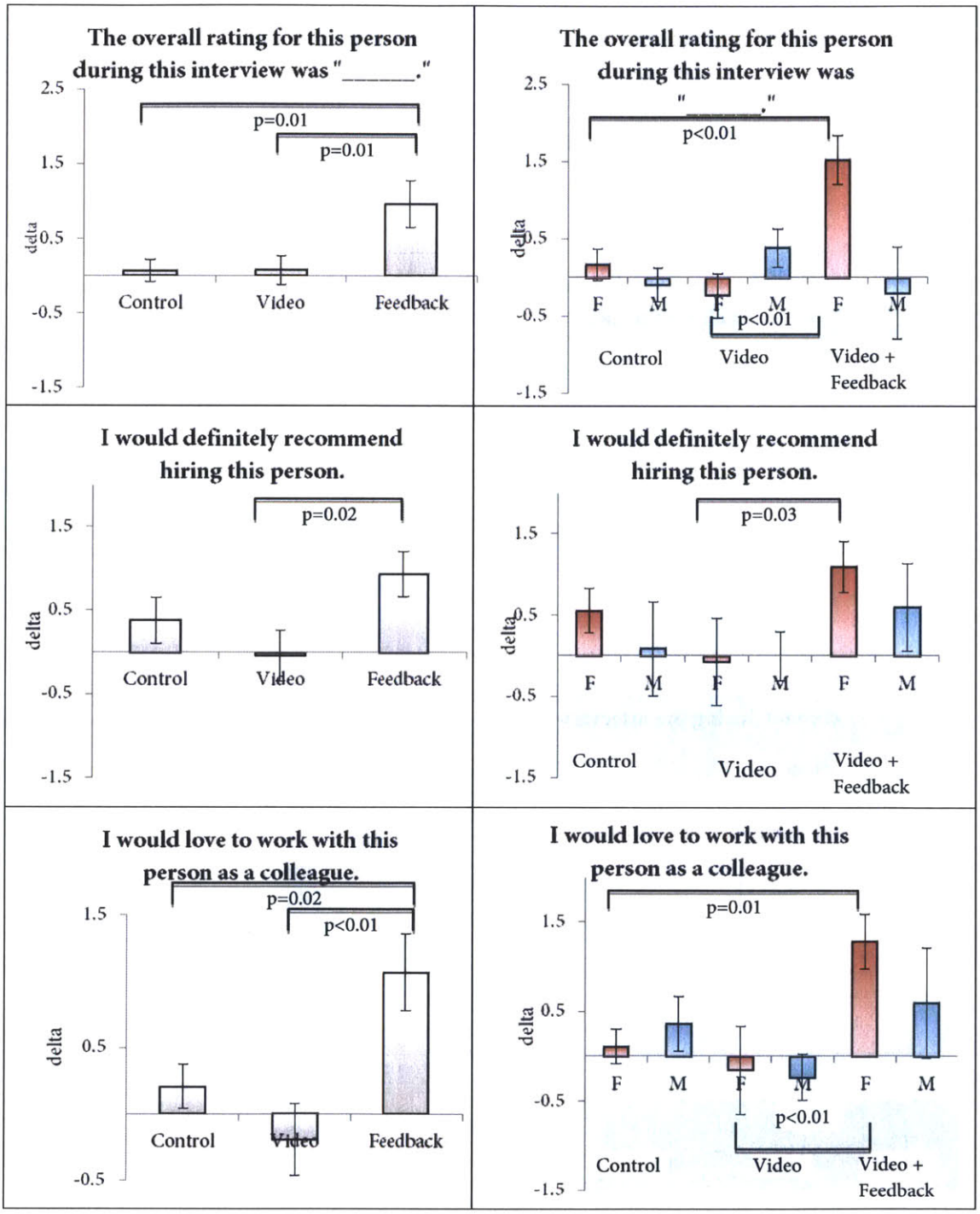
OTHER COUNSELORS (conditions and gender)

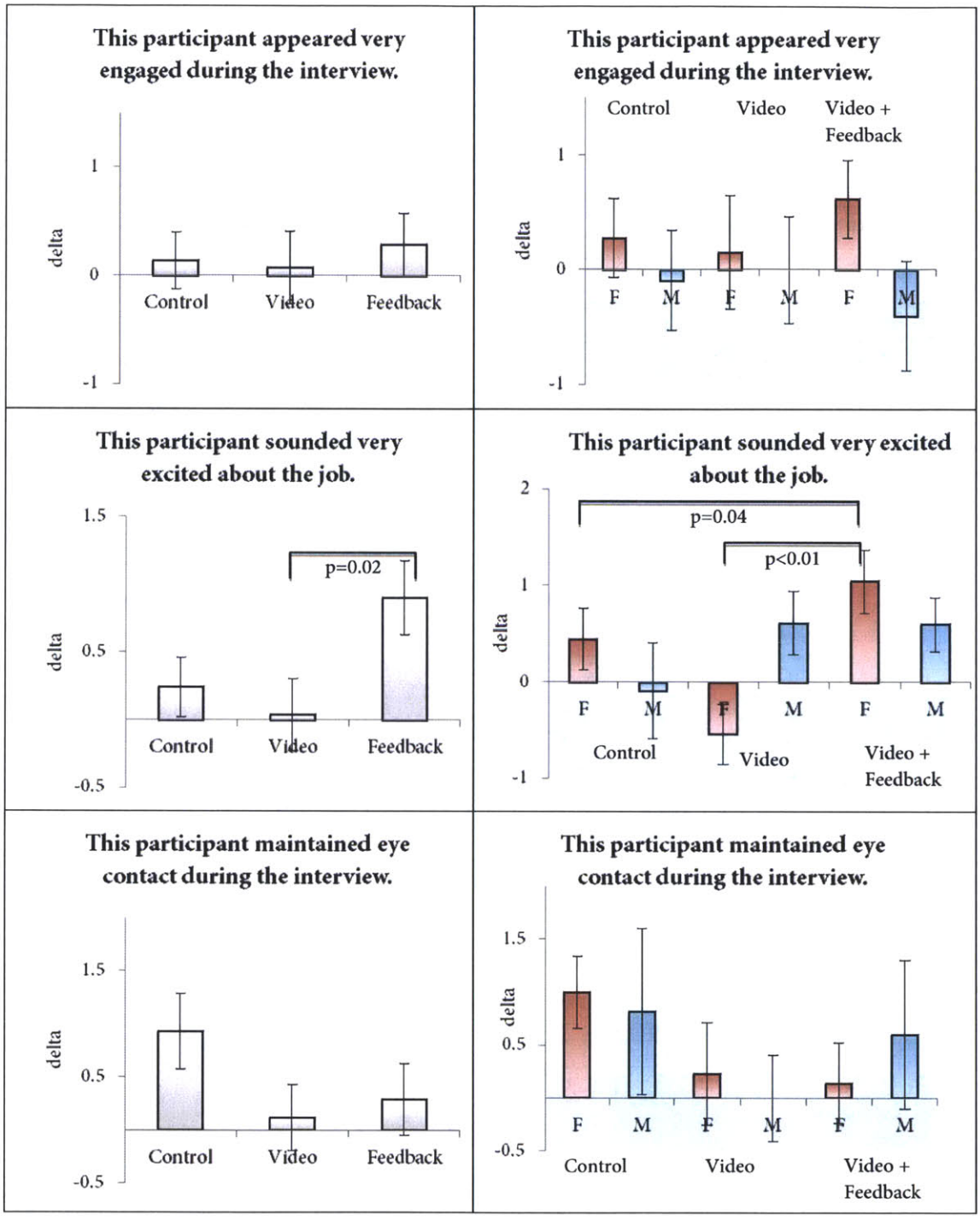
	Effect Test	Female Feedback vs. Video	Female Feedback vs. Control	Female Control vs. Video	Male Feedback vs. Video	Male Feedback vs. Control	Male Control vs. Video
Overall	**F(2,80)=6.6701, p<0.01**	**F(1,80)=7.7146, p=0.01**	**F(1,80)=12.7932, p=0.01**	F(1,80)=0.8545, p=0.36	F(1,80)=1.3844, p=0.24	F(1,80)=0.0447, p=0.83	F(1,80)=0.9655, p=0.33
Recommend Hiring	F(2,80)=0.2932, p=0.75	**F(1,80)=4.7326, p=0.03**	F(1,80)=1.2110, p=0.27	F(1,80)=1.2954, p=0.26	F(1,80)=0.8729, p=0.35	F(1,80)=1.2110, p=0.27	F(1,80)=0.0211, p=0.88
Colleague	F(2,80)=0.8487, p=0.43	**F(1,80)=9.1799, p=0.001**	**F(1,80)=7.3774, p=0.01**	F(1,80)=0.2923, p=0.59	F(1,80)=2.1521, p=0.15	F(1,80)=0.1614, p=0.69	F(1,80)=1.1614, p=0.28
Engagement	F(2,80)=2.1483, p=0.15	F(1,80)=0.7036, p=0.40	F(1,80)=0.4571, p=0.50	F(1,80)=0.0469, p=0.83	F(1,80)=0.3662, p=0.55	F(1,80)=0.2026, p=0.65	F(1,80)=0.0199, p=0.89
Excitement	**F(2,80)=3.3855, p=0.04**	**F(1,80)=11.4594, p=0.01**	F(1,80)=2.0005, p=0.16	**F(1,80)=4.1372, p=0.0453**	F(1,80)=0.0008, p=0.98	F(1,80)=1.4185, p=0.24	F(1,80)=1.6863, p=0.20
Eye Contact	F(2,80)=0.2925, p=0.75	F(1,80)=0.0183, p=0.89	F(1,80)=2.1028, p=0.15	F(1,80)=1.3189, p=0.25	F(1,80)=0.6009, p=0.44	F(1,80)=0.0736, p=0.77	F(1,80)=1.1778, p=0.28
Smile	F(2,80)=0.1747, p=0.84	F(1,80)=2.6565, p=0.11	F(1,80)=0.1379, p=0.71	F(1,80)=1.5691, p=0.21	F(1,80)=0.5303, p=0.47	F(1,80)=0.1078, p=0.74	F(1,80)=0.1581, p=0.69
Speaking Rate	F(2,80)=2.4452, p=0.09	F(1,80)=0.2971, p=0.59	F(1,80)=2.1391, p=0.15	F(1,80)=0.5810, p=0.45	F(1,80)=0.9905, p=0.32	F(1.80)=1.0263, p=0.31	**F(1,80)=4.4198, p=0.04**

