


3-25-2015

Learning Data-Driven Models of Non-Verbal Behaviors for Building Rapport Using an Intelligent Virtual Agent

Reza Amini

Florida International University, ramin001@fiu.edu

Follow this and additional works at: <http://digitalcommons.fiu.edu/etd>

 Part of the [Artificial Intelligence and Robotics Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

Recommended Citation

Amini, Reza, "Learning Data-Driven Models of Non-Verbal Behaviors for Building Rapport Using an Intelligent Virtual Agent" (2015). *FIU Electronic Theses and Dissertations*. Paper 1765.
<http://digitalcommons.fiu.edu/etd/1765>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

LEARNING DATA-DRIVEN MODELS OF NON-VERBAL BEHAVIORS FOR
BUILDING RAPPORT USING AN INTELLIGENT VIRTUAL AGENT

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Reza Amini

2015

To: Dean Amir Mirmiran
College of Engineering and Computing

This dissertation, written by Reza Amini, and entitled Learning Data-Driven Models of Non-verbal Behaviors for Building Rapport Using an Intelligent Virtual Agent, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

S. S. Iyengar

Tao Li

Armando Barreto

Ubbo Visser

Jeffrey F. Cohn

Christine Lisetti, Major Professor

Date of Defense: March 25, 2015

The dissertation of Reza Amini is approved.

Dean Amir Mirmiran
College of Engineering and Computing

Dean Lakshmi N. Reddi
University Graduate School

Florida International University, 2015

© Copyright 2015 by Reza Amini

All rights reserved.

DEDICATION

To my family for their love and support
my father, my mother, my brothers, and my wife

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Christine Lisetti, for her great support and advice during all these years. Thank you for always being available and constructive, and for providing great advice.

I also would like to thank other members of my Ph.D Committee, Dr. S. S. Iyengar, Dr. Tao Li, Dr. Armando Barreto, Dr. Ubbo Visser, and Dr. Jeffrey F. Cohn, for reviewing my dissertation, for their support, and for their valuable feedback.

Special thanks are directed to the Affective Social Computing Laboratory (ASCL) members for the very nice and supportive working atmosphere.

Last but not least, I owe my dear friends and family, especially my daddy and mommy who were fully present despite the geographical distance.

ABSTRACT OF THE DISSERTATION
LEARNING DATA-DRIVEN MODELS OF NON-VERBAL BEHAVIORS FOR
BUILDING RAPPORT USING AN INTELLIGENT VIRTUAL AGENT

by

Reza Amini

Florida International University, 2015

Miami, Florida

Professor Christine Lisetti, Major Professor

There is a growing societal need to address the increasing prevalence of behavioral health issues, such as obesity, alcohol or drug use, and general lack of treatment adherence for a variety of health problems. Excessive alcohol use is the third leading preventable cause of death in the United States, and is responsible for a wide range of health and social problems. On the positive side though, these behavioral health issues can often be prevented with relatively simple lifestyle changes, such as learning how to reduce alcohol consumption. Medicine has therefore started to move toward finding ways of preventively promoting wellness, rather than solely treating already established illness.

Evidence-based patient-centered Brief Motivational Interviewing (BMI) interventions have been found particularly effective in helping people find intrinsic motivation to change problem behaviors after short counseling sessions, and to maintain healthy lifestyles over the long-term. Lack of locally available personnel well-trained in BMI, however, often limits access to successful interventions for people in need. To fill this accessibility gap, Computer-Based Interventions (CBIs) have started to emerge. Success of the CBIs, however, critically relies on insuring engagement and retention of CBI users so that they remain motivated to use these systems and come back to use them over the long term as necessary.

Because of their text-only interfaces, current CBIs can therefore only express limited empathy and rapport, which are the most important factors of health interventions. Fortunately, in the last decade, computer science research has progressed in the design of simulated human characters with anthropomorphic communicative abilities. Virtual characters interact using humans' innate communication modalities, such as facial expressions, body language, speech, and natural language understanding.

To facilitate successful communication and social interaction between artificial agents and human partners, it is essential that aspects of human social behavior, especially empathy and rapport, be considered when designing human-computer interfaces. Hence, the goal of the present dissertation is to provide a computational model of rapport to enhance an artificial agent's social behavior, and to provide an experimental tool for the psychological theories shaping the model. Parts of this thesis were already published in [LYL⁺12, AYL12, AL13, ALYR13, LAYR13, YALR13, ALY14].

TABLE OF CONTENTS

CHAPTER	PAGE
1. Introduction	1
1.1 Statement of the Problem	5
1.2 Specific Research Questions and Objectives	6
1.3 Outline of the Dissertation	8
2. Literature Review	10
2.1 Rapport	10
2.1.1 Modalities of Communicating Rapport	12
2.2 Empathy	13
2.2.1 Motor (Parallel) Empathy	15
2.2.2 Emotional (Affective) Empathy	16
2.2.3 Cognitive Empathy	17
2.2.4 Verbal Empathy	17
2.2.5 Compound Empathy	18
2.3 Computational Modeling of Rapport and Empathy	19
2.3.1 Modeling Rapport	22
2.3.2 Modeling Motor Empathy (Mimicry)	25
2.3.3 Modeling Emotional (Affective) Empathy	28
2.3.4 Modeling Cognitive Empathy	29
2.3.5 Modeling Verbal Empathy	34
2.3.6 Modeling Compound Empathy	37
2.4 Automatic Gesture/Expression Generation	44
2.4.1 Rule-Based Approaches	45
2.4.2 Machine Learning Approaches	59
2.5 Facial Expression Generation	63
2.5.1 Facial Action Coding System (FACS)	63
2.5.2 Facial Expression Datasets	65
2.5.3 Virtual Character Animation	68
2.5.4 Haptek Avatar System	72
2.6 Summary of Literature Review	73
3. Haptek Character Animation	76
3.1 HapFACS	76
3.1.1 HapFACS Functionalities	78
3.1.2 Registers, Switches and Action Units	80
3.1.3 Integration with Other Software	84
3.1.4 HapFACS Validation	86
3.2 HapGest: Haptek Gesture Synchronizer	116

4. On-Demand Virtual Counselor (ODVIC)	120
4.1 Introduction	121
4.2 Motivational Interviewing (MI) and Brief MI	125
4.3 Health Counselor System Architecture	127
4.3.1 System Overview	127
4.3.2 Avatar-Based Multi-Modal User Interface	129
4.3.3 Dialog Module	129
4.3.4 Psychometric Instruments	130
4.3.5 Utterance Planner	131
4.3.6 Score Evaluator	132
4.3.7 Empathy Model	132
4.4 ODVIC Evaluation	137
4.5 Hypotheses	137
4.5.1 Procedure	138
4.5.2 Questionnaire	140
4.6 Results and Discussion	144
4.6.1 User Acceptance Results	144
4.6.2 Agent Evaluation Results	153
4.7 Summary	157
5. Modeling Rapport Using Machine Learning	159
5.1 Overview	160
5.2 Data Collection	161
5.3 Annotation Schema	163
5.3.1 Visual Features	163
5.3.2 Textual Features	164
5.4 Data Annotation	166
5.4.1 Face, Head Movement, and Eye Gaze Recognizer	167
5.4.2 Part of Speech Tagger	171
5.4.3 Dialog Act Tagger	172
5.4.4 Phrase Boundary Tagger	174
5.4.5 Sentence Valence Tagger	176
5.4.6 New-Word Tagger	178
5.5 Data Pre-Processing	179
5.6 Data Alignment	180
5.7 Feature Selection	180
5.8 Model Induction	185
5.8.1 Probability of an Observed Sequence	190
5.8.2 Maximum Likelihood State Assignment: Viterbi	192
5.8.3 Parameter Learning for HMMs	193
5.9 Runtime Operation	194
5.9.1 Modeling Non-verbal Rapport Communication	197
5.10 Evaluation and Validation - Hypotheses Testing	198

5.10.1 Objective Evaluation of the Non-verbal Models	198
5.10.2 Subjective Evaluation of the Character	200
5.11 Summary	214
6. Conclusions	216
6.1 Summary	216
6.2 Future Directions	217
Bibliography	220
VITA	251

LIST OF TABLES

TABLE	PAGE
2.1 Synthesis of the state of the art.	21
3.1 Haptেক registers for a full-body character.	81
3.2 Mapping between FACS action units and Haptেক facial variables.	85
3.3 Individual AU recognition rates in Experiment 1.	91
3.4 Inter-rater correlation (Cronbach α) for individual AU recognitions.	91
3.5 Individual AU recognition results in AU combinations of Experiment 2.	92
3.6 AU combination recognition rates. AUs in parentheses show false-positive recognitions.	93
3.7 AUs involved in emotional expressions.	94
3.8 Subjects' demographic data in each group of Experiment 3.	95
3.9 Emotion recognition percentages of the 100%-intensity colored videos.	96
3.10 Emotion recognition ratings of the 50%-intensity colored videos.	97
3.11 Believability (sd = std. deviation) and Cronbach α for Experiment 3.	98
3.12 Subjects' demographic data in each group of Experiment 4.	99
3.13 Emotion recognition ratings of the 100%-intensity gray-scaled videos.	100
3.14 Emotion recognition ratings of the 50%-intensity gray-scaled videos.	101
3.15 Believability (sd = std. deviation) and Cronbach α for Experiment 4.	102
3.16 Subjects' demographic data in each group of Experiment 5.	102
3.17 Emotion recognition ratings of the 100%-intensity colored images.	103
3.18 Emotion recognition ratings of the 50%-intensity colored images.	104
3.19 Believability (sd = std. deviation) and Cronbach α for Experiment 5.	105
3.20 Subjects' demographic data in each group of Experiment 6.	106
3.21 Emotion recognition ratings of the 100% intensity gray-scaled images.	106
3.22 Emotion recognition ratings of the 50% intensity gray-scaled images.	107

3.23	Believability (sd = std. deviation) and Cronbach α for Experiment 6.	109
3.24	Subjects' demographics in Experiment 7.	110
3.25	Emotion recognition ratings of the videos in Experiment 7.	110
3.26	Believability (sd = std. deviation) and Cronbach α for Experiment 7.	112
3.27	Subjects' demographics in Experiment 8.	113
3.28	Comparison between recognition rates and believability of the static emotional expressions generated in FACSGen and HapFACS. FACSGen believability rates are scaled from 7 to 5-scale for comparison.	114
4.1	AUDIT score interpretation.	131
5.1	Frequency of the speaker (i.e., counselor) and the listener (i.e., client) non-verbal behaviors in the dataset.	162
5.2	Part of speech tags of the Stanford Natural Language Toolkit.	165
5.3	Recognition accuracies reported by SightCorp for the InsightSDK.	169
5.4	Feature comparisons used to recognize emotional facial expressions.	170
5.5	Feature comparisons used to recognize smile.	170
5.6	Max. frequency features for speaker models (numbers show frequencies).	182
5.7	Max. frequency features for listener models (numbers show frequencies).	183
5.8	Final set of features selected for speaker models.	186
5.9	Final set of features selected for listener models.	187
5.10	Objective evaluation results of the speaker models.	199
5.11	Objective evaluation results of the listener models.	200
5.12	Subjective evaluation mean value comparison and Chi-Square test results.	210
5.13	Comparing evaluations of decision tree and machine learning (ML) approaches.	212

LIST OF FIGURES

FIGURE	PAGE
2.1 Experiment setup used in [GWO07, WG09].	23
2.2 System architecture of the Rapport Agent 2.0 [HMG11].	25
2.3 Game experiment in [GSM ⁺ 11].	26
2.4 Story telling experiment setup used in [RR08].	27
2.5 Story telling experiment setup in [HSW ⁺ 06].	29
2.6 Job interview experiment setup used in [PI05].	30
2.7 Gaming experiment setup used in [PBA06, BBA07].	31
2.8 Gaming experiment setup used in [PLM ⁺ 11].	32
2.9 Internal emotion mimicry in [BW11].	33
2.10 A snapshot of the FitTrack [BP05].	35
2.11 Health application setup used in [BPJ09].	35
2.12 Virtual coach setup used in [MDI ⁺ 11].	37
2.13 Web credibility experiment setup used in [NM09].	38
2.14 Sequence diagram of the information flow in [BAW09].	40
2.15 Empathy model and test environment of used in [RM09].	41
2.16 Environment setup used in [MRP08].	42
2.17 The 3D character used in [MEM12].	44
2.18 The BEAT architecture [CVB01].	46
2.19 Example BEAT application snapshot [CVB01].	49
2.20 The system architecture and interface snapshot in [LMR06].	51
2.21 The system architecture and interface snapshot in [NPI07b].	53
2.22 Output expression in [FO08].	56
2.23 Geneva Emotion Wheel and appraisal dimensions used in [LM12].	57

2.24	Overall framework proposed by [LM12].	58
2.25	The stages of the learning process proposed by [LM09].	59
2.26	Data construction process in [LM09].	60
2.27	Sample AUs of the FACS.	64
3.1	HapFACS interface snapshot.	79
3.2	Haptek facial and head registers.	83
3.3	HapFACS work flow.	86
3.4	Models used for evaluation.	89
3.5	Emotion recognition percentages of the 100%-intensity colored videos.	96
3.6	Emotion recognition percentages of the 50%-intensity colored videos.	97
3.7	Emotion recognition percentages of the 100%-intensity gray-scaled videos.	100
3.8	Emotion recognition percentages of the 50%-intensity gray-scaled videos.	101
3.9	Emotion recognition percentages of the 100%-intensity colored images.	103
3.10	Emotion recognition percentages of the 50%-intensity colored images.	104
3.11	Emotion recognition percentages of the 100%-intensity gray-scaled images.	107
3.12	Emotion recognition percentages of the 50%-intensity gray-scaled images.	108
3.13	Validation of speaking characters showing emotional faces.	111
3.14	HapGest architecture.	117
4.1	ODVIC Amy in her office.	124
4.2	System Architecture.	128
4.3	Ethnicity and gender concordance.	129
4.4	A sample piece of the decision tree used in cognitive module.	136
4.5	Mean value comparison of experimental conditions for user acceptance features. Percentages show the empathic character's improvement over the text-only system.	153
4.6	Mean value comparison of experimental conditions for the character features. Percentages show the empathic character's improvement over the text-only system.	157

5.1	The overview of the modeling phases.	161
5.2	Snapshot of the implemented face recognizer.	167
5.3	Face recognizer design.	168
5.4	POS tagger design.	172
5.5	Dialog act tagger design.	174
5.6	Phrase boundary tagger design.	175
5.7	Sentence valence tagger design.	178
5.8	New-word tagger design.	179
5.9	10-fold cross validation.	184
5.10	Sample Hidden Markov Model.	189
5.11	Runtime process for the virtual character's speaker role.	195
5.12	Runtime process for the virtual character's listener role.	195
5.13	New architecture of the ODVIC.	201
5.14	Snapshot of the virtual counselor.	202
5.15	Class diagram of the ODVIC.	203

CHAPTER 1

Introduction

There is a growing societal need to address the increasing prevalence of *behavioral health issues*, such as obesity, alcohol or drug use, and general lack of treatment adherence for a variety of health problems. The statistics, worldwide and in the USA, are daunting. Excessive alcohol use is the third leading preventable cause of death in the United States [Nat11] (with 79,000 deaths annually), and is responsible for a wide range of health and social problems (e.g., risky sexual behavior, domestic violence, loss of job). Alcoholism is estimated to affect 10-20% of US males, and 5-10% females sometime in their lifetimes. Similar risks exist with other forms of substance abuse. In 2010, the *World Health Organization (WHO)* reported that obesity – worldwide – has more than doubled since 1980. In 2011, 1.5 billion adults in the world were overweight, of which 500 million were obese, and 43 million children under the age of five were overweight [WHO11]. In the USA alone, obesity affects 33.8% of adults, 17% (or 12.5 million) of children and teens, and it has tripled in one generation. These behavioral issues place people at risk of serious diseases; e.g., obesity can lead to diabetes, alcoholism to cirrhosis, physical inactivity to heart disease.