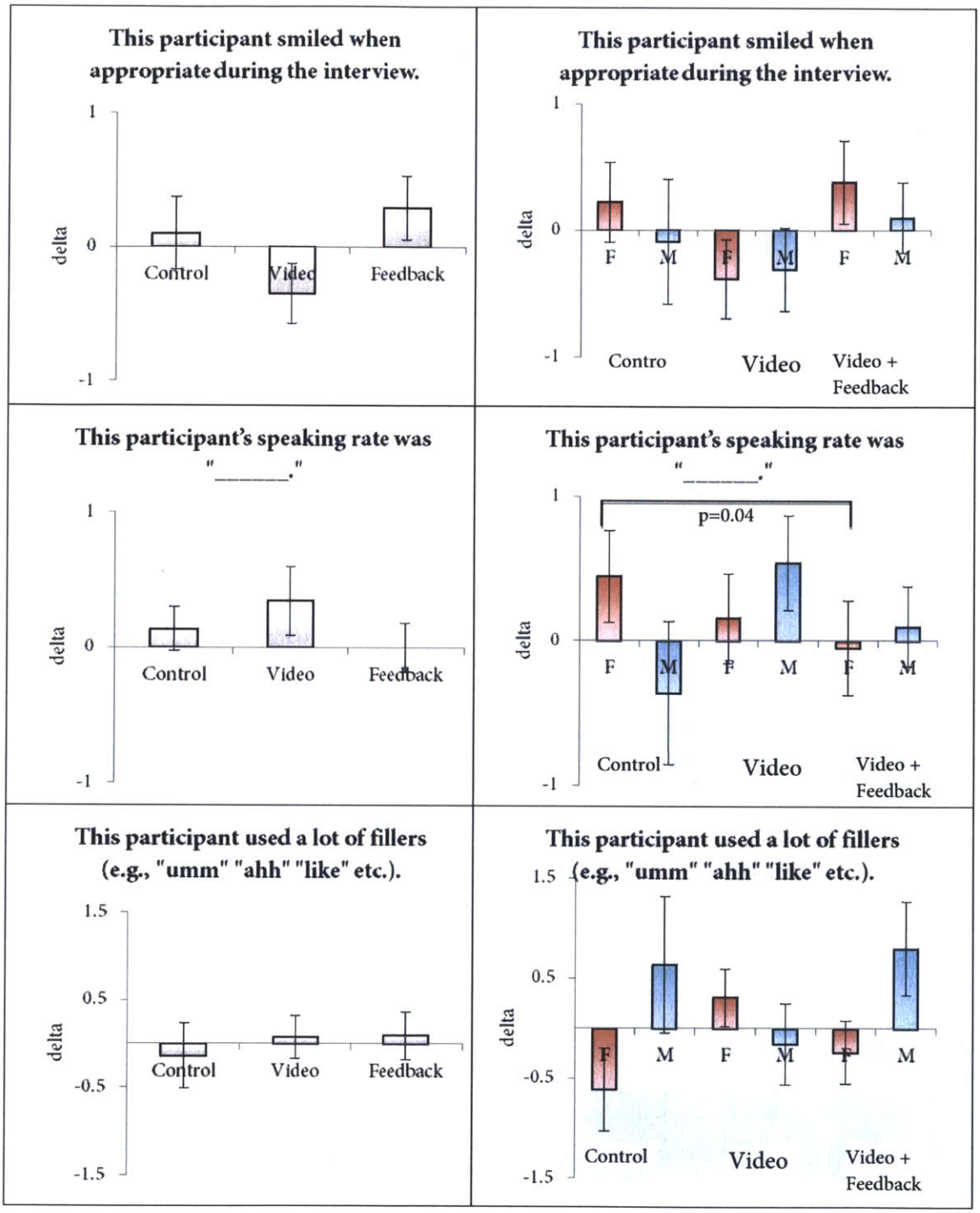
Filler	F(2,80)=2.2645, p=0.11	F(1,80)=0.9440, p=0.33	F(1,80)=0.5323, p=0.47	F(1,80)=2.5151, p=0.12	F(1,80)=2.0296, p=0.16	F(1,80)=0.0554, p=0.81	F(1,80)=1.4684, p=0.23
Friendly	F(2,80)=1.3953, p=0.25	F(1,80)=0.1335, p=0.72	F(1,80)=1.3612, p=0.25	F(1,80)=0.4561, p=0.50	F(1,80)=1.3866, p=0.24	F(1,80)=1.3481, p=0.25	F(1,80)=0.0009, p=0.98
Pauses	F(2,80)=0.5990, p=0.55	F(1,80)=0.0940, p=0.76	F(1,80)=1.3942, p=0.24	F(1,80)=0.5546, p=0.46	F(1,80)=2.7837, p=0.10	F(1,80)=2.9495, p=0.09	F(1,80)=0.0141, p=0.91
Engaging Tone	F(2,80)=3.0499, p=0.053	F(1,80)=2.0728, p=0.15	F(1,80)=0.3374, p=0.56	F(1,80)=0.7802, p=0.38	F(1,80)=0.7989, p=0.05	F(1,80)=0.2722, p=0.60	F(1,80)=2.0872, p=0.15
Structured	**F(2,80)=5.1724, p=0.01**	F(1,80)=3.2871, p=0.07	F(1,80)=1.0581, p=0.31	F(1,80)=0.7227, p=0.40	F(1,80)=3.1953, p=0.08	**F(1,80)=8.6293, p=0.01**	F(1,80)=1.6840, p=0.20
Not Calm	F(2,80)=0.0591, p=0.94	F(1,80)=0.0010, p=0.98	F(1,80)=0.1462, p=0.70	F(1,80)=0.0945, p=0.76	F(1,80)=0.0143, p=0.91	F(1,80)=0.0014, p=0.97	F(1,80)=0.0264, p=0.87
Stressed	F(2,80)=1.2346, p=0.30	F(1.80)=0.6648, p=0.42	F(1,80)=0.1670, p=0.68	F(1.80)=1.3252, p=0.25	F(1,80)=0.1597, p=0.69	F(1.80)=0.3928, p=0.53	F(1,80)=1.1637, p=0.28
Focused	F(2,80)=1.3257, p=0.27	F(1,80)=0.1780, p=0.67	F(1,80)=0.0063, p=0..94	F(1,80)=0.1150, p=0.74	F(1,80)=2.7649, p=0.10	F(1,80)=2.0928, p=0.15	F(1,80)=0.0270, p=0.87
Authentic	F(2,80)=1.8429, p=0.17	F(1,80)=2.7393, p=0.10	F(1.80)=0.8018, p=0.37	F(1,80)=0.6634, p=0.42	F(1,80)=1.2481, p=0.27	F(1,80)=0.2455, p=0.62	F(1,80)=0.3827, p=0.54
Awkward	F(2,80)=1.0277, p=0.36	F(1.,80)=1.8992, p=0.17	F(1,80)=2.2387, p=0.14	F(1,80)=0.0002, p=0.99	F(1,80)=0.3785, p=0.54	F(1,80)=0.0544, p=0.82	F(1,80)=0.1456, p=0.70