On the positive side though, these behavioral health issues (and associated possible diseases) can often be prevented with relatively simple lifestyle changes, such as losing weight with a diet and/or physical exercise, or learning how to reduce alcohol consumption. Medicine has therefore started to move toward finding ways of preventively promoting wellness rather than solely treating already established illness. In order to address this new focus on wellbeing, health promotion interventions aimed at helping people to change lifestyle have been designed and deployed successfully in the past few years.

Evidence-based patient-centered *Brief Motivational Interviewing (BMI) interventions* have been found particularly effective in helping people find intrinsic motivation to change problem behaviors (e.g., excessive drinking and overeating) after short counseling sessions, and to maintain healthy lifestyles over the long-term [ER01, DDR01]. A methodological review of clinical trials of 361 treatments showed that out of 87 treatment methods, the top two ranked treatment styles were: 1) Brief Interventions and 2) Motivational enhancement therapies [MW02]. It is reported that 5 minutes of advice and discussion about behavioral problems (e.g., alcohol or drug use) following a screening can be as effective as more extended counseling, and that a single session can be as effective as multiple sessions [BG92].

Lack of locally available personnel well-trained in BMI, however, often limits access to successful interventions for people in need. Yet, the current epidemic nature of these problems calls for drastic measures to rapidly increase access to effective behavior change interventions for diverse populations. To fill this accessibility gap, evidence has accumulated about the general efficacy of *Computer-Based Interventions (CBIs)* [Hes97, BTB⁺08, Ski94, Cun99, PSSJC08].

Success of the CBIs, however, critically relies on insuring engagement and retention of CBI users so that they remain motivated to use these systems and come back to use them over the long term as necessary (e.g., for booster sessions, follow-ups, and lifestyle maintenance sessions). Whereas current BMI interventions delivered by computers have been found effective, high drop-out rates due to their users' low level of engagement during the interaction limit their long-term adoption and potential impact [PSSJC08, Ver10].

One crucial aspect positively affecting the health outcomes of BMIs (and most counseling techniques for that matter), involves the ability of the therapist to establish rapport and to express empathy [MR02]. *Empathy* can involve cognitive at-

tributes or affective attributes, which can also be combined during full-blown empathy [GM85]. *Cognitive attributes* of empathy involve cognitive reasoning used to understand another person’s experience and to communicate that understanding [Hoj07] (or putting oneself in someone else’s shoes). *Emotional* or *affective attributes* of empathy, on the other hand, involve physiological arousal and spontaneous affective expressive responses to someone else’s display of emotions [Wis87] (e.g., people often unconsciously mimic someone else’s perceived expressions of distress or joy). Someone can have a reflex-like affective physiological reaction to someone else’s experience (without cognitively understanding it), or a cognitive understanding of that person’s situation (without physically expressing it), or both.

Because of their text-based only interfaces, current CBIs can therefore only express limited rapport and empathy (mostly reflected in the choice of textual wording of the intervention). Fortunately, in the last decade, at the same time as CBIs are being developed and studied in healthcare, computer science research has progressed in the design of simulated human characters and avatars with anthropomorphic communicative abilities [CSPC00]. Expressive virtual characters and avatars are emerging technologies in multi-modal interfaces [ML09, UPI08, Hel04] that have become increasingly interesting user interfaces for a wide range of applications, such as tutoring systems [MEM12], health behavior change systems [LYL⁺12, SBS11], training interfaces [HFG03], and health applications [BPJ09].

Virtual characters who specifically focus on dialog-based interactions are called *Embodied Conversational Agents (ECAs)*, also known as Intelligent Virtual Agents (IVA). ECAs are digital systems created with an anthropomorphic embodiment (be it graphical or robotic), and are capable of having a conversation (albeit still limited) with a human counterpart, using some artificial intelligence broadly referred to as an “agent”. With their anthropomorphic features and capabilities, they interact using

humans' innate communication modalities, such as facial expressions, body language, speech, and natural language understanding, and can also contribute to bridging the digital divide for low reading and low health literacy populations, as well as for technophobic individuals [NK11, BPJ09].

In this dissertation, I posit that using well-designed rapport-enabled and empathic virtual characters for the delivery of BMIs has the potential to increase users' engagement and users' motivation to *continue* to interact with them.

In Human-Computer Interaction (HCI), the presence of contingent non-verbal feedback is shown to have two different types of effects: *Subjective* and *Behavioral*. Subjective effects include: (1) greater feelings of self-efficacy [KGWW08a]; (2) less tension [WG10] and less embarrassment [KGWW08b]; (3) greater feelings of rapport [WG10]; (4) greater sense of mutual awareness [PKG09]; and (5) greater feelings of trustworthiness about the agent [KGWW08b]. Behavioral effects include: (1) more disclosure of information including longer interaction times and more words elicited [GOL06, GWGF07, PKG09, WG10]; (2) more fluent speech [GOL06, GWGF07, PKG09, WG10]; (3) more mutual gaze [WG10]; and (4) fewer negative facial expressions [WG09].

Rapport plays a major role in human-human and human-computer social interactions and motivational behaviors. As described in Chapter 2, research has shown that creating rapport and empathizing with the users can improve a virtual character's user acceptance and engagement. To facilitate successful communication and social interaction between artificial agents and human partners, it is essential that aspects of human social behavior, especially empathy and rapport, be considered when designing human-computer interfaces. Hence, the goal of the present thesis is to provide

a computational model of rapport to enhance an artificial agent’s social behavior, and to provide an experimental tool for the psychological theories shaping the model.

1.1 Statement of the Problem

In human face-to-face communication, non-verbal behaviors, such as gaze, facial expressions, gestures, and body postures, improve the effectiveness of the communication and help create a smooth relationship between the interlocutors. In HCI, also, a key issue to create ECAs with believable non-verbal behavior involves the creation of models that can accurately reflect human-human communicative clues. Moreover, due to the major role of the rapport in human-human and human-computer social interactions and motivational behaviors, rapport can improve the virtual character’s user acceptance and engagement. Most of the current approaches, so far, use rule-based systems, in which rules are created by ECA designers from their literature research in social sciences. Whereas some of these rule-based systems have been successful in creating a sense of rapport, they are limited to the designer’s expertise and how well the designer has designed these rules.

In this dissertation, I will address the limitations of rule-based approaches to modeling non-verbal rapport patterns. I modeled multimodal non-verbal rapport signals displayed by humans in the following steps: (1) annotate and extract both *verbal* information (from the lexical and syntactical structure of the surface text) and *non-verbal* information from a video corpora of counseling sessions; (2) model different non-verbal behaviors (e.g., facial expressions, head gestures, and hand gestures) of the counselor using machine learning techniques; (3) integrate the combination of these non-verbal behavior models to animate the rapport messages of a virtual character during the assessment portion of a virtual counseling intervention for behavior change; (4) build my system with the content of an existing computerized evidence-based

behavior change intervention, and (5) evaluate the impact of the virtual character's rapport in terms of user's acceptance and perceived sense of rapport.

Dissertation statement: *Enjoyable communication with virtual characters can be achieved when characters are animated using a model of non-verbal rapport-building communication.*

1.2 Specific Research Questions and Objectives

My main dissertation thesis can be addressed by answering the following research questions:

1. Can I extract information from the lexical and syntactical structure of the surface text to support the automatic generation of believable non-verbal behaviors?
2. Can I extract information from counseling video corpora to support the automatic generation of believable non-verbal behaviors?
3. Given a set of non-verbal behavior models (e.g., head nod model, head movement model, hand gesture model, and smile model), can I model non-verbal rapport for a virtual character using a combination of these non-verbal behavior models?
 - (a) In order to improve the facial expressiveness of a virtual character, can I map all the possible facial muscle movements of the human face in terms Action Units (AUs) of the Facial Action Coding System (FACS) [EF78, ELF83, EFH02] to the virtual character's face, as a standard method for facial expression generation at the facial muscle-level, and reversely for facial expression recognition?

- (b) In order to improve the expressiveness of a virtual character, can I map the head movements (e.g., turn left and turn up) and gestures (e.g., head node and head shake) of a virtual character to the AUs corresponding to these in the FACS?

To answer the research questions, the following **project objectives** are realized:

Objective 1: In this project, I explore different techniques for modeling non-verbal behaviors of a human. The overall goal is to **explore the possibilities of using machine learning techniques to move away from hand-crafted rule-based models employed in most of the current health-related dialogue systems (Discussed in Section 2), toward modeling human’s non-verbal behaviors based on the data derived from the video and text corpora of human-human communication.**

Objective 2: In order to improve the head movement and facial expressiveness of the virtual character, I generate a software, which maps all the possible neck and facial muscle movements of the human face in terms AUs of the FACS to the virtual character’s head and face. Then, I study the accuracy and expressiveness of this mapping through different experiments.

Objective 3: Given the individual non-verbal models of a human, I study the possibility of combining these individual models to generate an integrated non-verbal **rappor** communication model. Therefore, I apply the generated non-verbal rapport-building communication model to a virtual character (e.g., in a virtual health counseling context), and study the improvements in the user acceptance of the character (e.g., perceived rapport, believability, likability, enjoyability, and usefulness) and perceived character features (e.g., animacy and perceived intelligence).

Some of the **main challenges** of this project are:

- Annotating and extracting verbal information from the lexical and syntactical structure of the surface text, and non-verbal information from a video corpora of real human-human counseling sessions.
- Using machine learning techniques to model the non-verbal behaviors using the data gathered from the annotated videos and surface text, and evaluating the objective performance of the models.
- Applying the combination of the non-verbal models to the virtual character and evaluate the subjective performance of the combination.

1.3 Outline of the Dissertation

- *Chapter 2*: this chapter reviews the literature and discusses the relevant background of the (1) psychological aspects needed to understand the theoretical concepts underlying the computational modeling of rapport; (2) previous research on computational modeling of rapport and empathy; (3) previous works on automatic gesture generation; and (4) related research on facial expression generation.
- *Chapter 3*: this chapter describes my approaches for facial expression generation (called HapFACS) and gesture generation (called HapGest) on Haptek characters.
- *Chapter 4*: this chapter describes the On-Demand Virtual Counselor (ODVIC), which I implemented as the framework for applying the non-verbal behavior models. This framework is used for evaluating whether simple rapport-building and empathizing techniques, such as facial expression adaption, can improve the user experience with an interactive system.

- *Chapter 5*: this chapter covers my approach for modeling non-verbal behaviors and rapport using machine learning.
- *Chapter 6*: this chapter summarizes and concludes my contributions in this thesis. The chapter ends with future directions for research.

CHAPTER 2

Literature Review

In my survey of the literature, I first explain some of the concepts related to human communication of rapport and empathy, then I discuss some of the latest approaches taken to model these concepts computationally.

2.1 Rapport

In a conversation, the *feeling* of **flow** and **connection** is formally known as *rapport*. Also, rapport is mostly correlated with non-verbal behaviors during the face-to-face interactions. Research shows that non-verbal behaviors are more indicative of rapport than verbal signals in human-human interactions [Gra99].

According to Tickle-Degnen and Rosenthal [TDR90], the three essential components of rapport are *mutual attentiveness* (e.g., mutual gaze, mutual interest, and focus during interaction), *positivity* (e.g., head nods, smiles, friendliness, and warmth) and *coordination* (e.g., postural mirroring, synchronized movements, balance, and harmony). In this dissertation I use the Tickle-Degnen and Rosenthal theory of rapport.

Knowing the definition of rapport, we need to know how people perceive and express rapport. Grahe [Gra99] tested the hypothesis that rapport can be perceived through visual channels. Grahe stimulated five display conditions: transcript, audio, video, video+transcript, and video+audio. Results show that perceivers with access to non-verbal visual information were the most accurate perceivers of rapport and the transcript condition produced the least accurate judgments.

In a human-human interaction, listeners frequently nod and use para-verbals, such as “uh-huh” and “mm-hmm”, when someone is speaking. Such behaviors are called *back-channel continuers*, which are considered by a speaker as signals that the com-

munication is working and that she/he should continue speaking. *Nods*, *postural mirroring*, and *mirroring of head gestures* (e.g., gaze shifts) are a few examples of back-channel continuers [CB99]. Moreover, different dis-fluency signals, such as repetition, spurious words, pauses and filled pauses (e.g., ehm, um, un), show that the speaker is experiencing processing problems or high cognitive load [CW98], to which the listener responds with a “take your time” feedback [WT00], posture shift, gaze shift or frown [HMG10b].

Furthermore, Tickle-Degnen and Rosenthal [TDR90] believe that *positive emotions* are also a part of the fundamental non-verbal behavior structure of rapport. The most universal and powerful expressions of positive emotions are *head nods*, *positive facial expressions* (e.g., happiness, surprised), *smile* and *eye contact*. Eye contact is indicative of positive feelings, yet in personal and competitive conditions it may indicate aggressiveness [TDR90]. Similarly, smiling may be a positive expression of warmth or a negative expression of anxiety [EFA80]. Therefore, the use of non-verbal acts must be viewed as context-dependent. During non-helping interactions, positive relationships exist between participants’ evaluative impressions and their partners’ forward trunk lean, smiling, nodding, direct body orientation, uncrossed arms, directed gaze, and posture mirroring. During helping interactions (e.g., health interventions, tutoring), posture mirroring seems to have the most effect [TDR90].

An important aspect to be considered in developing rapport-enabled systems is to be able to measure the perceived rapport by the users. According to Tickle-Degnen and Rosenthal [TDR90], non-verbal behaviors are measured in two ways: **molecular** and **molar**. The molecular measures consist of *counts* or *durations* of specific behaviors, such as head nodding or eye contact. Molecular level would be appropriate for measuring the attention and positivity components. The molar measures are defined in terms of the psychological impression, gestalt image, or perceived function

they create. Molar level would be more appropriate for measuring the coordination component.

Subject's rapport can also be assessed through *subjective* measures, such as social presence, helpfulness, distraction and naturalness [BH02], or *behavioral* measures, such as the speak length (as a measure of engagement), fluency of speech (e.g., number of repeated words, broken words, and filled pauses per minute), intimacy of disclosure, facial expressions produced, and amount of mutual gaze [GKW10].

2.1.1 Modalities of Communicating Rapport

The value of rapport comes not only from understanding others' feelings, but also from what we express as rapport reaction. The ability to understand other people's emotions from external signals, such as facial expressions, voice, and bodily gestures, is a core ingredient for communicating rapport and empathy [EF74, ZWRE92, EF74, BFS87, Fes87]. In comparison with the verbal modalities, non-verbal modalities might be more trustworthy because it's easier to shut off the words than it is the face, eyes, or body [Fus02]. An other important characteristic of the non-verbal modalities is that they accompany verbal information without disrupting the natural flow of speech. Also, Noller [Nol85] argues that constructs containing affect are communicated more quickly via non-verbal behaviors.

Although some researchers [KCC96] may believe that the amount of information that conversational gestures convey is very small relative to the information conveyed by speech, gestures tell us about the concepts underlying our communicative intentions that we want to express verbally. So, they enhance the communicativeness of speech, not by conveying information, but by helping the speaker formulate her/his speech to convey more adequate information [KCC96].

Non-verbal means of communicating rapport are facial expression, motor mimicry, head gestures, head movements, hand gestures, lean direction, eye gaze, and vocal features. Facial expressions and body/hand gestures are the most important modalities in human behavioral judgments [AR92]. Different modalities can cause different impressions, for instance, posture sharing indicates group rapport, body positioning shows liking and understanding, and behavioral mimicry (mirroring) creates rapport and increases liking [LJC03].

Furthermore, each modality can indicate one feature of the emotion better than other modalities [Fus02], for example, *vocal* modalities are better in indicating the intensity; *facial* modalities better show valence (i.e., positive vs. negative) information; and *body gestures* are good for action readiness. Expressing rapport in one modality can ease its expression in another, in the same way that gesturing can facilitate verbal encoding of messages [Fus02].