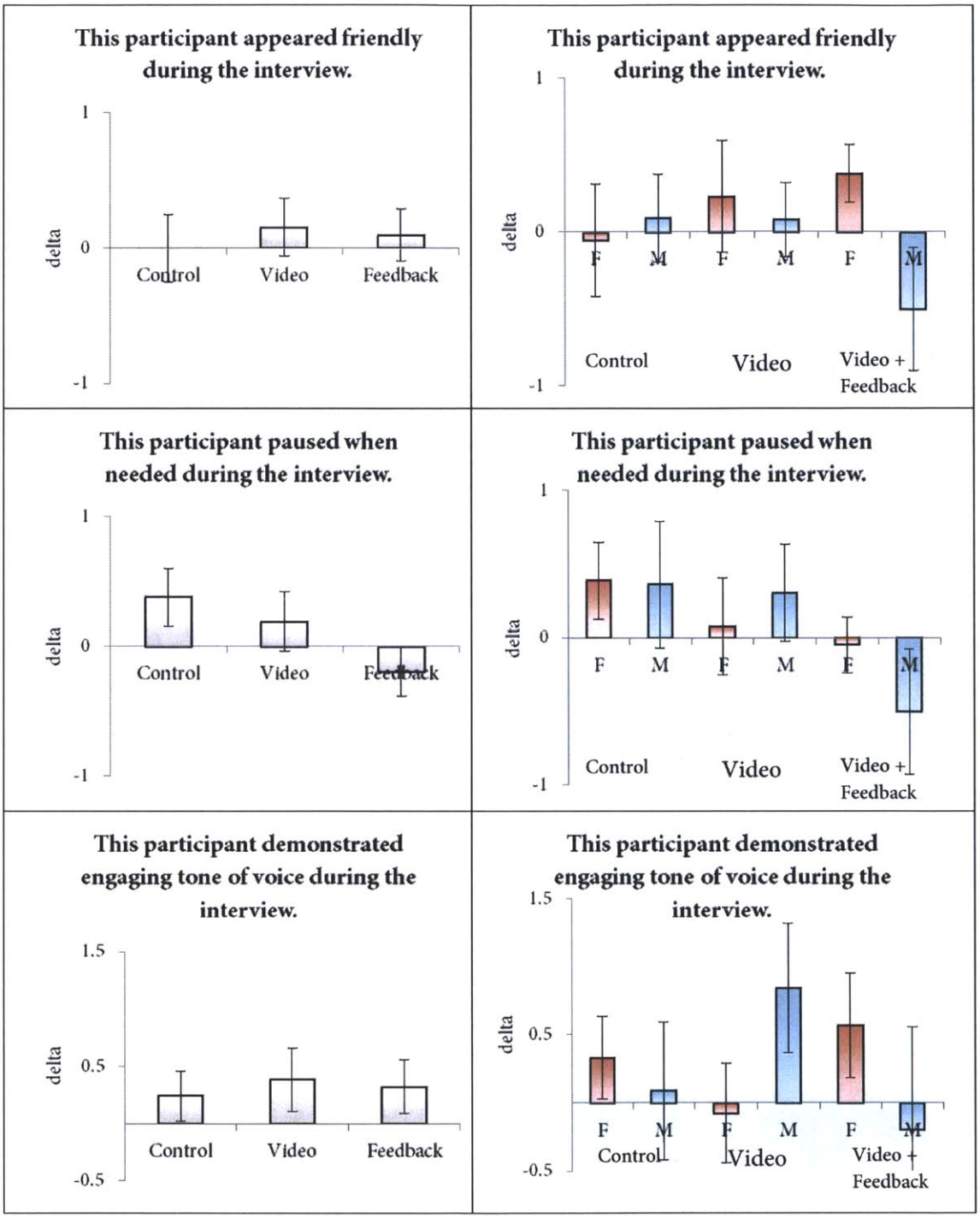
APPENDIX E-3: GRAPHICAL ANALYSIS BASED ON THE OTHER COUNSELOR'S RATINGS

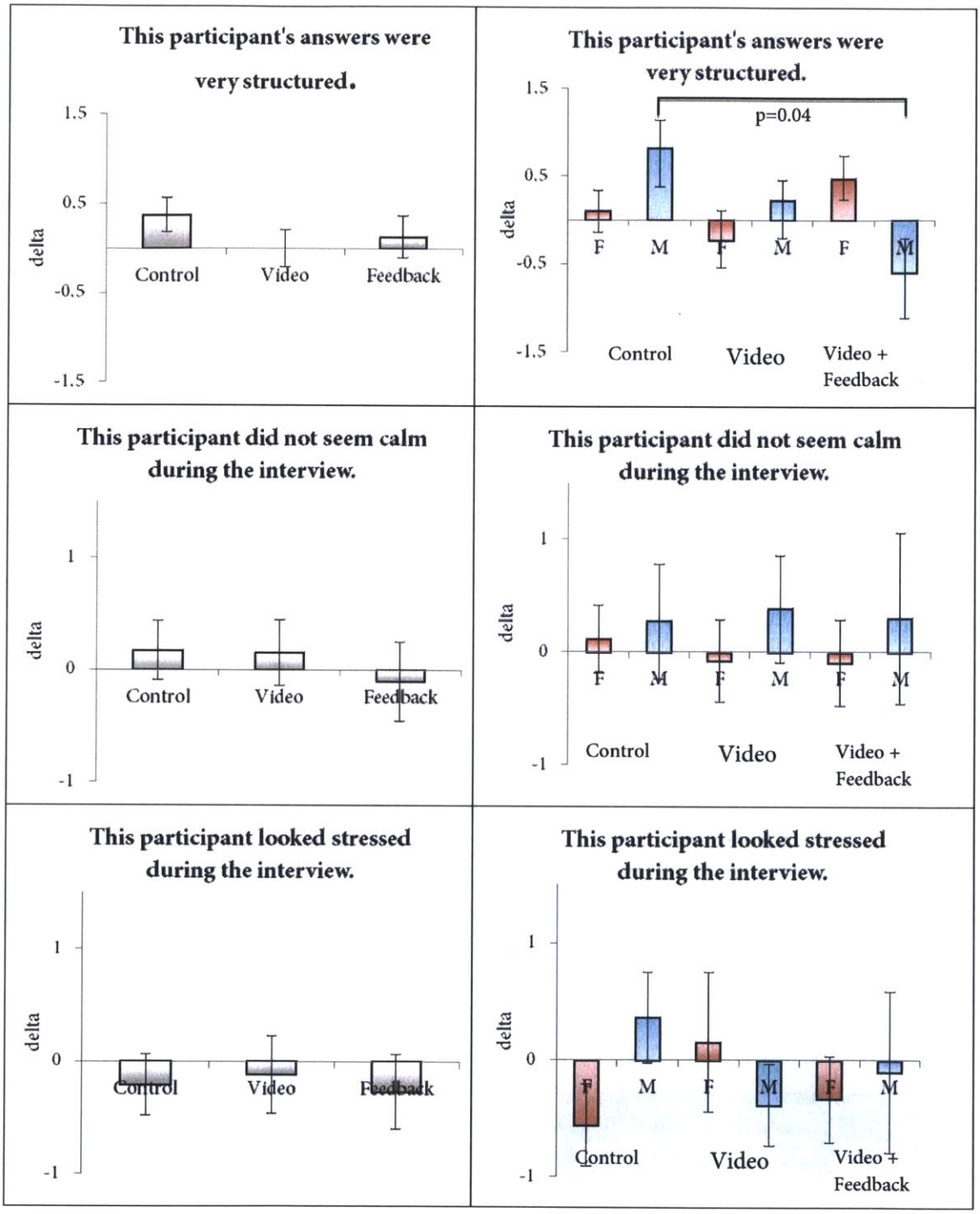
Other Counselors

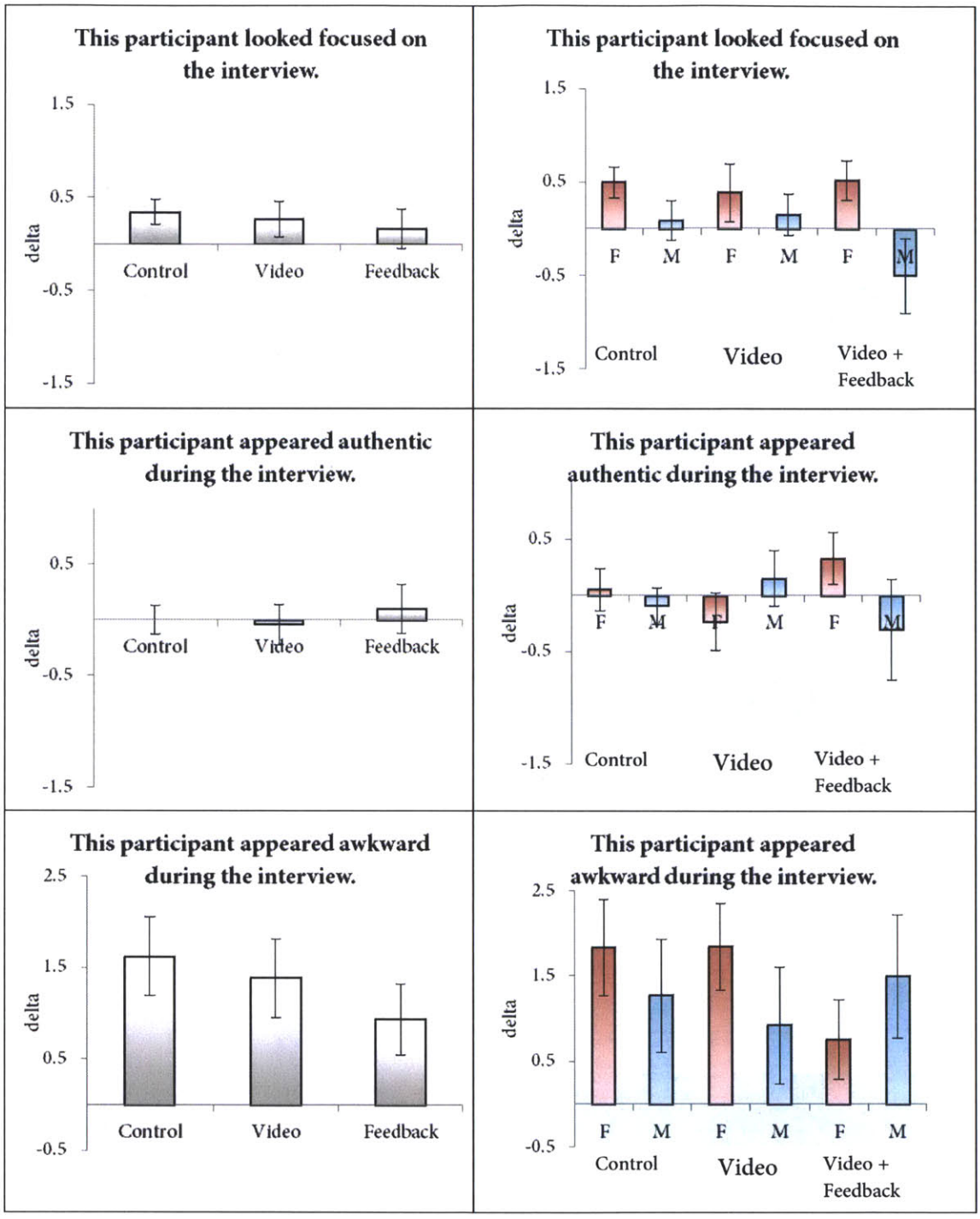



This participant appeared very engaged during the interview.
delta
1
0
-1
Control
Video
Feedback
This participant appeared very engaged during the interview.
Control
Video
Video + Feedback
delta
1
0
-1
F
M
F
M
F
M
This participant sounded very excited about the job.
delta
1.5
0.5
-0.5
p=0.02
Control
Video
Feedback
This participant sounded very excited about the job.
delta
2
1
0
-1
p=0.04
p<0.01
F
M
F
M
F
M
Control
Video
Video + Feedback
This participant maintained eye contact during the interview.
delta
1.5
0.5
-0.5
Control
Video
Feedback
This participant maintained eye contact during the interview.
delta
1.5
0.5
-0.5
F
M
F
M
F
M
Control
Video
Video + Feedback


This participant smiled when appropriate during the interview.
delta
1
0
-1
Control
Video
Feedback
This participant smiled when appropriate during the interview.
delta
1
0
-1
F
M
F
M
F
M
Contro
Video
Video + Feedback
This participant's speaking rate was "_____."
delta
1
0
-1
Control
Video
Feedback
This participant's speaking rate was "_____."
p=0.04
delta
1
0
-1
F
M
F
M
F
M
Control
Video
Video + Feedback
This participant used a lot of fillers (e.g., "umm" "ahh" "like" etc.).
delta
1.5
0.5
-0.5
-1.5
Control
Video
Feedback
This participant used a lot of fillers (e.g., "umm" "ahh" "like" etc.).
delta
1.5
0.5
-0.5
-1.5
F
M
F
M
F
M
Control
Video
Video + Feedback


This participant appeared friendly during the interview.
delta
1
0
-1
Control
Video
Feedback
This participant appeared friendly during the interview.
delta
1
0
-1
F
M
F
M
F
M
Control
Video
Video + Feedback
This participant paused when needed during the interview.
delta
1
0
-1
Control
Video
Feedback
This participant paused when needed during the interview.
delta
1
0
-1
F
M
F
M
F
M
Control
Video
Video + Feedback
This participant demonstrated engaging tone of voice during the interview.
delta
1.5
0.5
-0.5
Control
Video
Feedback
This participant demonstrated engaging tone of voice during the interview.
delta
1.5
0.5
-0.5
F
M
F
M
F
M
Control
Video
Video + Feedback


This participant's answers were very structured.
delta
1.5
0.5
-0.5
-1.5
Control
Video
Feedback
This participant's answers were very structured.
p=0.04
F
M
Control
Video
Video + Feedback
This participant did not seem calm during the interview.
1
0
-1
This participant looked stressed during the interview.


This participant looked focused on the interview.
1.5
0.5
-0.5
-1.5
delta
Control
Video
Feedback
This participant looked focused on the interview.
1.5
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
This participant appeared authentic during the interview.
0.5
-0.5
-1.5
delta
Control
Video
Feedback
This participant appeared authentic during the interview.
0.5
-0.5
-1.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback
This participant appeared awkward during the interview.
2.5
1.5
0.5
-0.5
delta
Control
Video
Feedback
This participant appeared awkward during the interview.
2.5
1.5
0.5
-0.5
delta
F
M
F
M
F
M
Control
Video
Video + Feedback

APPENDIX F-1: ONE-WAY ANOVA ANALYSIS BASED ON THE TURKER'S RATINGS (CONDITIONS)

TURKERS (conditions)