2.2 Empathy

Rapport and empathy are so similar and inter-connected as being interpreted the same sometimes. However, whereas rapport is referred to the behaviors that convey the feeling of flow and connection in a conversation, empathy conveys more cognition and understanding.

Empathy is an ambiguous concept, which was, to the best of my knowledge, first discussed by Robert Vischer in 1873. He used the German word *Einfühlung* to describe an observer's feelings elicited by works of art [Hun67, Jac92]. Then later in 1909, psychologist Edward Bradner Titchener introduced the new English word "empathy" as the translation of *Einfühlung* [Hoj07]. Empathizing is the ability to detect what others feel and to experience that emotion ourselves. The human ability to recognize others' emotional states from external signals, such as facial expressions

and bodily gestures, is a core ingredient of an empathic communication [EF74, BFS87, Fes87, ZWRE92].

Empathy plays a major role in human social interaction, such as motivating, cooperative behavior, and moral acts (e.g., helping, caring, and justice) [Hof00]. Empathy is interpreted as a **cognitive** attribute, an **emotional (affective)** attribute, or a combination of both, where cognition is the mental activities involved in acquiring and processing information for better understanding. Cognitive processing involves reasoning and appraisal, whereas emotional processing involves arousal and spontaneous affective responses. When we talk about cognition here, it should not be confused with the human's cognition processes.

Psychologists [Dav83, Cli02, LMMR97] believe that empathy has four components: (1) the *Perspective Taking* sub-scale (i.e., tendency to adopt the views of others spontaneously); (2) the *Empathic Concern* sub-scale (i.e., tendency to experience the others' feelings and to feel sympathy and compassion for unfortunate people); (3) the *Fantasy* sub-scale (i.e., tendency to imagine oneself in a fictional situation); and (4) the *Personal Distress* sub-scale (i.e., tendency to experience others' distress).

At the same time, empathy is known as the most important tool of promoting positive outcomes in psychotherapy [BM91] and it is the cause of improvement in 25%-100% of patients [MR02, Rog59]. Communication without empathy does not deliver the desired results [Stu06, Stu08]. Heimgärtner et al. [HTW11] believe that since a successful communication depends crucially on the empathizing capability of the people involved [Den87], in addition to the human-human interaction, empathy is an essential prerequisite for successful inter-cultural communication between human and computer. It promotes successful inter-cultural usability and good user experience.

Although empathy is a complex phenomenon with no unique definition, there seems to exist a general agreement among psychologists that empathy can be cat-

egorized under five categories: motor empathy (or mimicry), emotional (affective) empathy, cognitive empathy, verbal empathy, and different combinations of these empathy types.

2.2.1 Motor (Parallel) Empathy

Motor or parallel empathy (sometimes referred to as mimicry or mirroring), is often considered as a kind of primitive empathy [BBLM86, BBLM87, CB99, WKP⁺03]. Mimicry is an innate part of the human-human interaction [CML05], which improves relationships [Van03] with unconscious mirroring of the others' non-verbal behaviors.

The mirrored motor behaviors are not necessarily emotional behaviors. Motor empathy occurs when someone mirrors (mimics) the observed motor behaviors of someone else, such as speech patterns (e.g., accent, rates, rhythm, tone), gestures, head movements, hand gestures, postures, facial expressions, emotions, mannerisms, and idiosyncratic movements [HCR94].

Bavelas et al. [BBLM86] observed that non-verbal mirroring occurs too rapidly for much prior cognitive processing. Consistently, Chartrand and Bargh [CB99] hypothesized that perception of another's non-verbal behavior primes a perception-behavior link that is unconscious and automatic.

Mimicry is highly related to the brain's mirror neurons, that are the brain cells that enable us to "mirror" others' behaviors. Mirror neurons are activated when someone performs a motor behavior and observes someone else mimicking the same motor behavior (i.e., observing an on-going mirroring behavior, activates the mirror neurons and increases the tendency to mimic that perceived behavior) [CID⁺03, Gal03, Hoj07, Sch11, ST11]. Chartrand and Bargh [CB99] believe that perception of another person's behavior can automatically increase the imitating probability of the perceived behavior. They described this phenomenon as the "chameleon effect".

Mimicking facial expressions can also actually result in adopting the emotions and moods of others [HCR94], that approves Zajonc’s research results, which shows that imitating an emotional facial expression can alter the emotional state of a person to the imitated emotion [ZMI89].

Non-verbal mirroring, when not exaggerated to the point of mocking, positively influences the interlocutors in different ways: (1) facilitates communication and may increase the mirrored person’s attention [LB76]; (2) creates liking, rapport, and affiliation [LB76, Laf79, Laf82, Wal95, CB99, LJC03] (e.g., in health interventions, counselors mimic their clients’ behaviors to create rapport and affiliation and receive more information from them consequently [MT83]); (3) increases the persuasiveness level of a speaker, and the mirrored people feel more confident to talk [Laf79, LB76, Van03]; (4) results in perception of a pleasant and natural conversation, positively influencing the emotional state of the mirrored person [VbHKK04, WMS⁺87]; and (5) plays a major role in empathy perception by the person being mirrored [SbJS03].

The Perception-Action Model (PAM) of empathy [Pre07] is developed based on the motor empathy. According to the PAM, empathy is defined as “a shared emotional experience occurring when one person (i.e., subject) comes to feel a similar emotion to another (i.e., object) as a result of perceiving the other’s state. This process results from the fact that when a person (subject) pays attention to the emotional states of another person (object), subject automatically feels and expresses the same emotions as of the object [Pre07].

2.2.2 Emotional (Affective) Empathy

Contrary to motor empathy, in which motor behaviors are mirrored (whether they are emotional or not), in emotional empathy, one responds specifically to the *emotional* states of others.

Wispé [Wis87] defines empathy as “an observer reacting emotionally because he perceives that another is experiencing, or about to experience, an emotion”. The emotional empathy are can be a reaction to an *emotional display* of the other, or a response to any other *emotional stimuli*, such as a verbal phrase.

It should be noted that the term *sympathy* is also used to refer to what we just described above as emotional empathy. Psychologists believe that empathy is associated more with cognition and understanding, whereas sympathy is associated more with emotions [Hoj07, GM86]. Between sympathy and empathy, there is “compassion”, which has both cognition and emotion equally.

2.2.3 Cognitive Empathy

Cognitive empathy (theory of mind) is the communication of the cognitive understanding of the other’s emotions. Hojat [Hoj07] defines empathy as “a cognitive (rather than an emotional) attribute that involves an understanding (rather than feeling) of experiences, concerns and perspectives of the person”, combined with the communication of this understanding. This empathy type, in which one represents the internal mental state of another, is the ability to represent the perceived thoughts, desires, beliefs, intentions, and knowledge of the others [Fri03, Les87, PW78].

2.2.4 Verbal Empathy

Verbal empathy is the verbal reactions to someone’s verbal statements [Bla05]. When hearing about one’s situation, one can simulate the other’s state internally (i.e., role taking) and react to it verbally, which does not necessarily need a cognition behind that, and can be a simple verbal comment or reflection.

The following statements represent some simple approaches to the verbal expression of empathy [MSB93, ER01, NM09]: **validation** (e.g., I understand this is a difficult problem); **self-disclosure** (e.g., I’ve been in this situation before, ...); **rephrasing** (e.g., Let’s see if I understood correctly, ...); **sympathy** (e.g., It is so sad to be in such a situation); **metaphorical** (e.g., I remember the day that I was ...); **using a small talk** to build trust and relationship (e.g., A beautiful weather works like an energy drink for me!); **being polite and friendly** with acknowledging opinions (e.g., Thanks for sharing this information with me); **offering means** to enable the users to correct the character’s judgments (e.g., You seem to be sad, right?); expression of **empathic understanding** (e.g., I understand your feelings in such a difficult situation); **avoiding judgments and comparison**; **giving the speaker enough time** to speak; **focusing on the speaker** without distraction [Gor85]; and **reflective listening**, in which the listener confirms his/her perceptions of the utterance with reflecting it back to the speaker (e.g., speaker’s speech content, feeling, meaning, or summary) [Rog59, KL85] (e.g., – Speaker: I have one drink a day. – Listener: So, you have seven drinks per week).

2.2.5 Compound Empathy

Compound empathy is a combination of two or more of the above empathy types. For example, Goldstein [GM85] and Feshbach [Fes87] define empathy as the combination of the affective responses and cognitive understanding, while Blair [Bla05] define it as an overlap of the cognitive, motor, and emotional empathy. Blair believes that, motor and emotional empathy are mostly automatic and may occur simultaneously, especially in terms of facial expressions. Goldstein [GM85] defines empathy as a social requirement involving perspective taking, understanding of non-verbal signals, being

sensitive to the other’s affective state changes, and communication of a feeling of caring.

2.3 Computational Modeling of Rapport and Empathy

Humans continuously perceive others’ situation, modify their own affective state, and express a response (whether empathic or not). Interactive virtual agents can do the same, however, the virtual agents need to decide how to empathize given what they perceive, i.e., they need a computational model of rapport and empathy.

In different research projects, input features to the rapport and empathy models might be selected based on different modulation theories. For example, De-Vignemont [dVS06] believes that empathy is modulated through different factors, such as (1) relation between the interlocutors (i.e., affective link, familiarity, similarity, and communicative intentions), (2) situation context (i.e., appraisal of the situation), (3) features of the empathizer (e.g., mood arousal, personality, gender, age, and emotional regulation capacities), and (4) features of observed emotion (i.e., valence, intensity, saliency, and primary vs. secondary emotion). Emotions might be categorized into primary emotions and secondary (intermediate) [Dam94]. Typical primary emotions refer to emotions which are supposed to be innate. These include joy, sadness, anger, fear, disgust and surprise. Secondary emotions arise from higher cognitive processes, based on an ability to evaluate expectations over outcomes (e.g., hope, relief). They can be represented by mixtures of the primary emotions.

Based on Ortony et al. [OCC88], often called OCC, the intensity of an empathic emotion is prone to modulation through factors, such as: *desirability-for-self* (i.e., degree to which the desirable/undesirable event for the other is desirable/undesirable for the empathizer), *desirability-for-other* (i.e., degree to which the event is presumed to be desirable/undesirable for the other person), *deservingness* (i.e., degree to which

the other person deserves/not deserves the event), and *liking* (i.e, degree to which the other person is liked/disliked).

Table 2.1 shows some of the most recent research in rapport and empathy modeling. Items in Table 2.1 are ordered in terms of the type of the models. In the following sub-sections we study each research in more details. For each modeling approach, we indicate different aspects of the modeling process in the table:

- Type of interface agent embodiment, including 2D or 3D characters, and robots.
- Input: agent’s perception and recognition from the user, including facial expressions, which may or may not be based on FACS [EF78, ELF83, EFH02], user self-report, voice, head movement, gaze direction, body gesture, user’s pleasure/arousal/dominance, user’s characteristics, game/system/environment parameters, physiological signals, menu-based inputs, user performance, and user’s text input.
- Output: agent’s capabilities to express itself, including facial expressions, voice, head movement, gaze direction, body gesture, breathing behavior, and text output.
- Decision making methods, emotion theories, and intervention approaches: including Wizard of Oz (WOZ) experiments, hand-crafted rules, different machine learning techniques, Belief-Desire-Intention (BDI); OCC, PAM, and Pleasure-Arousal-Dominance (PAD), role taking, and Motivational Interviewing (MI).
- Chronology of various projects, i.e., the hierarchy of building the projects on top of previous projects.

Table 2.1: Synthesis of the state of the art.

Reference	Context	Method	Agent	Input Modality	Output Modality	Model	Based-On
[GWO07, GWGF07]	Story telling	Hand-crafted rules	3D	Voice, Gaze, Body, Head	Gaze, Body, Head	Rapport	-
[WG09]	Story telling	Hand-crafted rules, SVM	3D	Voice, Gaze, Body, Head, Face	Body, Head, Gaze	Rapport	[GWO07, GWGF07]
[HMG11]	Interviewing	CRFs	3D	Smile, Voice, Gaze, Head	Smile, Voice, Gaze, Head	Rapport	[WG09]
[GSM ⁺ 11]	Gaming	Hand-crafted rules	Robot	Face, Voice	Face, Voice	Motor	-
[RR08]	Story telling	Hand-crafted rules, WOZ	Robot	Face, Mouth	Head, Mouth	Motor	-
[HSW ⁺ 06]	Story telling	Hand-crafted rules, Naïve Bayes	Robot	Voice	Face	Emotional	-
[PI05]	Job interview	Bayesian Networks	2D	Self report, Physio. signals	Text, Voice	Cognitive	-
[PBA06]	Gaming	Rule-based	3D	Self report, Physio. signals	Face, Voice	Cognitive	[PI05]
[PLM ⁺ 11]	Gaming	Role taking	Robot	Game parameters	Verbal	Cognitive	-
[BW11, Bou13]	Virtual guide	Belief, Desire, Intention (BDI)	3D	PAD, Face	Voice, Face, Eyes, Breath	Cognitive	[BAW09]
[BP05]	Health	Rule-based	2D	Env. parameters, Self report	Voice, Face, Body	Verbal	-
[BPJ09]	Health	Rule-based	2D	Self report	Voice, Text	Verbal	[BP05]
[SBS11]	Health	Motivational Interviewing (MI)	2D	Text	Voice, Text	Verbal	[BP05]
[MDI ⁺ 11]	Virtual coach	Rule-based, MI	2D	Menu	Text	Verbal	-
[NM09]	Web credibility	Rule-based, MI	3D	Performance, Self report	Voice, Body, Face	Verbal	-
[JL04]	Health, Training	Rule-based	2D	Performance	Body, Face, Voice	Verbal	-
[LYL ⁺ 12, LAYR13, AL13, ALY14]	Health	Decision tree	3D	Head, Face, Performance	Voice, Head, Face	Compound	-
[BBA07]	Gaming	BDI	3D	Self report, Physio. Signals	Face, Voice	Compound	[PBA06]
[BAW09]	Gaming	BDI, OCC	3D	Self report, Physio. signals	Face, Voice	Compound	[BBA07]
[RM09]	Tutoring	PAM	3D	Face, Voice, Body	Face, Voice, Body	Compound	-
[MRP08]	Gaming	Naive Bayes, Dec. tree, SVM	3D	Characteristics, Self report, Game params, Physio. Signals	Text, Voice	Compound	-
[OSP10]	Mail system	BDI	3D	System parameters	Face	Compound	-
[MEM12]	Tutoring	WOZ	3D	Face	Face, Voice	Compound	-

2.3.1 Modeling Rapport

As discussed in Section 2.1, empathy and rapport are very interconnected. For example, they have mimicry as a common core component. In a research by Gratch et al. [GWO07], rapport is explored as a feeling of connectedness raised from positive feedbacks between the interlocutors associated with their emotional states. They perform a story telling experiment, during which the virtual agent provides feedback with non-verbal signals, such as nodding, and postural mirroring. The agent (called Rapport Agent) produces the feedback without understanding the meaning of the monologs. They use a set of fixed hand-crafted rules to map speakers' head gestures (nods, shakes, rolls), posture shifts (lean left or right), gaze direction (straight, up, down, left, right), speech acoustic features (pitch, intensity, range) to the agent's behaviors. Silence of the speaker is mapped to the agent's gaze up/straight. Agent nods in the cases of speaker's raised loudness, back-channel, and asking question. Speaker posture shifts, gaze aways, and nods (or shakes) are mimicked by the agent. Therefore, as a part of their rapport modeling, they also model motor empathy.