	Effect Test	Feedback vs. Video	Feedback vs. Control	Control vs. Video
Overall	**F(2,66)=4.01, p=0.02**	**F(1,66)=7.07, p=0.01**	**F(1,66)=4.42, p = 0.04**	F(1,66)=0.29, p=0.59
Recommended Hiring	F(2,66)=2.66, p=0.07	**F(1,66)=4.70, p=0.03**	F(1,66)=0.10, p=0.75	F(1,66)=3.23, p=0.08
Colleague	F(2,66)=0.49, p=0.62	F(1,66)=0.11, p=0.74	F(1,66)=0.44, p=0.51	F(1,66)=0.93, p=0.34
Engaged	F(2,66)=0.15, p=0.86	F(1,66)=0.30, p=0.58	F(1,66)=0.07, p=0.79	F(1,66)=0.08, p=0.78
Excited	**F(2,66)=3.09, p=0.05**	**F(1,66)=5.12, p=0.02**	**F(1,66)=3.88, p=0.05**	F(1,66)=0.08, p=0.78
Eye Contact	F(2,66)=0.07, p=0.80	F(1,66)=0.07, p=0.80	F(1,66)=0.27, p=0.61	F(1,66)=0.57, p=0.45
Smiled	F(2,66)=2.62, p=0.08	F(1,66)=0.35, p=0.56	F(1,66)=2.81, p=0.10	**F(1,66) =4.84, p=0.03**
Speaking	F(2,66)=0.50, p=0.61	F(1,66)=0.44, p=0.51	F(1,66)=0.12, p=0.73	F(1,66)=0.96, p=0.33
No Filler	F(2,66)=0.15, p=0.86	F(1,66)=0.30, p=0.59	F(1,66)=0.10, p=0.75	F(1,66)=0.05, p=0.83
Friendly	F(2,66)=0.06, p=0.95	F(1,66)=0.08, p=0.78	F(1,66)=0.00, p=0.97	F(1,66)=0.09, p=0.76
Paused	F(2,66)=0.23, p=0.63	F(1,66)=0.23, p=0.63	F(1,66)=0.25, p=0.63	F(1,66)=0.90, p=0.35
Engaging	F(2,66)=0.88, p=0.42	F(1,66)=0.00, p=0.95	F(1,66)=1.30, p=0.26	F(1,66)=1.37, p=0.25
Structure	F(2,66)=1.00, p=0.37	F(1,66)=1.89, p=0.17	F(1,66)=0.90, p=0.35	F(1,66)=0.17, p=0.68
Calm	F(2.66)=0.16, p=0.85	F(1,66)=0.31, p=0.58	F(1,66)=0.02, p=0.89	F(1,66)=0.16, p=0.69
Not Stressed	F(2,66)=1.00, p=0.37	F(1,66)=1.44, p=0.23	F(1,66)=0.01, p=0.92	F(1,66)=0.16, p=0.21
Focused	F(2,66)=1.22, p=0.30	F(1,66)=0.40, p=0.53	F(1,66)=0.94, p=0.34	F(1,66)=2.40, p=0.13
Authentic	F(2,66)=0.34, p=0.71	F(1,66)=0.68, p=0.41	F(1,66)=0.11, p=0.74	F(1,66)=0.23, p=0.63
Not Awkward	F(2,66)=1.74, p=0.18	F(1,66)=3.36, p=0.07	F(1,66)=1.35, p=0.25	F(1,66)=0.43, p-0.52

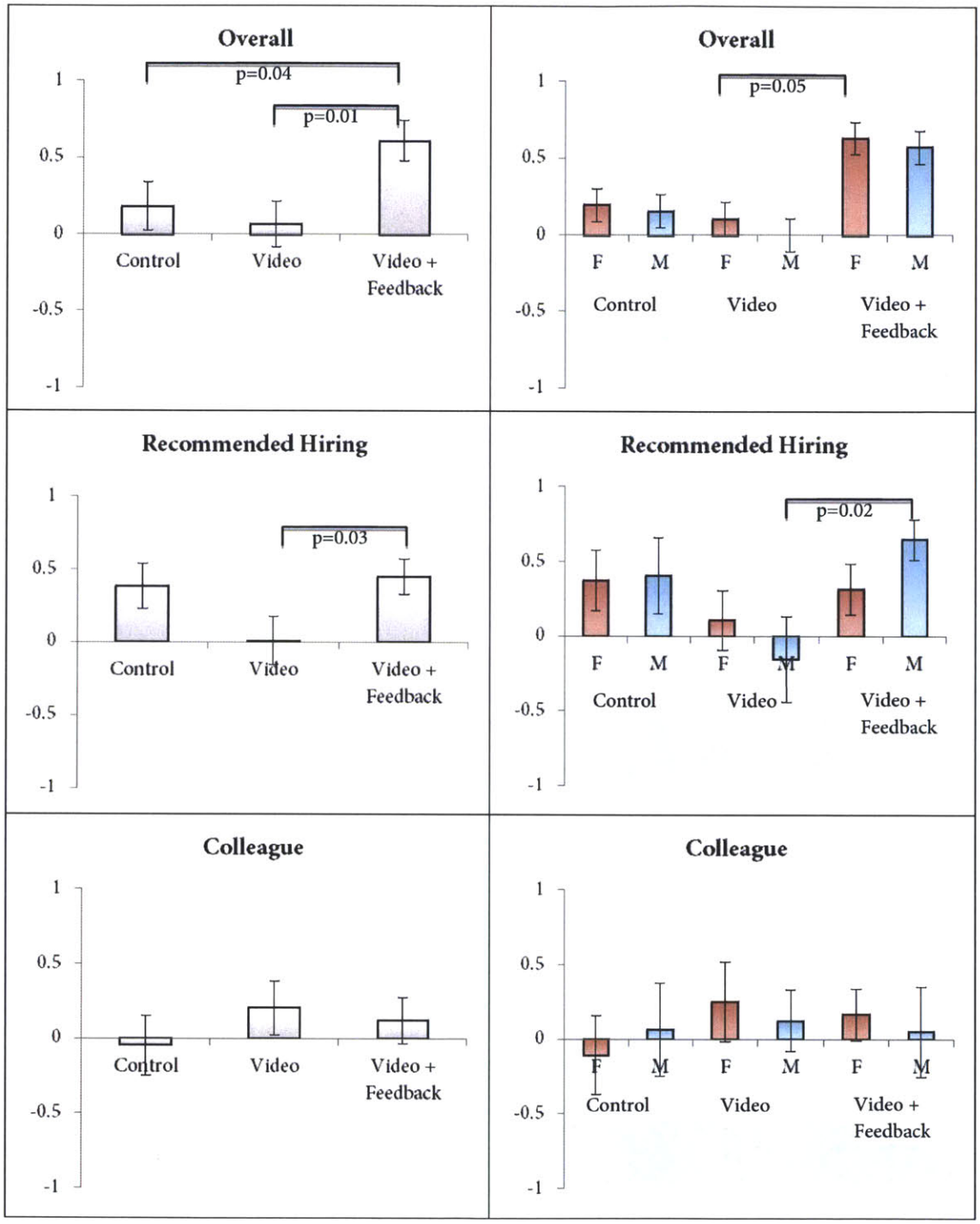
APPENDIX F-2: TWO-WAY ANOVA ANALYSIS BASED ON THE TURKER'S RATINGS (CONDITIONS WITH GENDER)

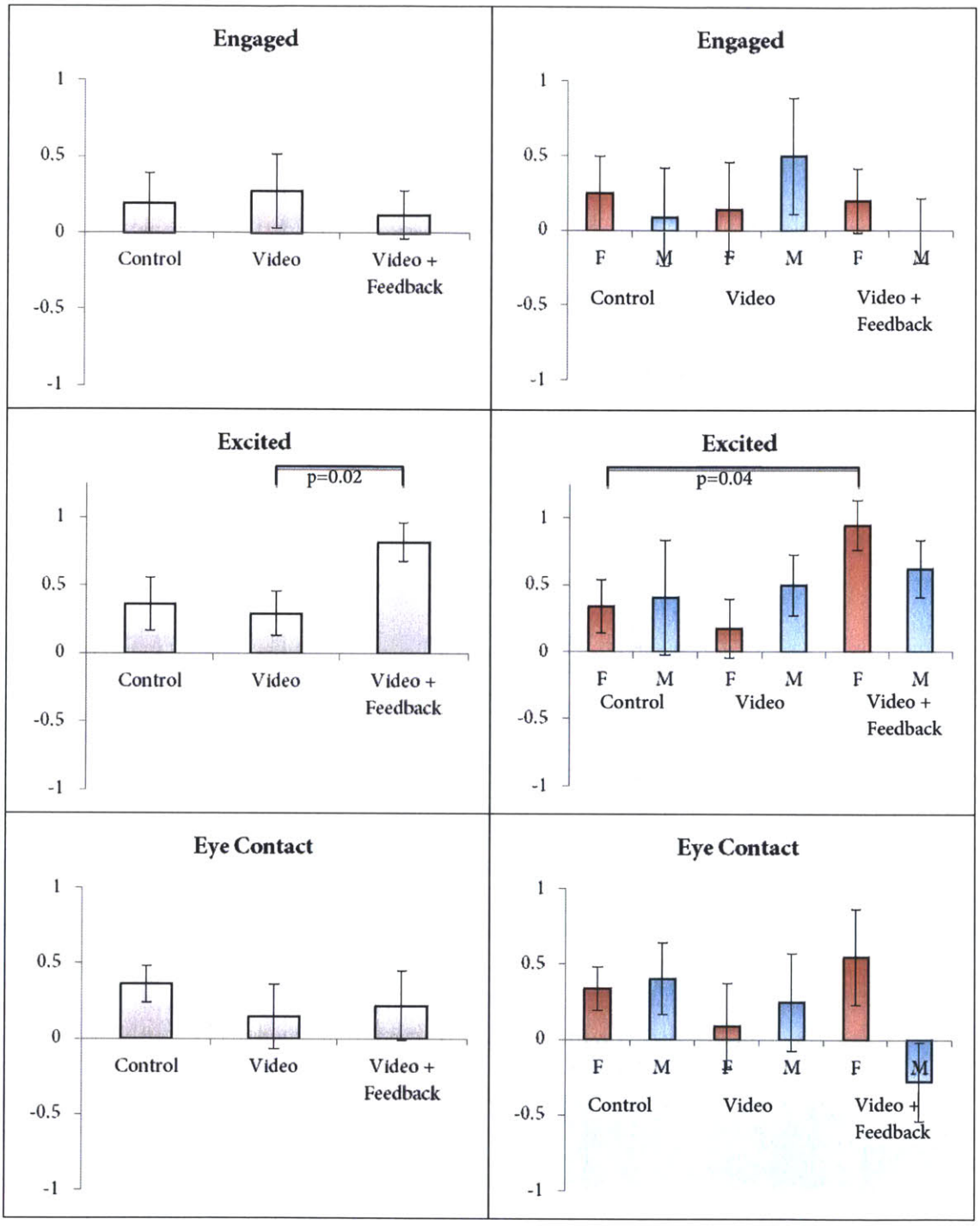
Turkers (conditions and gender)