In an experiment setup shown in Figure 2.1, they designed a human-human and a human-computer interaction. In the human-human interaction experiment, participants were grouped into pairs and assigned to the roles of speaker and listener randomly. The speaker watches a video clip and then retells the story to the listener. In the human-computer interaction experiment, participant sits in front of a computer monitor seeing a virtual agent representing the human listener (the human listener also sits in front of a TV and listens to the speaker). Participants are randomly divided into three groups, each of which interacts with one of the three virtual agent configurations: (1) "good virtual listener", which gazes at the speaker and shows attentive listening with head nods, smiling, posture mimicry and posture shifts; (2) "not responsive listener", which gazes at the speaker and blinks randomly, but does

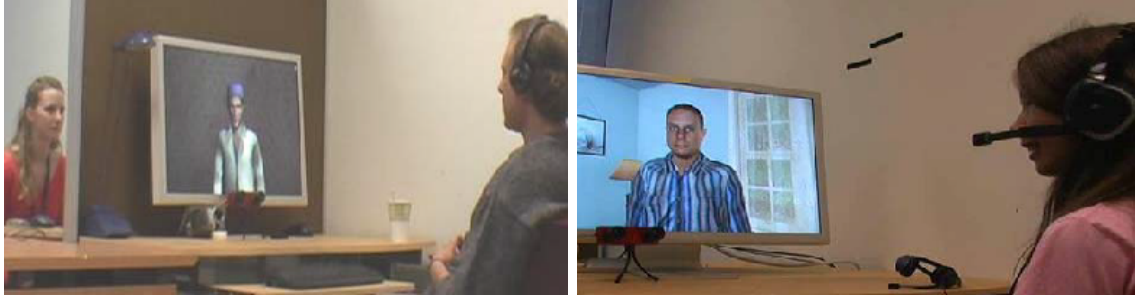


Figure 2.1: Experiment setup used in [GWO07, WG09].

not provide attentive listening; and (3) “ignoring listener”, which does not maintain gaze with the speaker and does not provide attentive listening (See Figure 2.1).

Results show that, the total time to tell the story and the count of the words used to retell the story in the first condition is longer than a face-to-face condition. Therefore, a simple virtual character with positive listening feedback (rapport and motor empathy) can be **more engaging** than a face-to-face human listener. Authors [GWO07] suggest that, such rapport-enabled and empathic agents can improve the computer-mediated systems used in the learning and health interventions to have socially desirable outcomes. They believe that rapport and motor empathy can lead to effective communications, better learning outcomes, and improved acceptance of medical advice in Human-Computer Interaction (HCI).

Among the visual channels, facial expressions are the most important in the human judgment of behavioral cues [AR92]. Human observers seem to be mostly accurate in their judgments when looking at the face. This fact indicates that people rely on the facial expressions to recognize someone’s behavioral changes. Thus, human affect analysis would better to include facial expressions as a modality [CMK⁺06].

Wang and Gratch [WG09] extended the previous research [GM04, GWO07] by adding facial expression recognition with the focus on the positivity component of rapport [TDR90]. During a similar experiment performed by Gratch et al. [GWO07], the virtual agent recognizes and analyzes the participants facial expressions (using

CERT toolkit [BLL⁺04] and a SVM classifier), speech (back-channels, dis-fluencies, questions, and loudness), head motions (head nods, shakes), gaze shifts, and body posture shifts in real-time and provides non-verbal feedbacks to the speaker. The feedback includes back-channel continuers (nods), postural mirroring, and mimicry of certain head gestures (e.g., gaze shifts and head nods). Obviously, similar to their previous research, motor empathy is included in their rapport model. They debriefed the participants via a questionnaire about the content of the video and the story retelling.

Results show that (1) negative facial expressions, such as disgust, are significant predictors of the lack of rapport and positive facial expressions are good indicators of rapport; (2) presence of listener’s nods enhances rapport; and (3) rapport-enabled and empathic virtual human listeners can be **more engaging** than human listeners.

Later, Huang et al. [HMG11] enhanced the Rapport Agent and developed the Virtual Rapport 2.0, in which the simple behavior rules are replaced by three probabilistic Conditional Random Field (CRF) models of the *backchannel* prediction, *end-of-turn* (turn-taking opportunity), and *affective feedback* (smile), based on the data driven from a video corpora. The CRF models predict when to give feedback and how to give such feedback. The input features used in modeling are silence, head nod, eye gaze and smile, where the non-verbal output features (i.e., generated feedback) are *smile* and *head nod*. Figure 2.2 shows the system architecture of the Rapport Agent 2.0.

The Rapport Agent 2.0 is tested in an interview setting, in which the agent interviews the human interviewees while creating rapport with them. Results show that, the **mutual attention, coordination, positive emotion communication** (i.e., affective response), **rapport, naturalness, and backchannel prediction** of the Rapport Agent are improved in Rapport Agent 2.0. However, the agent is still

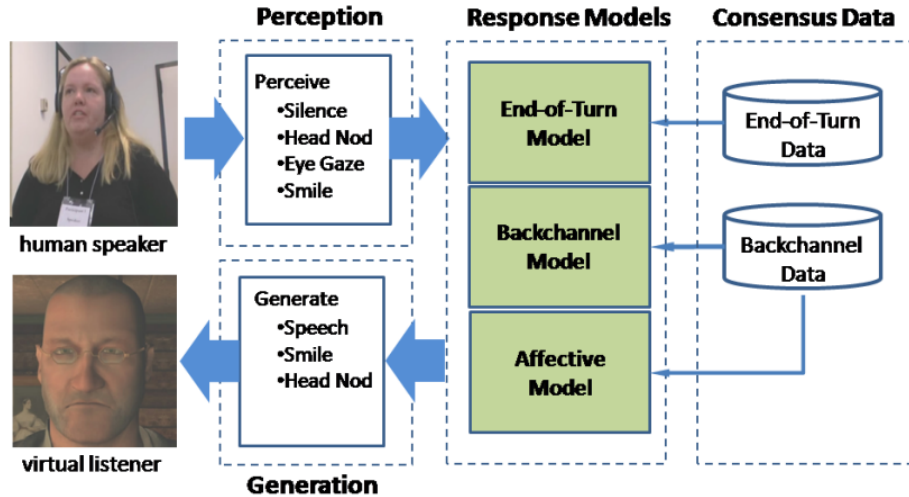


Figure 2.2: System architecture of the Rapport Agent 2.0 [HMG11].

limited in different aspects: (1) only head nod, smile, and turn-taking are modeled, (2) the dialog content is not used at all, and the character expresses non-verbal feedbacks without having any ideas about the content of the dialog, (3) few interactive input features are taken into account, and (4) non-verbal behaviors are only modeled for the listener role, but character needs to be able to express gestures while speaking as well. These limitations are addressed, in this dissertation, by extending the model learning process to other behaviors in both listener and speaker roles and by using new input features extracted from the dialog (e.g., part of speech, dialog acts).

2.3.2 Modeling Motor Empathy (Mimicry)

While rapport is mostly modeled using different non-verbal behaviors, different empathy types are also modeled non-verbally, such as the motor empathy (mimicry). In the Human-Robot Interaction (HRI) field, Gonsier et al. [GSM⁺11] found that not only a robot can empathize with the users, but also that a robot’s behavior influences the extent of the human’s empathy toward the robot. Their robot generated similarity (of personal attitudes) with the user by mirroring the user’s facial expressions, so

that their shared emotional state triggers the mirror neuron system of the user and thus evokes user’s empathy for the robot. They proposed a system, which recognizes the user’s facial expressions through camera, calculates the corresponding emotional facial expression according to the FACS, and reacts with the *same* emotional facial expressions on the robotic head called EDDIE (shown in Figure 2.3). EDDIE can display 13 Emotional FACS (EmFACS) action units (AU) [FE83], and it can recognize 7 AUs (i.e., outer brow raiser, brow lowerer, upper lid raiser, lid tightener, lip corner depressor, yaw drop, and eyes closed).

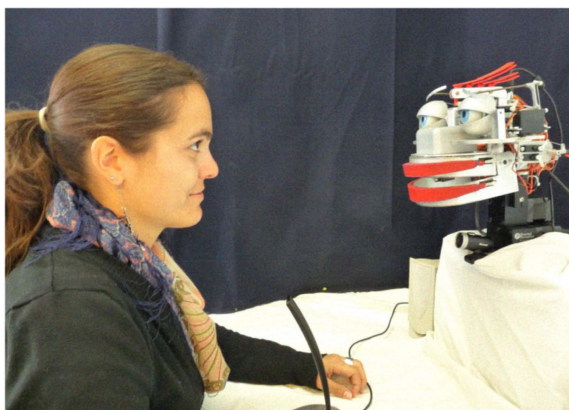


Figure 2.3: Game experiment in [GSM⁺11].

In a game scenario experiment (called Akinator), shown in Figure 2.3, EDDIE reacts to the human’s facial expressions either by ignoring them or by mirroring them. After the game, the Bartneck’s “five key concepts in HRI” questionnaire [BKC08] is used to evaluate the robot’s anthropomorphism, animacy, likability, perceived intelligence, and perceived safety. Furthermore, the user acceptance of the system was evaluated according to the Heerink’s measure [HKEW09] using five constructs: *turst*, *perceived sociability*, *social presence*, *perceived enjoyment*, and *intention to use*. Users’ responses (55 subjects) to the questionnaire showed that **users’ empathy is induced toward the mirroring robot**, and that the **mirroring improves the**

HRI regarding the Bartneck’s five key concepts by 48%, as well as the **user’s acceptance** of the robot.

In another HRI research, Riek and Robinson [RR08] used a robot with a chimpanzee head (named Virgil), shown in Figure 2.4, to model motor empathy by mimicking the subjects’ facial expression during a story telling session. The robot has degrees of freedom (DOFs) for eye movements, eyebrows, lower jaw, upper lip, and head, from which only the mouth open/close and head nods are used. Virgil uses its mouth open/close and head nods as output modalities. The robot’s facial expression recognition and expression are controlled manually in a Wizard-Of-Oz (WOZ) setup. The robot’s expressions are evaluated in two experimental conditions: (1) random expressions, and (2) mimicry, in which it mimics the open/close mouth expressions and performs head nods. Results show that, subjects in the facial-mimicking group rate their interaction with Virgil as more **satisfactory** than those in the random expression group. However, subjects reported that it is difficult to feel strongly engaged with the robot due to the fact that it did not speak nor acknowledged their statements, which shows the importance of having verbal modality in the output.

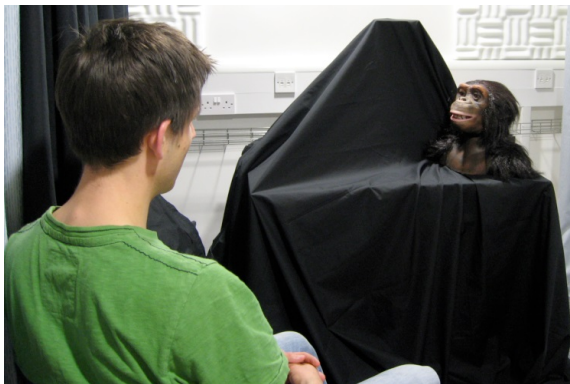


Figure 2.4: Story telling experiment setup used in [RR08].

Therefore, as shown in the mimicry related research, a simple one-to-one mapping of the non-verbal behaviors of the user to the same non-verbal behaviors of the character improves the empathizing ability of the character, the user acceptance, and the

user satisfaction. I used this research results in implementing a rule-based empathy model (discussed in Chapter 4 and used head movement mimicry for enhancing the user's experience with the character.

2.3.3 Modeling Emotional (Affective) Empathy

In addition to just mimicking the users' non-verbal behaviors, characters can recognize the emotional non-verbal behaviors, such as emotional facial expressions, and respond adequately to them (i.e., affective empathy). Affective virtual agents can help motivating users, supporting them through stressful tasks, and increasing their abilities to recognize and regulate emotions [GM04].

In a story telling scenario, Hegel et al. [HSW⁺06] uses an anthropomorphic robot (called BARTHOD Jr.), which recognizes the user's emotional states from speech and then mirrors this state with a corresponding emotional facial expression. The speech recognizer takes speech features, such as pitch, energy, MFCCs, frequency, duration, and pauses, as input and classifies them into 6 emotional states namely happiness, fear, surprise, anger, sadness, and thinking using a Naïve Bayes model. The robot can move its jaw, mouth, eyes, eyebrows, and eye lids. They model the emotional empathy by mirroring the emotional states of the user and show that comparing to a robot with no emotional reactions, an emotional robot is perceived as **being able to react more adequately** to emotional aspects of a situation and to recognize the emotions better. Figure 2.5 shows the experiment setup used in this research. One limitation of this research is that emotion detection accuracy from vocal features is not the same for all emotions. So, to improve the accuracy of emotion detection other modalities can be added.



Figure 2.5: Story telling experiment setup in [HSW⁺06].

2.3.4 Modeling Cognitive Empathy

While motor empathy enables the character to mimic the non-verbal behaviors of the user, and emotional empathy involves recognizing the affective states of the user in providing the best non-verbal feedback to the user, cognitive empathy enables the communication of the character’s cognitive understanding of the user’s emotions. This understanding can be derived using different input features such as user self reports, environment parameters, physiological signals, and facial expressions. For example, Prendinger and Ishizuka [PI05] designed an animated embodied agent, enabled with cognitive empathy, that helps the users to set a virtual job interview. They use a 2D character created by the Microsoft Agent, which is capable of interacting verbally. This application gathers physiological data (skin conductance, heart rate, and electromyography) of the user in real-time, as well as user self-reports about her/his valence and arousal, and interprets the data as emotions in 2D emotion model [Lan95]. A decision making system using Bayesian networks classifies these inputs to 3 user affective states (relaxed, joyful, and frustrated). Then, the character responds verbally to the user knowing his/her affective state. They concluded that an empathic agent can **reduce user’s level of stress or frustration**, and **undo**

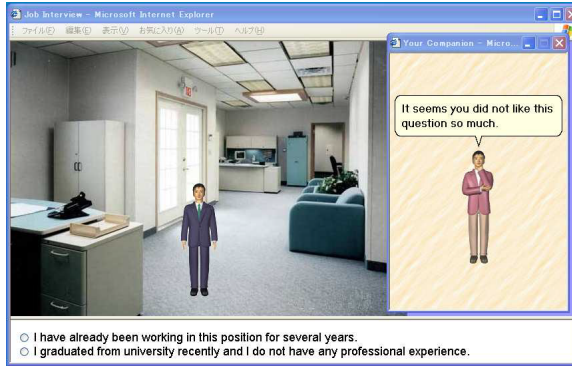


Figure 2.6: Job interview experiment setup used in [PI05].

negative emotions using empathic feedback. Figure 2.6 shows a snapshot of their experiment.

Prendinger and Becker-Asano [PBA06] extended the empathy model in the previous research [PI05] and applied it to a card game. They used a 3D lifelike character (called MAX) implemented in Poser, which provides empathic feedback in multiple modalities, such as facial expressions, affective verbal expressions, body gestures, eye blinking, head nodes and breathing behavior. In this research, they use the 2D emotion model for recognition of the user emotions from ElectroMyoGraphy (EMG), skin conductance (SC), game situational context/parameters, and self-report of valence and arousal. The PAD (3D) emotion model is used for the emotional feedback expressions. The user's affective states are categorized into 3 categories of joyful, fearful, and sad. Then, based on the user's affective state, four types of agents are compared: (1) non-emotional (i.e., agent does not display any emotional behavior); (2) self-centered (i.e., agent only appraises its own game play, e.g., by displaying joy when it is able to move cards); (3) negative empathy (i.e., agent is self-centered, plus, it appraises the user's play and responds to the user in a negative way, e.g., happy about the user's distress); and (4) positive empathy (i.e., agent is self-centered, plus, it appraises the user's play and responds to the user in a positive way, e.g., sorry for the user's distress). The third and fourth conditions model the cognitive empathy.