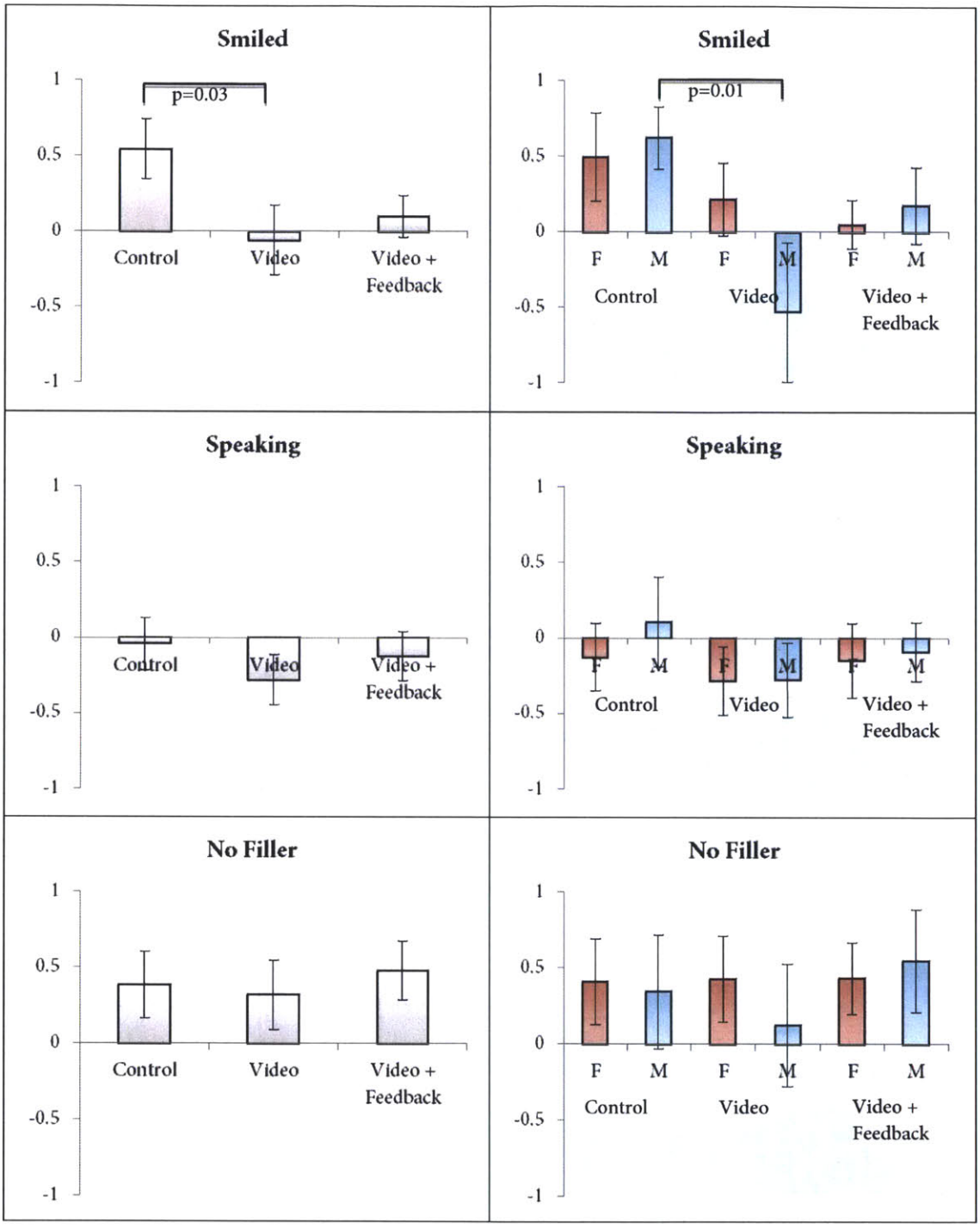
	Effect Test	Female Feedback vs. Video	Female Feedback vs. Control	Female Control vs. Video	Male Feedback vs. Video	Male Feedback vs. Control	Male Control vs. Video
Overall	F(2,63)=0.01, p=0.99	F(1,63)=3.95, p=0.05	F(1,63)=2.72, p=0.10	F(1,63)=0.11, p=0.74	F(1,63)=2.90, p=0.09	F(1,63)=1.54, p=0.22	F(1,63)=0.19, p=0.66
Recommended Hiring	F(2,63)=1.01, p=0.37	F(1,63)=0.65, p=0.42	F(1,63)=0.05, p=0.82	F(1,63)=1.03, p=0.31	**F(1,63)=5.93, p=0.02**	F(1,63)=0.54, p=0.46	F(1,63)=2.60, p=0.11
Colleague	F(2,63)=0.19, p=0.83	F(1,63)=0.07, p=0.80	F(1,63)=0.71, p=0.40	F(1,63)=1.17, p=0.28	F(1,63)=0.03, p=0.86	F(1,63)=0.00, p=0.98	F(1,63)=1.17, p=0.28
Engaged	F(2,63)=0.55, p=0.58	F(1,63)=0.03, p=0.87	F(1,63)=0.02, p=0.89	F(1,63)=0.09, p=0.77	F(1,63)=1.20, p=0.28	F(1,63)=0.04, p=0.84	F(1,63)=0.71 ,p=0.40
Excited	F(2,63)=0.92, p=0.40	**F(1,63)=6.74, p=0.01**	**F(1,63)=4.22, p=0.04**	F(1,63)=0.74, p=0.39	F(1,63)=0.06, p=0.82	F(1,63)=2.41, p=0.13	F(1,63)=1.65, p=0.20
Eye Contact	F(2.63)=1.92, p=0.16	F(1,63)=1.76, p=0.19	F(1,63)=0.37, p=0.55	F(1,63)=0.50, p=0.48	F(1,63)=1.40, p=0.24	F(1,63)=2.36, p=0.13	F(1,63)=0.11, p=0.74
Smiled	F(2,63)-1.62, p=0.21	F(1,63)=0.24, P=0.63	F(1.63)=1.80, p=0.19	F(1,63)=0.70, p=0.41	F(1,63)=2.71, p=0.10	F(1,63)=1.10, P=0.30	**F(1,63) = 6.55, p=0.01**
Speaking	F(2,63)=0.11, p=0.90	F(1,63)=0.20, p=0.66	F(1,63)=0.00, p=0.95	F(1,63)=0.25, p=0.62	F(1,63)=0.23, p=0.64	F(1,63)=0.26, p=0.62	F(1,63)=0.88, p=0.36
No Filler	F(2,63)=0.23, p=0.80	F(1,63)=0.00, p=0.99	F(1,63)=0.00, p=0.95	F(1,63)=0.00, p=0.96	F(1,63)=0.74, p=0.39	F(1,63)=0.18, p=0.68	F(1,63)=0.18, p=0.67
Friendly	F(2,63)=0.50, p=0.61	F(1,63)=0.51, p=0.48	F(1,63)=0.24, p=0.62	F(1,63)=0.05, p=0.83	F(1,63)=0.25, p=0.62	F(1,63)=0.53, p=0.47	F(1,63)=0.05, p=0.84
Paused	F(2,63)=0.39, p=0.68	F(1,63)=0.64, p=0.43	F(1,63)=0.00, p=0.98	F(1,63)=0.58, p=0.45	F(1,63)=0.10, p=0.75	F(1,63)=0.88, p=0.35	F(1,63)=0.35, p=0.56
Engaging	F(2,63)=1.58, p=0.21	F(1,63)=0.69, p=0.41	F(1,63)=0.04, p=0.85	F(1,63)=0.39, p=0.53	F(1,63)=0.71, p=0.40	**F(1,63) = 4.15, p=0.04**	F(1,63)=1.28, p=0.26
Structure	F(2,63)=0.10, p=0.90	F(1.63)=0.85, p=0.36	F(1,63)=0.76, p=0.39	F(1,63)=0.00, p=0.96	F(1,63)=1.09, p=0.30	F(1,63)=0.17, p=0.68	F(1,63)=0.36, p=0.55

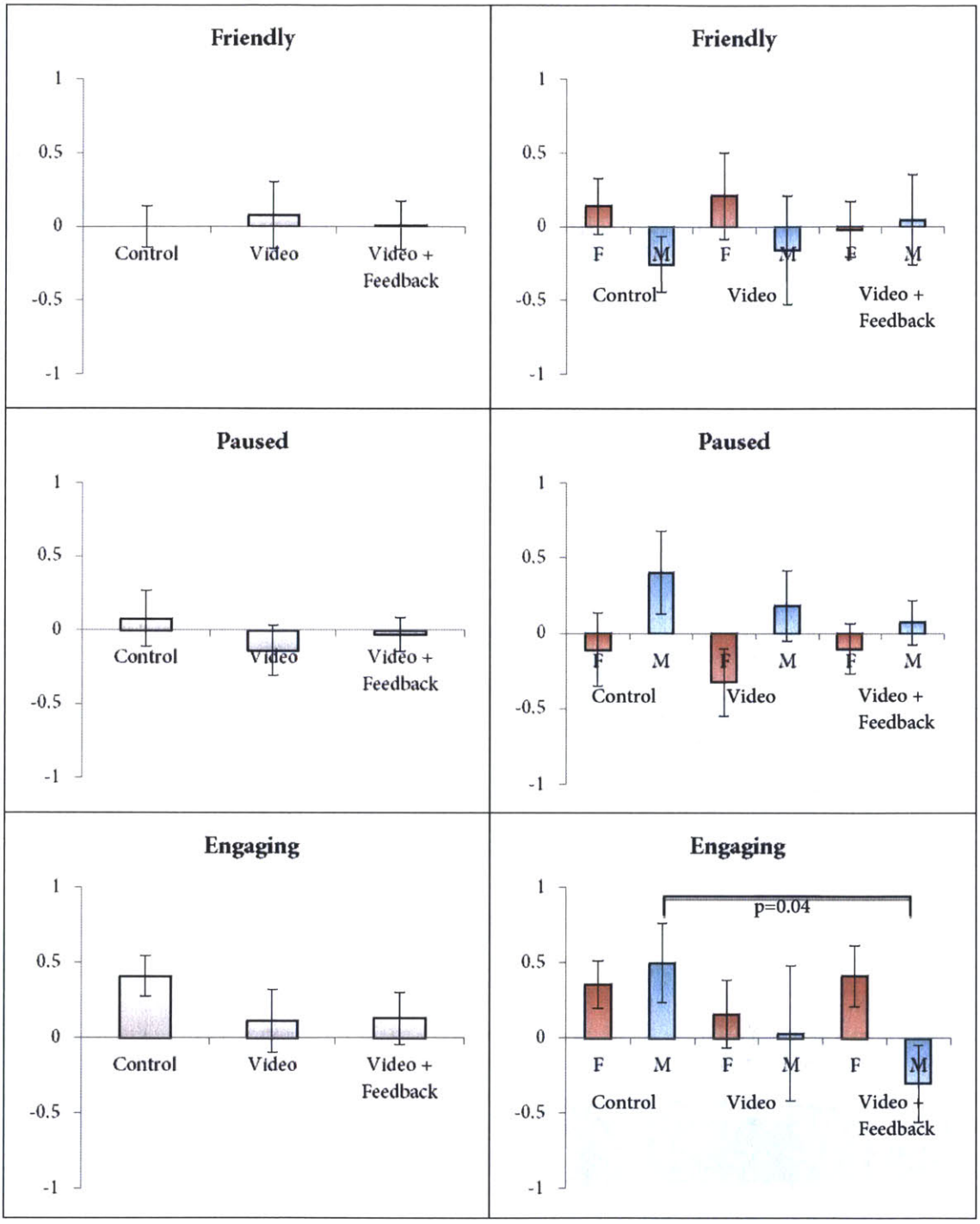
Calm	F(2.63)=0.20, p=0.82	F(1,63)=0.58, p=0.45	F(1,63)=0.19, p=0.66	F(1,63)=0.10, p=0.75	F(1,63)=0.01, p=0.93	F(1,63)=0.11, p=0.74	F(1,63)=0.06, p=0.81
Not Stressed	F(2,63)=0.33, p=0.72	F(1,63)=0.46, p=0.50	F(1,63)=0.03, p=0.87	F(1,63)=0.26, p=0.62	F(1,63)=1.27, p=0.26	F(1,63)=0.12, p=0.73	F(1,63)=1.95, p=0.17
Focused	F(2,63)=0.44, p=0.64	F(1,63)=0.00, p=0.97	F(1,63)=1.07, p=0.30	F(1,63)=0.96, p=0.33	F(1,63)=1.36, p=0.25	F(1,63)=0.04, p=0.84	F(1,63)=1.68, p=0.20
Authentic	F(2,63)=1.14, p=0.33	F(1,63)=0.98, p=0.33	F(1,63)=0.10, p=0.76	F(1,63)=1.63, p=0.21	F(1,63)=0.01, p=0.93	F(1,63)=1.06, p=0.31	F(1,63)=0.90, p=0.37
Not Awkward	F(2,63)=0.59, p=0.56	F(1,63)=1.65, p=0.18	F(1,63)=2.16, p=0.15	F(1,63)0.01, p=0.92	F(1,63)=1.85, p=0.18	F(1,63)=0.01, p=0.94	F(1,63)=1.48, p=0.23