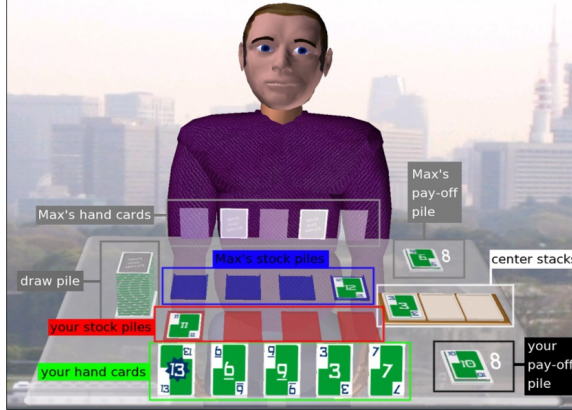


Figure 2.7: Gaming experiment setup used in [PBA06, BBA07].

Results show that, agent behavior should be adequate with respect to the context. For example, within a competitive game scenario, showing positive emotions is more arousing or stressful than displaying negative emotions. Figure 2.7 shows their experiment setup.

Empathy is a social requirement involving **perspective taking** (role-taking), the understanding of non-verbal cues, sensitivity to the other’s affective state and communication of caring [GM85]. Pereira et al. [PLM⁺11] believe that, in human-robot interaction, the more robots can socially interact with humans, the more people accept it. Therefore, they implemented a scenario, in which a social robot (iCat [vBY05]) watches, reacts empathetically, and comments a chess match played by two human players. The robot puts itself into the user’s situation (i.e., role-taking) to determine her affective state and decide about the empathic response. Figure 2.8 shows the experiment setup used in this research.

The robot’s affective state depends on the state of the game in the perspective of the robot’s companion. Two sets of utterances are considered for each affective state of the iCat: (1) **empathic**, which is used when the iCat is commenting its companion’s moves. Empathic utterances often contain references to the possible companion’s



Figure 2.8: Gaming experiment setup used in [PLM⁺11].

emotions, and try to encourage the companion (e.g., “you’re doing great, carry on!”); and (2) **neutral**: used when the iCat is commenting its opponent’s moves.

They measure the participant’s friendship toward the iCat using the Mendelson’s friendship questionnaires [MA99] involving six functions: (1) stimulating companionship (i.e., doing enjoyable/exciting things together); (2) help (i.e., providing guidance); (3) intimacy (i.e., being sensitive to the other’s needs/states and being open to honest expressions of thoughts, feelings and personal information); (4) reliable alliance (i.e., remaining available and loyal); (5) self-validation (i.e., reassuring, encouraging, and otherwise helping the other maintain a positive self-image); (6) emotional security (i.e., providing comfort and confidence in novel or threatening situations). Results show that, with the exception of the help dimension, all other dimensions were rated higher in the empathic condition than the neutral condition.

According to Polajnar et al. [PDP11], in the same way that the emotional intelligence and the empathy improve the effectiveness of human teamwork, the empathy has a significant role in developing robust artificial agents for tasks requiring practical reasoning. Having the same belief, Boukricha and Wachsmuth [BW11, Bou13] enabled a virtual human (called EMMA) to empathize with another virtual human (called MAX) in a museum guide context. They follow Davis’s [Dav94] empathy

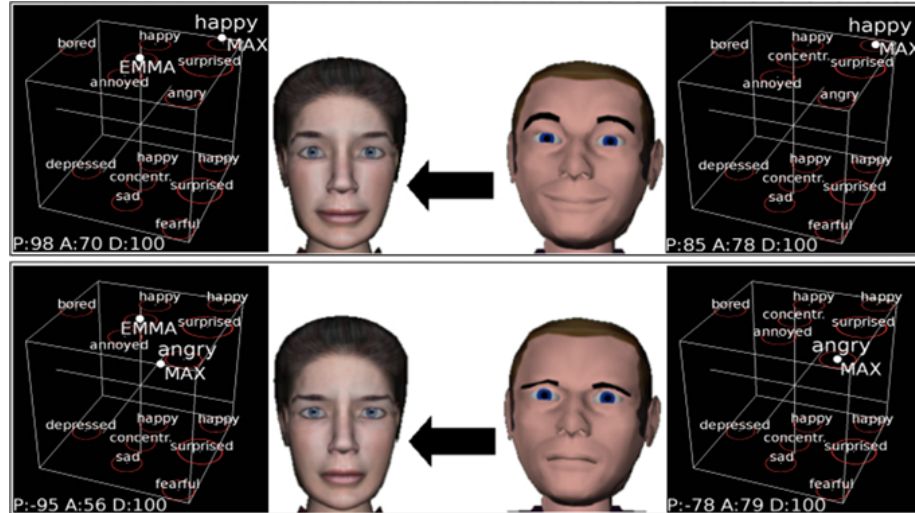


Figure 2.9: Internal emotion mimicry in [BW11].

model. Davis believes that between assessment and outcome, there is another step called *intra-personal*, in which cognitive and affective responses are produced in the observer but not showed to the other. Boukricha and Wachsmuth model the cognitive empathy in three steps: (1) *empathy mechanism*; (2) *empathy modulation*; and (3) *expression of empathy*. The *empathy mechanism* module accepts the emotional facial expressions as its input, then internally simulates and imitates them to come up with an internal emotional feedback that represents the empathic emotion in PAD space. This output is passed to the *empathy modulation* to be modulated through factors, such as agent’s mood and relationship to the other (e.g., familiarity and liking). In the *expression* step, the modulated empathic emotion activates multiple expression modalities, such as facial expression, verbal expression, eye blinking, and breathing behavior. Figure 2.9, shows the mimicry of emotions in this research.

As shown in the cognitive empathy related research, involving more understanding, in the empathy communication in comparison to the affective and motor empathy, reduces the stress, frustration level, and negative emotions; also it increases the companionship, intimacy, alliance, self-validation, and emotional security.

2.3.5 Modeling Verbal Empathy

While the above discussed empathy models used mostly the non-verbal behaviors to communicate empathy, in this section, I discuss a few of the related research which focus on communicating empathy through the verbal channel. The earliest computer application, which was trying to communicate with its users in a natural way using the Reflective Listening (discussed in Section 2.2.4) was developed by Joseph Weizenbaum at MIT in 1966 [Jos66] and called Eliza. The communication with Eliza is completely textual. The earliest Eliza was representing a psychotherapist who used the Reflective Listening techniques for engaging the clients (i.e., psychiatric patients). Eliza conveyed the sense of being intelligent and sometimes emotionally supportive [Tur95] to the users by parsing the user's sentences, performing word matching algorithms, and reflecting back to the user. Although Eliza was a great achievement in 1960's, it was highly limited in the dialog abilities and modalities (text).

More recently, again at MIT, the MIT FitTrack [BP05] was developed, which used an ECA to investigate the ability of an avatar-based system to establish and maintain a long-term working alliance with the users in a behavior-change context. It used a 2D lifelike character (called Laura) created with LiteBody. They model verbal empathy and rapport by applying different models of personal relationship, such as social dialogs, empathic dialogs, humor, continuity behaviors, politeness, and non-verbal behaviors (high/low immediacy). Figure 2.10 shows a snapshot of the FitTrack. A client-server design is used to enable users interact with FitTrack on their home computers on a daily basis during a one-month intervention, with each interaction taking ten minutes. They use self-report techniques to measure the users' performance including Working Alliance Inventory (WAI) questionnaire, which measures the therapist-patient trust. Results show that compared to an equivalent

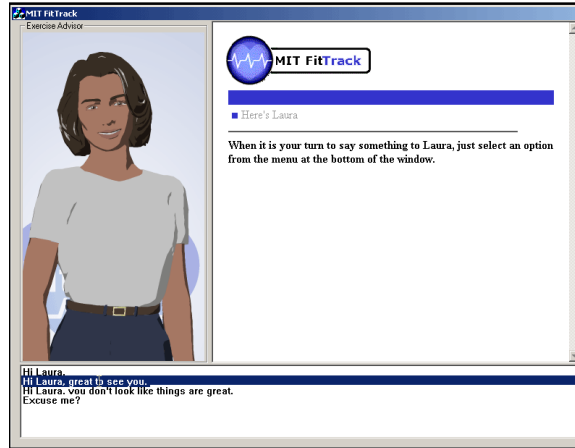


Figure 2.10: A snapshot of the FitTrack [BP05].

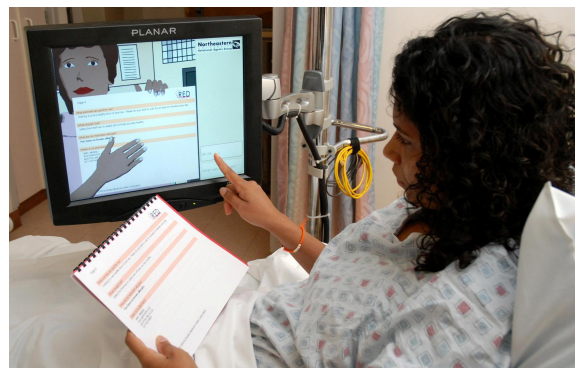


Figure 2.11: Health application setup used in [BPJ09].

agent without any deliberate social-emotional or relationship-building skills, their agent was **more respected, liked, and trusted**, even after four weeks of interaction.

The FitTrack system is used to develop the Virtual Hospital Discharge Nurse [BPJ09] to explain written hospital discharge instructions to the patients with low health literacy. It is made available to the patients on their hospital beds to help to review the material before discharging from the hospital. It also notifies the unresolved issues for the human nurse to be clarified for the patient. Results indicate that hospital patients with low health literacy found the system easy to use, reported **satisfaction**, and said they preferred receiving the discharge information from the agent over a human nurse. Figure 2.11 shows the experiment setup in this research.

Recently, Motivational Interviewing (MI) [MR02] has been identified as particularly useful for health dialog systems [Lis08, LW08a, LYL⁺12] because it intends to increase the likelihood of making change in unhealthy lifestyles through support and motivation. MI is a face-to-face patient-centered counseling style, which respects the patient's pace during the interaction toward the behavior change. It involves a brief assessment followed by an empathic feedback about the assessment.

Schulman et al. [SBS11] designed a conversational agent (designed with LiteBody) as a virtual counselor for health behavior change. They enabled the character to provide empathic verbal statements and techniques drawn from the MI to enhance client motivation and confidence to change. To model and implement these techniques, they use the Dtask dialog manager based on a domain-specific taxonomy of dialog acts. They perform a preliminary experiment, in which the virtual counselor gives counseling to two groups of people in exercising and diet contexts. Then, users are asked to rate the counselor's MI spirit, empathy, and their satisfaction. Results show that, only following the MI spirit satisfied the users and cause them to rate the empathy and MI spirit positively.

Magerko et al. [MDI⁺11] presents a 2D virtual coach agent, called Dr. Vicky, and training environment (called the Virtual BNI Trainer) for learning how to correctly talk with medical patients who have substance abuse issues. They designed a menu-based dialog interaction and engage the users in the conversations according to the Brief Negotiated Interview (BNI) techniques. This system follows four main steps during the interactions: (1) raise the subject: including introduction, and asking if the patient would mind talking about the substance use; (2) provide empathic verbal feedback; (3) enhance motivation: including asking the patient to rate his/her readiness to change in scale of 1-10 and respond accordingly verbally; and (4) negotiate and advice. A snapshot of the Dr. Vicky is shown in Figure 2.12.



Figure 2.12: Virtual coach setup used in [MDI+11].

Nguyen and Masthoff [NM09] used a 3D human-like Haptik avatar in a web credibility judgment test. In that application the agent empathizes using verbal phrases. Figure 2.13 shows a snapshot of this system. Nguyen and Masthoff used different validated [BS07, BP05, KMR02] approaches to the **verbal** expression of empathic understanding in interaction based on MI, sociology, and communication theories to increase the believability and build trust in the users. Some examples of these techniques are provided in Section 2.2.4. Results show that users have a positive attitude to the empathizing avatar. Also, they concluded that systems represented by a human-like representations are expected to behave more like humans and to be empathic.

As shown above, although the verbal channel solely seems very limited in delivering empathic signals, research shows that enabling a character with the verbal empathizing abilities improves the users' respect, likability, trust, satisfaction, and attitude toward the character.

2.3.6 Modeling Compound Empathy

Given the positive effects of each individual empathy type on users of empathic virtual characters, some research projects model different combinations of the empathy types



Figure 2.13: Web credibility experiment setup used in [NM09].

for a virtual character and study their effects on the users. Below, I discuss the design of some of these studies and their effect on the users.

Boukricha and Becker-Asano [BBA07] extended the empathy model proposed by Prendinger and Becker-Asano [PBA06] with an intelligent empathy model comprised of two components: a Belief-Desire-Intention (BDI) based cognitive component (which models the cognitive empathy) and an affective component (which models the emotional empathy). The cognitive component understands how emotions occur in human users, and the affective component simulates the emotion dynamics of the agent the same as the human user. In this research, a 3D character (called MAX) generates a hypothesis about the emotional state of the user by appraising user's situation in a card game environment, and displays his affective state by appropriate emotional facial expressions namely happy, sad, board, depressed, concentrated, surprised, angry, annoyed, and fearful (modeled in PAD emotion model). They conducted the same 4-condition experiments as in [PBA06] and confirmed the results

of the previous research. Also, they showed that the valence of the human player's emotion is congruent with the valence of the expressed emotion by the agent. Figure 2.7 shows the game environment of their experiment.

Becker-Asano and Wachsmuth [BAW09] used the same system as in their two previous research [PBA06, BBA07]. In this research a BDI reasoner (1) reasons about the current game state of the agent and the human player, and (2) recognizes the possible emotional state of the human player based on OCC emotion structure. They extended the previous research by adding secondary emotions to the agent's responses. This system is called WASABI (Affect Simulation for Agents with Believable Interactivity). In an experiment with the same card game scenario, the agent responds in 2 different ways to the human's appraised emotions: (1) the agent expresses primary emotions (i.e., happy, bored, surprised, concentrated, depressed, angry, annoyed, sad, fearful) by facial expressions based on PAD; the agent acknowledges the human's actions verbally and appraises them negatively; the agent appraises his own progress positively; and the agent feels dominant when it is his turn, and when correcting the human's mistakes; (2) the agent expresses the secondary emotions (namely hope, fears-confirmed, relief) verbally, in addition to the primary emotions. Figure 2.14 shows the information flow in their system.

Since secondary emotions need more cognition abilities than primary emotions, they are called "adult" emotions. Therefore, Becker-Asano and Wachsmuth hypothesize that players who played with an agent who expresses both primary and secondary emotions should perceive the agent older than the one that expresses only primary emotions. This hypothesis is supported by their results.

Rodrigues and Mascarenhas [RM09] model the cognitive and motor empathy between the synthetic characters. They use the empathy modulation factors proposed by De-Vignemont and Singer [dVS06] discussed in Section 2.3 (specifically similarity,

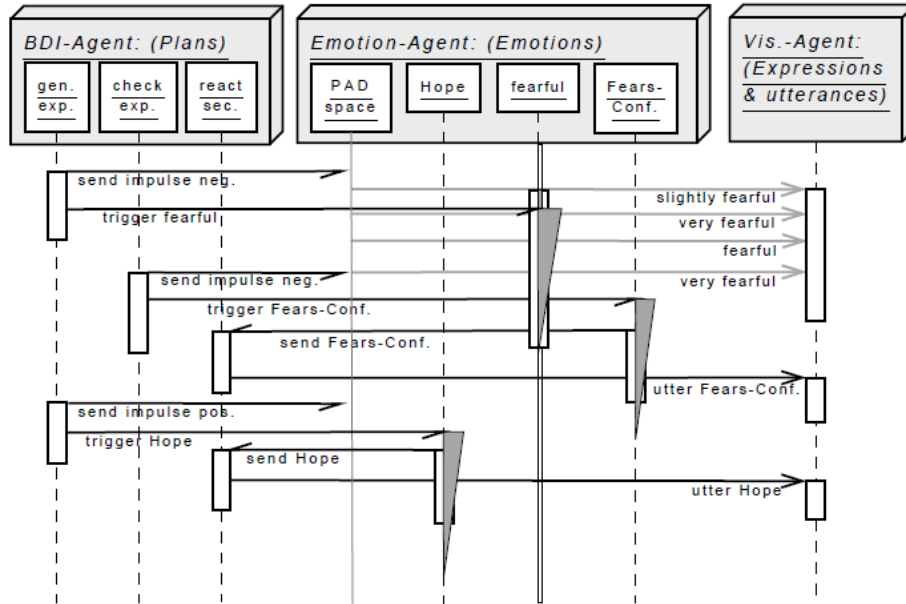


Figure 2.14: Sequence diagram of the information flow in [BAW09].

affective link, personality, and mood) and Perception-Action Model (PAM) [Pre07] of empathy (see Section 2.2.1). Rodrigues and Mascarenhas model empathy in two steps: (1) *Empathic appraisal*, which takes place when an agent perceives a new event that raises an emotional cue in another agent. The emotional cue can be facial expression, body posture or voice tone. The empathic agent recognizes the emotional cue by an emotion recognizer and candidates some emotions as the other agent’s possible emotion. Also, by appraising the other agent’s situation via self-projection (i.e., role taking), it elicits another list of candidate emotions. The two lists are compared and the strongest emotion is selected as the perceived emotion of the other agent; (2) *Empathic response*, in which the intensity of the emotion is determined using the fore-mentioned modulation factors. Then, based on the perceived emotion and its intensity, an empathic action is selected to be expressed.

In an experiment, two versions of an educational system (called FearNot!) is tested, one with empathic characters, another with non-empathic characters. The interaction between the characters is video recorded and shown to the subjects. The

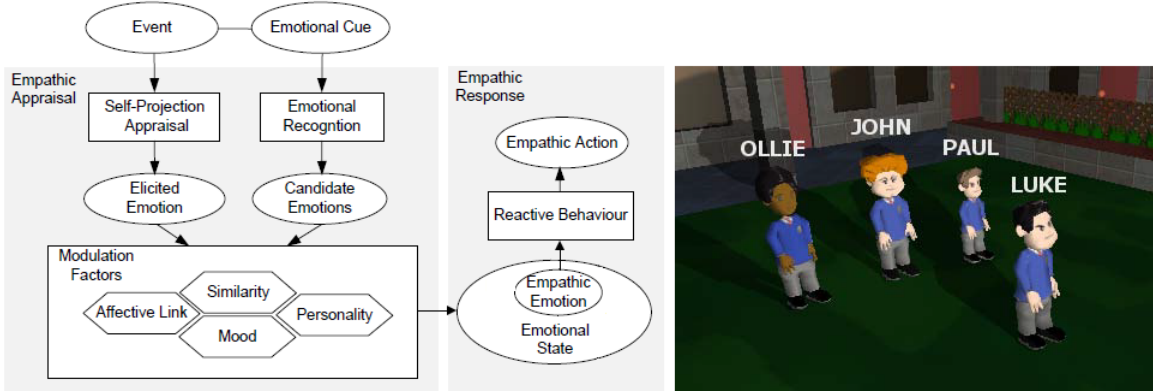


Figure 2.15: Empathy model and test environment of used in [RM09].

subjects are asked to rate the characters’ likability, relation between the characters in terms of like/dislike, perceived emotions felt by the characters, and their friendship. Results show that, users were capable to perceive the empathic responses elicited by the model. The empathy model and the experiment environment of this research is shown in Figure 2.15.

McQuiggan et al. [MRP08] proposed a framework (called CARE) for modeling a combination of the affective and cognitive empathy. CARE has two modes: (1) empathy model induction, which acquires the training data and learns the empathy models from the training users in a game environment; and (2) runtime operation, in which the induced model is used to select between the affective and the cognitive empathy.

In the training session, they used the game situation data, the affective states (namely anger, anxiety, boredom, confusion, delight, excitement, fear, flow, frustration, and sadness), the physiological responses (heart rate and galvanic skin response (GSR)), the characteristics (age, gender, user’s empathic nature measured by Interpersonal Reactivity Index [Dav83], and the user’s goal orientation measured by Elliot and McGregor’s goal inventory [EM01]) as the input vector. In the model induction mode, they modeled the empathy in three states: (1) antecedent, in which the user’s

affective and situational data is captured; (2) assessment, in which the data is processed and an empathic outcome is selected; and (3) empathic outcome, in which one of the cognitive or affective empathy is expressed. When the affective empathy is selected, the user's affective state is replicated by the agents. When the cognitive empathy is selected, the agent shows a higher level of cognition of the game situations and the agent's empathic reaction is not necessarily the same as the user's affective state. For modeling, they used Naïve Bayes, decision tree, and SVM classification methods. During the interaction, after each empathic feedback message they evaluate the agent's empathy through user's self-report. A snapshot of the game environment used in their experiment is shown in Figure 2.16.



Figure 2.16: Environment setup used in [MRP08].

Ochs et al. [OSP10] proposed a BDI-like model of emotions (namely satisfaction, sadness, frustration, irritation, and anger) based on empirical and theoretical analyses of the users' conditions of emotions elicitation. Using the emotion model, they guess the most probable user emotion every time an event happens. Then they mimic the same emotion with the corresponding emotional facial expressions of a 3D character (developed in Orange Labs). Therefore, they modeled a combination of the emotional and cognitive empathy. They performed an experiment in three conditions of empathic, non-emotional, and non-congruent emotional (i.e., the character

expresses an emotion with the opposite valence to the user's emotion). Results show that (1) the empathic character is perceived more jovial, expressive, and cheerful than non-emotional version; (2) the non-congruent version is perceived less pleasant, compassionate, expressive, jovial, and cheerful than the empathic version; (3) the non-congruent virtual agent is perceived more irritating, strange, cold, and stressful than the empathic version; (4) the non-congruent agent's facial expressions are less appreciated than non-emotional and empathic versions; while (5) the empathic agent's facial expressions are perceived more natural, and less perturbing and exaggerated than the incongruent agent.

Moridis et al. [MEM12] implemented a virtual tutoring system, in which a 3D character provides emotional and cognitive empathy with the students during a problem solving experiment. The virtual tutor recognizes the emotional facial expressions of the student (namely happy, sad, and fear) using the FaceReader¹ software. At the same time, two experts recognize user's facial emotions manually (i.e., Wizard-Of-Oz) and if all three recognizers agree on an emotion, the empathy model is triggered so the character empathizes with the student. In this research, three empathy models are compared: (1) ECA displays neutral facial expression and vocal tone; (2) ECA displays emotional empathy with replicating the user's affective states using facial expressions and vocal tone; or (3) ECA displays cognitive empathy by encouraging behaviors in face and voice tone. So, if the user is sad or feared, ECA first displays the same emotion, then happy. If the user is happy, ECA first displays the same emotion, then neutral. Figure 2.17 shows the 3D character used in this research. Results show that, the emotional empathy reinforces the student's emotion. In the emotional empathy case, students empathized back to the agent by expressing the

¹<http://www.noldus.com/human-behavior-research/products/facereader>

same emotions again. Also, when the ECA performs a combination of emotional and cognitive empathy, it induced fear emotion to neutral.



Figure 2.17: The 3D character used in [MEM12].

The above studies showed that combining different individual empathy types and applying that to virtual characters affects positively on the users' affective states, likability of the character, perceived emotions from the character, expressiveness of the character, perceived pleasure from the character, and stress reduction.

2.4 Automatic Gesture/Expression Generation

Two main strategies are reported in the literature for automating the gesture generation:

1. **rule-based** approach, in which the recorded human behaviors are analyzed manually, from which rules are hand-crafted for gesture selection [CVB01, LMR06, NPI07a, BPI07, LM12]. For example, the facial displays for the Greta agent [DCP02] were selected using manually hand-crafted rules technique, in which rules to map from emotional states to facial displays were derived from the literature on facial expressions of emotion. Similarly, Cassell et al. [Cas01] selected gestures and facial expressions based on the rules derived from North American non-verbal behavior studies. In this type of modeling, behavior models are generated based on average behaviors of a range of people.

2. **machine learning** approaches, in which recorded human behaviors and surface-texts are annotated (manually or automatically), from which models of different gestures/expressions are inducted [LM09, LPNM09, FO08]. Such systems are able to produce more naturalistic output than a rule-based system [FO08, LPNM09], and can also easily model a single individual. Cassell et al. [CNB⁺01], for example, used this technique to choose posture shifts for the REA agent based on the annotated behaviors of speakers. More recently, Kipp [Kip05] used a similar technique to generate agent gestures based on annotated videos of speakers.

A multi-modal corpus is used in both strategies, which is an annotated collection of coordinated content on communication channels, such as transcript, speech, gaze, hand gesture, and body language. The multi-modal corpus is generally based on recorded human behaviors. The pragmatic context, under which each item of the corpus was created, must be known, i.e., the corpus must include all contextual information that the generator might use to choose among alternatives in a given situation. Also, the content of different channels must be linked to each other so that the generator can produce properly coordinated output.

In the next two sections, I will review some of the latest related works, in which the above approaches are applied.

2.4.1 Rule-Based Approaches

Cassell et al. [CVB01] developed the Behavior Expression Animation Toolkit (BEAT), which allows animators to input typed text wished to be spoken by a virtual character, and to obtain as output appropriate non-verbal behaviors synchronized with the speech. Since many of the state of the art in automatic gesture generation have implemented their systems based on the BEAT, we discuss the BEAT in details.

BEAT uses linguistic and contextual information contained in the text to control the movements of the hands, arms, face, and voice intonation. Modeling is performed by rules derived from the research in non-verbal conversational behaviors. For example, speakers express gestures along with the words. Also, listeners nod when the speaker’s gaze shifts. BEAT is written in Java and is based on an input-to-output pipeline approach with XML as the primary data structure. The system is real-time (i.e., time to produce an utterance is less than the natural pause between speaker turns, 500 - 1000 ms). An overview of the BEAT is shown in Figure 2.18.

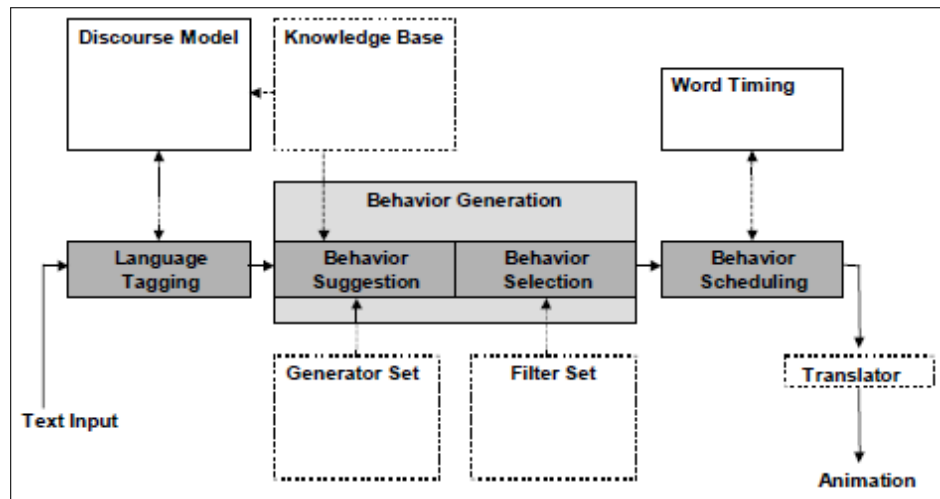


Figure 2.18: The BEAT architecture [CVB01].

The main processing modules of the BEAT project are as follow:

Knowledge Base: (1) adds the basic knowledge about the world to what we can understand from the text itself, and therefore allows (2) inferring from the text, and consequently (3) specifying the gestures representing it, and the times when emphasis is needed. Common gestures include the beat (i.e., a formless flick of the hand), deictic (i.e., pointing gesture), contrast, and iconic gesture (represents some object or action). These gestures are added to the database by the animator.

Language Tagging: the largest language unit is the *utterance* (i.e., an entire paragraph of input), which is broken into *clauses* (each of which represents a propo-

sition). Clauses are further divided into the *theme* (i.e., part that creates a coherent link with a preceding clause) and *rheme* (i.e., part that contributes some new information to the discussion) [Hal77]. Identifying the rheme is important since gestural activity is usually found within the rheme of an utterance [DRSV02]. The language tagging module uses the location of verb phrases within a clause and information about which words have been seen before in previous clauses to assign information structure. The next smallest unit is the *word* phrase, which either describes an *action* (i.e., verb phrase) or an *object* (i.e., noun phrase). The language module uses action and object databases. If an exact match of the verbs and objects are not found, an embedded word ontology module (WordNet [MBF⁺90]) is used to create a set of hypernyms (i.e., a related, but a more generic or broader term). The tagger uses the EngLite parser from Conexor² to supply word categories and lemmas for each word. The module also keeps track of all previously mentioned words and marks each new noun, verb, adverb or adjective as *new* if it has not been seen before. This “word newness” helps to determine which words should be emphasized by intonation, eyebrow motion or hand gesture. Also, if two words are in contrast with each other, that pair is tagged with the *contrast* tag.

Behavior Suggestion: operates on XML trees produced by the *Language Tagging* module by augmenting them with suggestions for appropriate non-verbal behavior. Any non-verbal behavior that is possibly appropriate is suggested independent of any other. The resulting generated behaviors will be filtered down in the next stage of processing to the final set to be animated.

Gesture Generator Set: a set of behavior generators are included in this module:

²www.conexor.fi

1. **Beat Gesture Generator:** beats are default gestures that are used when no information is available to generate a more specific gesture. Research shows that beats occur in 50% of the gestures observed in most contexts [McN92]. Thus, beats are given the lowest priority, so they will only be selected in absence of other gestures. The beats occur mostly when the speaker is introducing new material to the listener (rheme).
2. **Surprising Feature Iconic Gesture Generator:** determines if any of the *objects* identified by the *Tagger* within the *rheme* have unusual features, and for each generates an iconic (representational) gesture.
3. **Action Iconic Gesture Generator:** determines if there are any actions (verb phrase roots) occurring within the *rheme*, for which gestural descriptions are available in the action knowledge base. For each such action, an iconic gesture is suggested.
4. **Contrast Gesture Generator:** if there are exactly two objects being contrasted, special contrast gesture is suggested. Otherwise beats are suggested for contrast items.
5. **Eyebrow Flash Generator:** eyebrow raises can indicate new material [PBS96]. This generator suggests eyebrow raises when the character is introducing *objects* within the *rheme*.
6. **Gaze Generator:** Cassell et al. [CTP99] studied the relationship between eye gaze, theme/rheme, and turn-taking, and defined a few rule for controlling the gaze behavior of a conversational character (e.g., at beginning of the utterance gaze away from user).
7. **Intonation Generator:** assigns accents and boundary tones based on a theme-rheme analysis described by Prevost and Steedman [PS94] (e.g., high intonation on new objects).

Behavior Selection: analyzes the tree of gesture suggestions and filters them down to the set that will actually be used in the animation. In general, filters can reflect the personalities, affective state and energy level of characters by regulating how much non-verbal behavior they exhibit. For example, a Priority Threshold Filter removes all behavior suggestions whose priority falls below a user-specified threshold.

Behavior Scheduling and Animation: in general, there are two ways to achieve synchronization between a character animation and a the character’s speech (either through a TTS engine or from recorded audio samples): (1) estimate word and phoneme timings and construct an animation schedule prior to execution; (2) assume the availability of real-time events from a TTS engine and compile a set of event-triggered rules to govern the generation of the non-verbal behavior. Both of these approaches are used in BEAT. BEAT was implemented on an example news reporter character shown in Figure 2.19. However, the system was not evaluated in the sense of user acceptance and naturalness of the character.



Figure 2.19: Example BEAT application snapshot [CVB01].