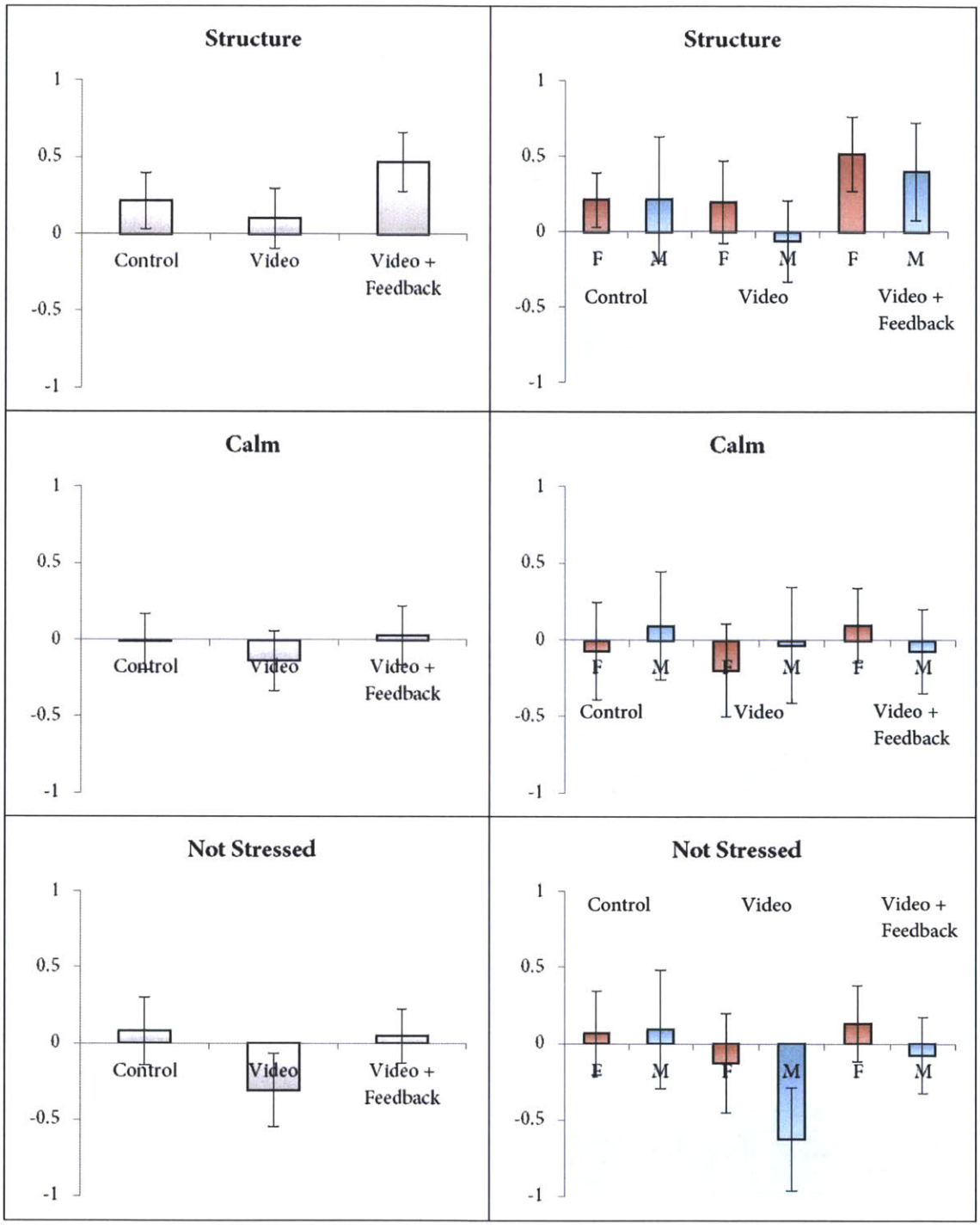
APPENDIX F-3: GRAPHICAL ANALYSIS BASED ON THE TURKER'S RATINGS

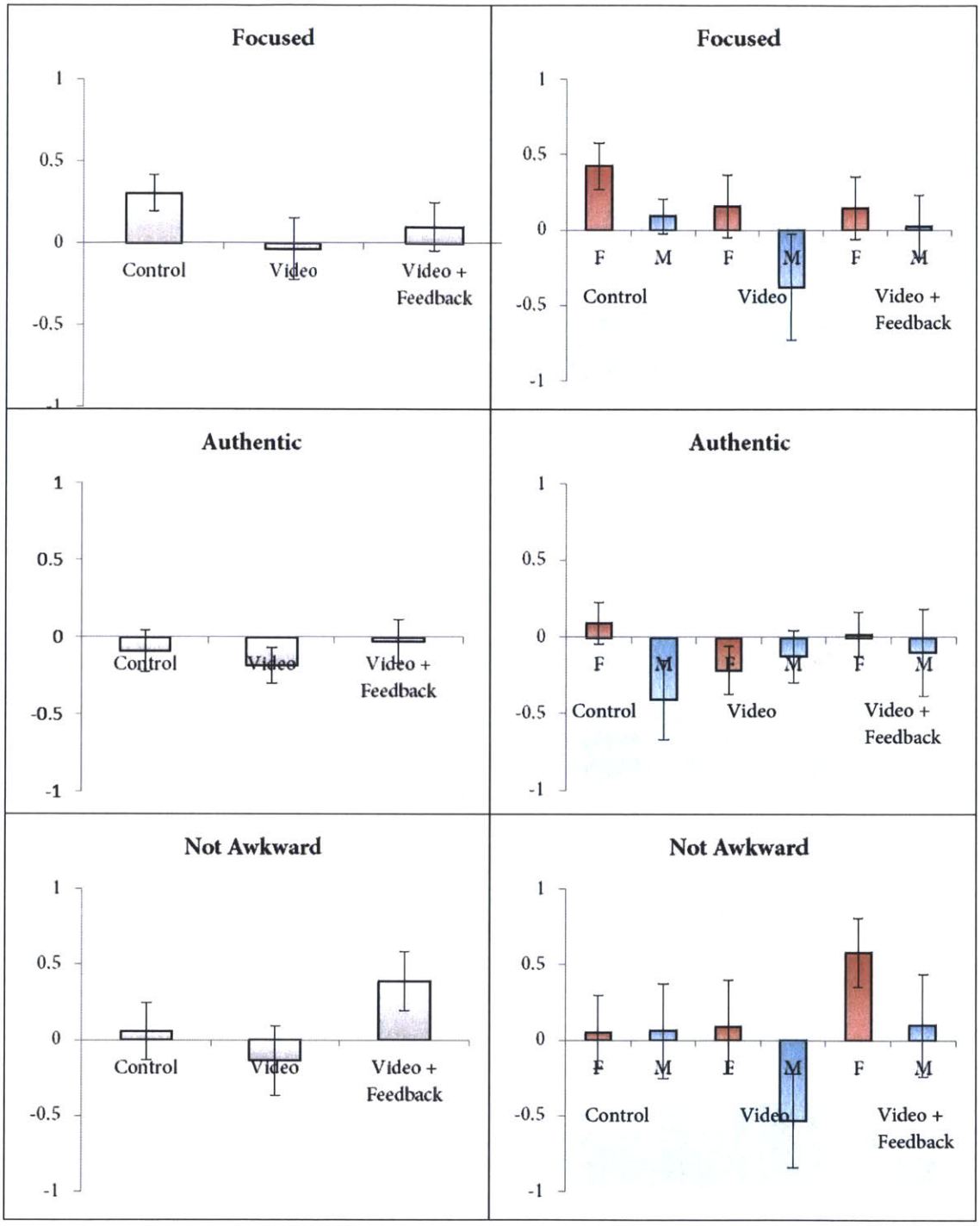



Engaged
1
0.5
0
-0.5
-1
Control
Video
Video + Feedback
Engaged
F
M
Control
Video
Video + Feedback
Excited
p=0.02
Excited
p=0.04
Eye Contact
Eye Contact


Smiled
p=0.03
1
0.5
0
-0.5
-1
Control
Video
Video + Feedback
Smiled
p=0.01
F
M
Control
Video
Video + Feedback
Speaking
Speaking
No Filler
No Filler


Friendly
1
0.5
0
-0.5
-1
Control
Video
Video + Feedback
Friendly
F
M
Control
Video
Video + Feedback
Paused
Paused
Engaging
Engaging
p=0.04
Engaging


Structure
1
0.5
0
-0.5
-1
Control
Video
Video + Feedback
Structure
F
M
F
M
F
M
Control
Video
Video + Feedback
Calm
Control
Video
Video + Feedback
Calm
Control
Video
Video + Feedback
Not Stressed
Control
Video
Video + Feedback
Not Stressed
Control
Video
Video + Feedback


Focused
1
0.5
0
-0.5
-1
Control
Video
Video + Feedback
Focused
F
M
Control
Video
Video + Feedback
Authentic
Authentic
Not Awkward
Not Awkward

APPENDIX G: CUSTOM INTERFACE DESIGNED FOR COUNSELORS AND TURKERS TO RATE THE VIDEOS

Links to videos and questionnaires: female

go to: male video list

Participant Number	Questionnaire Link
P13	Rate Video
P17	Rate Video
P15	Rate Video
P27	Rate Video
P77	Rate Video
P25	Rate Video
P26	Rate Video
P37	Rate Video
P15	Rate Video
P81	Rate Video
P33	Rate Video
P37	Rate Video
P57	Rate Video
P55	Rate Video
P72	Rate Video
P73	Rate Video
P21	Rate Video
P31	Rate Video
P17	Rate Video
P77	Rate Video
P22	Rate Video
P84	Rate Video
P25	Rate Video
P11	Rate Video
P57	Rate Video

Counselor Video Rating Form

* Required

Participant Number * P17

YouTube Video ID * _KrtF0YImCI

Gender * female

The overall rating for this person during this interview was "_____." *

	1	2	3	4	5	6	7	
very bad	○	○	○	○	○	○	○	very good

I would definitely recommend hiring this person. *

	1	2	3	4	5	6	7	
strongly disagree	○	○	○	○	○	○	○	strongly agree

I would love to work with this person as a colleague. *