More recently, Lee et al. [LMR06] used the BEAT toolkit to develop a rule-based non-verbal behavior generator that analyzes the surface text as well as the turn-taking and affective state of the ECA. They studied the uses of non-verbal behaviors in video clips of people conversing, and annotated the utterance dialog

acts (affirmation, negation, contrast, intensification, inclusivity, obligation, listing, assumption, possibility, response request, and word search).

based on the video analyses, they hand-crafted some rules, each of which associates a set of words with the non-verbal behaviors that are usually expressed along with them, and some priorities for each rule, which resolve conflicts between rules that could co-occur (numbers indicate the priority):

1. “*Interjection*: Head nod, shake, or tilt co-occurring with these words: yes, no, well.
1. *Negation*: Head shakes and brow frown throughout the whole sentence or phrase when these words occur: no, not, nothing, can’t, cannot.
2. *Affirmation*: Head nods and brow raise throughout the whole sentence or phrase when these words occur: yes, yeah, I do, I am, we have, we do, you have, true, OK.
3. *Assumption/Possibility*: Head nods throughout the sentence or phrase and brow frown when these words occur: I guess, I suppose, I think, maybe, perhaps, could, probably.
3. *Obligation*: Head nod when these words occur: have to, need to, ought to.
4. *Contrast*: Head lateral movement to the side and brow raise when these words occur: but, however.
4. *Inclusivity*: Lateral head sweep when these words occur: everything, all, whole, several, plenty, full.
4. *Intensification*: Head nod and brow frown when these words occur: really, very, quite, completely, wonderful, great, absolutely, gorgeous, huge, fantastic, so, amazing, just, quite, important.

4. *Listing*: Head lateral movement to the sides before and after the word ‘and’ when saying: X and Y.
4. *Response request*: Head moved to the side and brow raise when saying: you know.
4. *Word search*: Head tilt, brow raise, and gaze away when these words occur: um, uh, well.”

If there are two rules that overlap with each other, the one with a higher priority will be selected. They use Function Markup Language (FML) and Behavior Markup Language (BML) as part of the input and output messages. FML specifies the communicative and expressive intent of the agent (e.g., affect, coping, emphasis, turn). BML describes the verbal and non-verbal behaviors an agent executes (e.g., head, face, gaze, body, gesture, speech, lips, animation). The system architecture is presented in Figure 2.20.

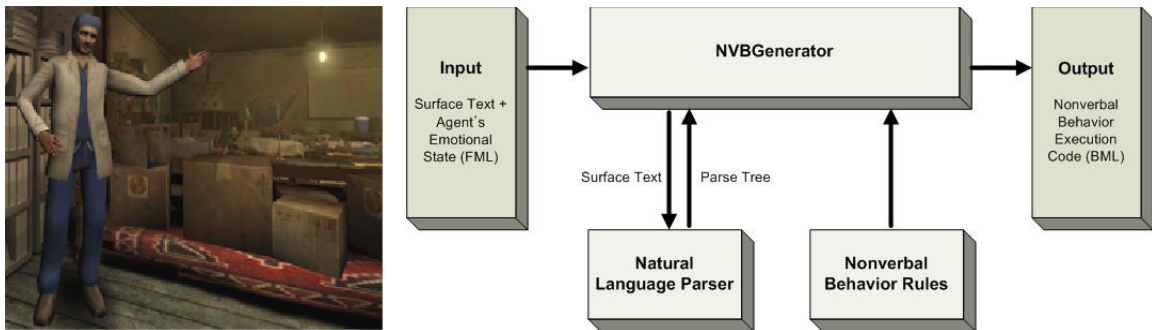


Figure 2.20: The system architecture and interface snapshot in [LMR06].

As shown in Figure 2.20, the module is incorporated into the SASO and Smart-Body character system as an example application, however, the user acceptance and naturalness of the gestures are not evaluated. Moreover, the system is limited in the sense that the features used in FML (e.g., affect) are hard-coded by the animator at the run-time and there is no realtime recognition, which limits the interacting ability of the character.

In a research by Neviarouskaya et al. [NPI07b], they extended the research done by Lee et al. [LMR06] and Cassell et al. [CVB01] with recognition and interpretation of affect communicated through text messaging. They developed an Affect Analysis Model to handle “not only correctly written text, but also informal messages written in abbreviated or expressive manner.” They proposed a rule-based approach which processes each sentence in different stages including word/phrase/sentence-level analyses. For affect categorization, the authors used not only affective words from WordNet-Affect [SV04], but also an affective lexicon derived from the evaluation of the semantic similarity between generic terms and affective concepts. They used a subset of emotional states defined by Izard [Iza77]: anger, disgust, fear, guilt, interest, joy, sadness (distress), shame, and surprise. The communicative function categories they used are greeting, thanks, posing a question, congratulation, and farewell.

They added the following information to the database: (1) words referring directly to emotions, mood, traits, cognitive states, behavior, attitude, sensations, (2) words that can express human affective states (e.g., beautiful, violate), (3) words showing dialog acts (functions), (4) interjections (e.g., wow, yay) that show an unexpected emotion, (5) 112 modifiers (e.g., very, extremely) that show the emotion strength. e.g., adverbs have an impact on neighboring verbs, adjectives [BIC⁺07], adverbs.

As shown in the system architecture presented in Figure 2.21, in the first stage, the sentence is tested for occurrences of emoticons, abbreviations, acronyms, interjections, “?” and “!” marks, repeated punctuation and capital letters. If found, no further analysis of affect in text is performed, but, if there are multiple emoticons or abbreviations in the sentence, the dominant emotion is selected based on these rules: “(1) when emotion categories of the detected emoticons (or abbreviations) are the same, the higher intensity value is taken for this emotion; (2) when they are different, the category (and intensity) of the emoticon occurring last is dominant.” If there are

no emotion-relevant emoticons in a sentence, the sentence is sent to the parser with the emoticons and abbreviations removed from the text; and non-emotional abbreviations and acronyms replaced by their complete form (e.g., “I’m [am] stressed bc [because] I have frequent headaches”).

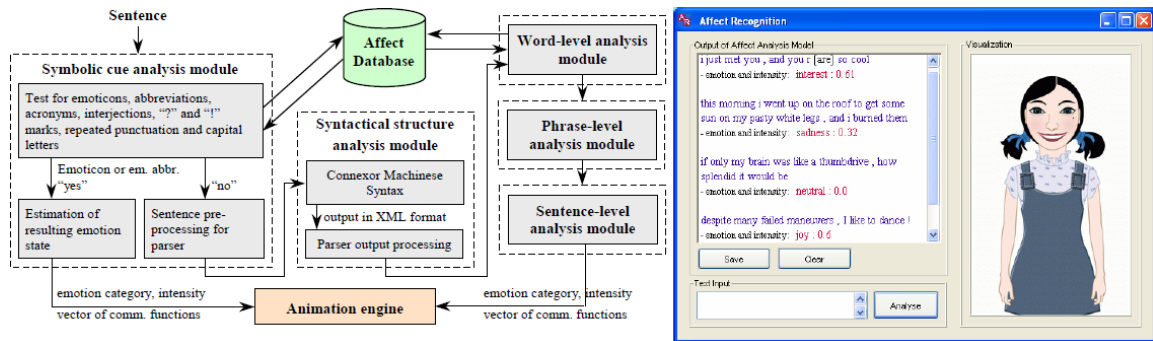


Figure 2.21: The system architecture and interface snapshot in [NPI07b].

The second stage is devoted to syntactical structure analysis using the Connexor Machine Syntax parser³, which analyzes the sentences, including word base forms, parts of speech, dependency functions, syntactic function tags, and morphological tags. In the third stage, for each word in the database, the affective features of a word are represented as a vector of emotional state intensities $e = [\text{anger, disgust, sadness, fear, guilt, interest, joy, shame, surprise}]$. In the fourth stage, phrase-level analysis is performed based on some predefined rules. The purpose of this stage is to detect emotions involved in phrases, and then in subject, verb, and object formations. During this stage, rules are applied and each formation is represented as a unified vector showing its emotion. In the fifth stage, the overall emotion of a sentence and its intensity are estimated. The emotional vector of sentences (or clauses) are generated from subject, verb, and object formation vectors using some pre-defined rules.

³ <http://www.connexor.com/>

In an evaluation experiment, in 79.4% of all sentences, the emotion category obtained by proposed algorithm matches with at least one of three human raters' annotation. In 70%, system output agrees with at least two annotators.

Breitfuss et al. [BPI07] introduce a system that automatically adds non-verbal behavior to a given dialog script between two virtual embodied agents. It transforms a dialog in text format into an agent behavior script enriched by *eye gaze* and *head nod*. The resulting annotated dialog script is then transformed into the Multi-modal Presentation Markup Language for 3D agents (MPML3D) [NPAI06], which controls the multi-modal behavior of human-like agents. An important feature of their system is that they generate the behavior not only for the speaker agent but also for the listener agent. This system consists of three modules: (1) Language Tagging module, which takes the input dialog text and uses the language module from the BEAT toolkit [CVB01] to annotate linguistic and contextual information, and to suggest appropriate non-verbal behaviors, (2) Non-Verbal Behavior Generation module, which adds *eye gaze* and *head nod* to the annotated input sentence, and (3) Transformation to MPML3D module, which produces an MPML3D file to control the behavior of the 3D agents. In order to avoid conflicts between certain gaze behaviors, like looking in two different directions at the same time, they assigned priorities to the behaviors. Gazing behaviors follow the same rules as in BEAT. In this system, a head nod is a basic gesture type for the listener. It is the gesture with the lowest priority and is used when no other, more specific gesture can be suggested.

Foster and Oberlander [FO08] presented a system that uses corpus-based selection strategies to automate the animation of the head and eyebrow movement of an ECA called RUTH [DRSV02]. Although their approach is data-driven, they do not use machine learning for modeling, but count the frequencies of behavior occurrences and either choose the behavior with the highest frequency or use a weighted choice.

The syntactic and pragmatic information they used include: (1) the user-preference evaluation of the object being described (positive or negative); (2) whether the fact being presented was previously mentioned in the discourse or is new information; (3) whether the fact is explicitly compared or contrasted with a feature of the previous tile design; (4) whether the node is in the first or second clause of a two-clause sentence; (5) the surface string, with words replaced by semantic classes or stems drawn from the grammar ; (6) and any pitch accents specified by the text planner. They annotated the speaker’s facial displays in each video to an XML document, considering five types of motion: eyebrow raising/lowering; eye narrowing; head nodding; head leans and turns.

They found the nodding and brow raising the most frequent and effective contextual features. In negative contexts, eyebrow raising, eye narrowing, and left leaning were more frequent; in positive contexts, right turns and brow raises had higher frequencies. In the first half of two-clause sentences, brow lowering and upward nodding were more frequent, while downward nodding and right turns were more frequent in the second clause. Output expressions are selected in one of two ways: taking the highest-probability option or making a weighted choice. As an example of the two generation strategies, consider a hypothetical context, in which the speaker made no motion 80% of the time, a nod 10% of the time, and a brow raise the other 10% of the time. In this context, the majority generation strategy would choose the majority option of no motion 100% of the time, while the weighted strategy would choose nothing with probability 0.8, a nod with probability 0.1, and a brow raise with probability 0.1. Figure 2.22 shows the expressions on the virtual character they used.

They compared the outputs produced by above two strategies in terms of precision, recall, F-measure, and accuracy. Results showed that the majority strategy scored higher on all measures, while the human subjects tended to prefer the output of the

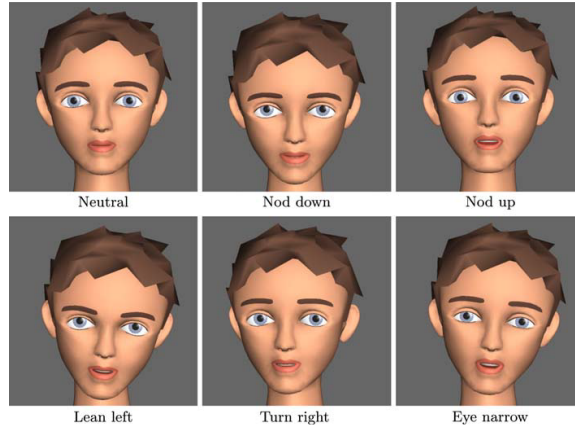
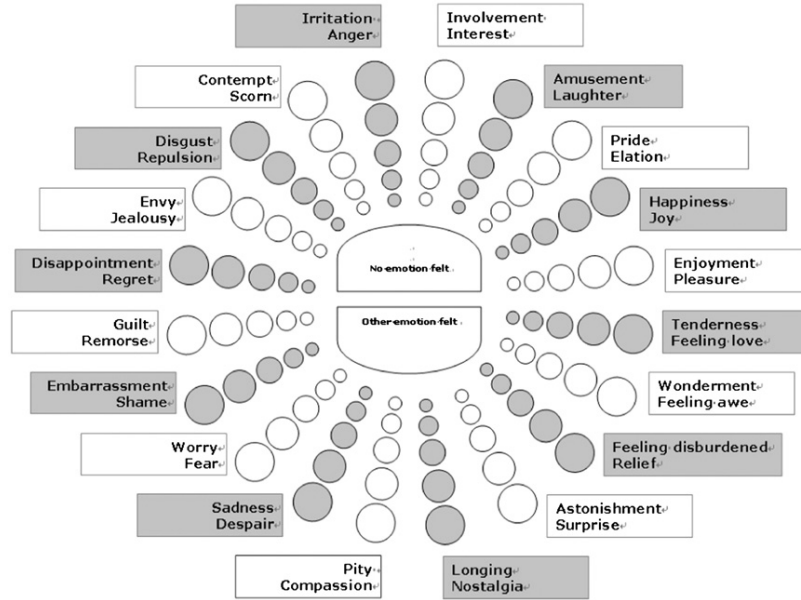


Figure 2.22: Output expression in [FO08].

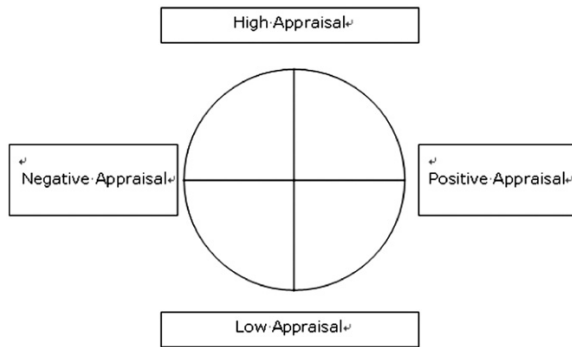
weighted strategy. This shows that the subjects would prefer generated output that reproduced more of the variation in the corpus, regardless of the corpus-similarity scores.

More recently, Li and Mao [LM12] proposed a computational framework that enables virtual agents to convey different emotional expressions to users through *eye movements*. They propose a rule-based approach to generate emotional eye movements based on the Geneva Emotion Wheel [Sch05] shown in Figure 2.23. Results show a high rate of recognition of the agent intended emotion. The Geneva Emotion Wheel (GEW) is a theoretically derived and empirically tested instrument to measure emotional reactions to objects, events, and situations [Sch05]. As shown in Figure 2.23, emotions are symmetrically arranged in a wheel shape with the axes being defined by two major appraisal dimensions: high/low appraisal and positive/negative appraisal.

The overall framework of this research is illustrated in Figure 2.24. The emotional eye movement synthesis is mainly composed in two phases: First, the Cohn-Kanade AU-Coded facial expression database is analyzed to derive Facial Animation Parameters (FAPs) values. The FAPs are defined to allow the definition of a facial shape and its animation for reproducing expressions, emotions, and speech pronunciation



(a) GEW primary and secondary emotins.



(b) Appraisal dimensions.

Figure 2.23: Geneva Emotion Wheel and appraisal dimensions used in [LM12].

[Ost98]. Second, for pupil size, blink rate and saccade, an eye tracker is used to capture and analyze real-time eye movement data, and derive FAPs values from raw eye-track data. The resulting FAPs values are used to realize the desired emotion by cheek, nose, eyebrow, eyelid and eyeball animation associated with the eye movement. Also, hand-crafted rules are employed to generate emotional eye movement for primary and intermediate (secondary) emotions of the virtual agent based on the GEW model.

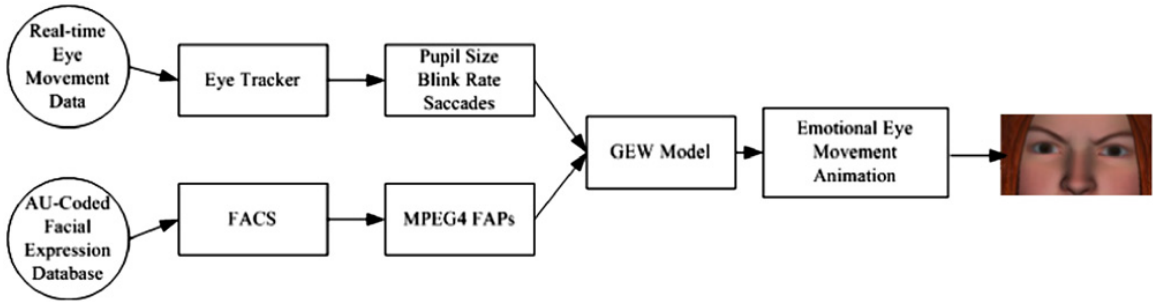


Figure 2.24: Overall framework proposed by [LM12].

Rules are extracted based on experiments with a pictorial stimuli (International Affective Picture System (IAPS) [LBC97]) accompanied by an audio stimuli (International Affective Digitized Sounds (IADS) [BL07]) in three emotion categories of neutral, negative arousing (negative valence), and positive arousing, on male and female subjects. Some of the rules include:

- Pupil size is larger in positive and negative state than in neutral state.
- The larger the valence is, whether positive or negative, the larger the pupil size is.
- The pupil size is larger for females than for male subjects during neutral stimuli.
- Blink rate decreases in positive and negative state compared to neutral state.
- Blink rate is slower during the negative and positive than during the neutral stimuli.
- Diagonal saccade movements occur more in negative emotions than in positive ones.
- Up, down, left, and right saccade happen more in positive emotions than in negatives.
- In neutral state, the gaze target is usually fixed to the straight direction.

2.4.2 Machine Learning Approaches

One major drawback of the previous rule-based systems is that the rules have to be hand-crafted, therefore, the person who creates the rules should have a broad knowledge of the modeling aspect. However, as more and more input features are added, it becomes harder to determine the effects of each feature on the overall outcome. To address this limitation Lee et al. [LM09] used a machine learning technique, Hidden Markov Model (HMM) [Rab89], to create a head nod model from annotated video corpora of face-to-face human interactions, based on the syntactical features of the dialog surface text. HMM is a statistical model that is used for learning data-driven models, in which a sequence of observations is available. For this work, the input is a sequence of feature combinations (vectors) representing each word. The sequential property of this problem is the reason of using HMMs to predict head nods.

Lee et al. [LM09] used the system proposed in [LMR06, NPI07b, NPI07a] and focus on the first step of the head movement generation process, which is predicting when the speaker should use head nods. The stages of the learning process are shown in Figure 2.25.

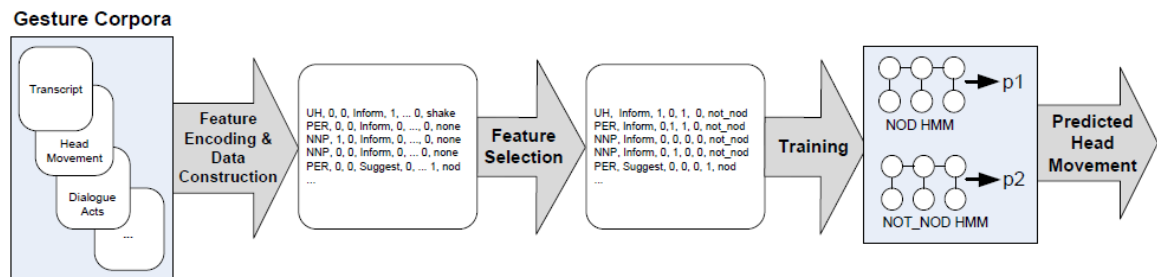


Figure 2.25: The stages of the learning process proposed by [LM09].

As shown in Figure 2.26, for data construction, they used the speaker's surface text, the utterance dialog acts, and the observed head movements (nod, shake, nod-shake, other, and none) during the utterance. They also obtained the part of speech tags, phrase boundaries (e.g., start/end of verb phrases and noun phrases), and key

lexical entities (i.e., keywords shown to be associated with head nods) by sending the utterances through a natural language parser (Charniak Parser [Cha00]). The number of features is reduced by counting the frequency of head nods that occurred with each feature and selecting a subset of features that have the highest co-occurrence frequency. The list of final features selected for training is: sentence start, noun-phrase start, verb-phrase start, and key lexical entities. After aligning each word of the utterances with the selected features, each sequence of three words are put together to form a set of trigrams. These trigrams compose the dataset. For each trigram, using the majority vote method, they determine the head gesture. For example, if two or three out of three words co-occurred with a nod, the trigram was classified as a nod, and the same for other head movements. To determine the classification category of each trigram, two HMMs are trained: a ‘*Nod*’ and a ‘*Not Nod*’ HMMs (i.e., other head movements than nod). The output of an HMM is a probability that a sample is classified as a specific head gesture.

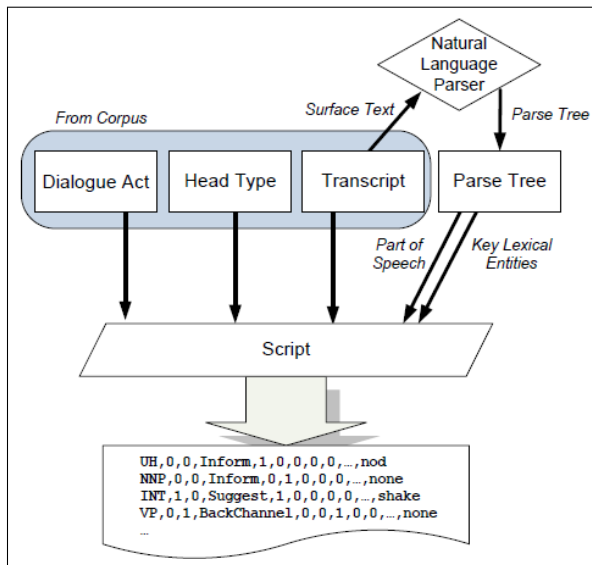


Figure 2.26: Data construction process in [LM09].

They show that the head nod model predicts head nods with accuracy of 0.8528, precision of 0.8249, recall of 0.8957, and F1-measure of 0.8588. This work can be

extended by learning the patterns of other non-verbal behaviors. Also, conducting evaluations with human subjects helps investigating if the head movements generated by the model are perceived to be natural. Moreover, adding some interactive features of the user (e.g., facial expressions) to the modeling process increases the naturalness and believability of the system.

Lee et al. [LPNM09] expanded the head nod model they proposed earlier [LM09] by using affective information during the learning process to build a domain-independent model of speaker's head movements, which predicts the speaker's head nods. They investigate the use of affective information during the learning of speaker head nod models. They perform this by using the detected emotion label of each word in the surface text as well as the emotion label over the whole sentence during training process. The Affect Analysis Model (AAM) presented in [NPI07b] was the rule-based system aimed for the recognition of ten emotions from text (i.e., anger, disgust, fear, guilt, interest, joy, sadness, shame, surprise, and neutral).

They trained the models with three different conditions: (1) using no affective information, (2) using emotion label of each word, and (3) using emotion label of the whole sentence. Results show that using affective information, especially in *sentence level*, improves the prediction metrics compared to using no affective information. Accuracy increased to 0.8957, precision increased to 0.8909, recall increased to 0.9018, and F1-measure increase to 0.8963. This research suggests using sentence-level affective features of the surface text in automatic modeling the non-verbal gestures.

Finally, a different hybrid approach is suggested in [Kip06], which uses both hand crafted rules and machine learning to generate the gestures. Author uses pre-defined rules to add some default gestures to the character based on textual features of the script. Then, machine learning is used to create more rules. Nearest neighbor clustering algorithm was used to cluster similar patterns to the rules already available in the

system, and find new rules. Output models are grouped into four channels of facial expression, gaze, manual gesture, and head movement. This approach was applied to the COHIBIT system (a mixed-reality museum exhibit), in which text is sent to two virtual characters to be uttered. Evaluation results show precision of 0.338 and recall of 0.136 for the male character and precision of 0.326 and recall of 0.321 for the female character.

Comparing the above manually generated rule-based systems and automatically inducted models, the machine-learning approaches have several advantages over the rule-based ones: (1) the process is automated; (2) having a good understanding of the phenomena is still important, however with this approach, it is no longer necessary for the author of the model to have a complete knowledge of the complex mapping between the various features and behaviors; and last but not least (3) it is flexible and can be customized to learn non-verbal behaviors in a specific context, culture, age, personality, etc. [KTB⁺08].

In the related research of the machine learning approach to automate gesture generation, there are still some limitations that I aimed to address in this dissertation, such as: (1) they either model the non-verbal behaviors of the characters in a speaker or listener role, while both roles are needed in order to create a believable flow of conversation; (2) mostly, they use either the textual features or the visual features of the conversation in order to decide about the best non-verbal behaviors of the character, while a combination of these features can give more information about the context and the user's state; and (3) they model very few non-verbal behaviors including smile, head nod, and eye gaze, which are not enough for creating a natural conversation.

2.5 Facial Expression Generation

2.5.1 Facial Action Coding System (FACS)

The Facial Action Coding System (FACS) [EF78, EFH02] is a widely used facial coding system for discussing and measuring all visible facial movements. FACS describes facial activities in terms of muscle Action Units (AU), each of which is associated with the underlying muscles that cause the movement. FACS helps to understand all possible physical movements of the human face. AUs are grouped based on their location on the face and the type of facial action involved. The “upper-face” AUs include the eyebrows, forehead, and eyelids muscles; the “lower-face” AUs include muscles around the mouth and lips, and the “head and eye movement” AUs include the neck muscles, which move the head and the eye muscles, which move the gaze direction.

AUs act as multi-level switches, which can create custom expressions depending on which AUs are activated/deactivated at a certain time. Since not all expressions require the farthest reach of a muscle, intensity levels are used to create subtle movements of the face. Intensities are annotated from “0” to “E”, where “0” is the neutral face without any activated AUs, “A” is the weakest trace of the AU, and “E” is the maximum intensity.

The muscle groups underlying all facial movements form 30 AUs for facial expressions, 14 AUs for head and gaze directions, and 2 AUs for head movements. Figure 2.27 shows facial muscles and their associated AU numbers.

Specific AUs in both sections can also be activated unilaterally to generate asymmetric expressions. Although a hard topic to master, trained FACS coders are known to identify movements of the human face so well as to recognize which AUs are activated at a certain frame (or point in time) of an expression, including the intensities

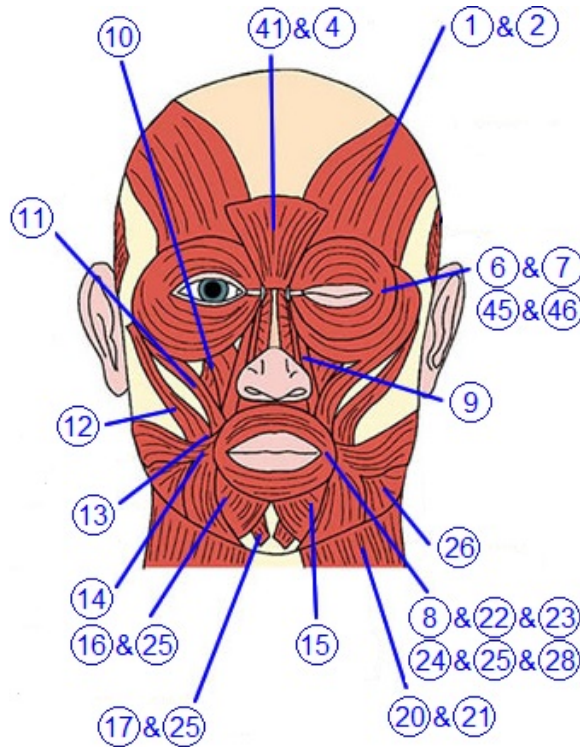


Figure 2.27: Sample AUs of the FACS.

of each AU. A FACS “score” of an expression consists of the list of AUs that produce it, accompanied with their corresponding intensities.

Emotional FACS (EmFACS) [FE83] is an extension of FACS, focused primarily on facial expression of the emotions. EmFACS provides subsets of AUs used to generate the six universal emotions identified by Ekman [ELF83], namely, fear, anger, surprise, disgust, sadness, and happiness, although, other emotional facial expressions, such as contempt, pride, and embarrassment, can also be depicted by combinations of FACS action units. Ekman and Friesen [FE83] studied people’s emotions around the world and found that people have an innate ability to generate and understand that set of facial expressions among all human cultures.

Expressive emotions that can captivate an audience are sought out by animators who want to create virtual characters, by psychologists who wish to understand a potential patient’s emotional distresses, and by affective computing researchers who

aim at generating believable virtual characters who can assist or entertain human beings. This has increased the importance of the universally accepted facial expressions, because well-rigged characters can be used to express the facial actions based on FACS to express cross-culturally readable expressions.

Although FACS is a widely accepted standard, it has a steep learning curve. Therefore, for researchers who are interested in controlling facial expressions, there is a need of easy-to-use learning tools.

An alternative to the FACS is the Facial Animation Parameters (FAPs) defined in the ISO MPEG-4 standard [Ost98, BCL00, Ost02] based on the Facial Definition Parameters (FDPs), that allow the definition of a facial shape and its expressions including emotional faces and speech pronunciation. FDPs represent 88 key points in a human face that are used to customize a model to a face and animation (i.e., FAPs that describe motion).

The FAPs are based on the study of minimal facial muscle actions that can represent a complete set of basic facial expressions. FAPs are expressed in terms of the Facial Animation Parameter Units (FAPUs), which correspond to fractions of distances between some key facial features. These units allow interpretation of the FAPs on any facial model.

2.5.2 Facial Expression Datasets

Researchers interested in human facial expressions - either the ones interested in correlating facial movements to the expression of emotions [EF78, EFH02, ELF83, EF74, Sch01], or the ones considering facial displays as social signals of intent [Cho91, Fri94] - typically base their research on large databases of human facial expression images and/or videos.

Facial databases are usually coded in a variety of meaningful ways, and used in a variety of disciplines, such as psychology or physiology, with research on facial expression recognition, facial expression generation, or emotion theories.

Several databases of human facial expressions have been developed. These databases provide standard sets of facial expression images and videos, including different emotional facial expressions and faces with specific activated AUs. For example, Karolinska Directed Emotional Faces [LFO98] is a set of 4,900 pictures of human facial expressions of emotions portrayed by 70 non-FACS certified individuals, each displaying 7 different emotional expressions (neutral, happy, angry, afraid, disgusted, sad, and surprised), each expression being photographed from 5 different angles.

Pictures of Facial Affect⁴ is another dataset consisting of 110 black and white photographs of FACS-based facial expressions.

UC Davis Set of Emotion Expressions [TRS09b] includes images of anger, embarrassment, fear, disgust, happiness, pride, sadness, shame, and surprise expressions portrayed based on FACS by 4 FACS-certified individuals – 2 females (1 Caucasian American, 1 African) and 2 males (1 Caucasian American, 1 African).

Montreal Set of Facial Displays of Emotion⁵ consists of emotional facial expressions portrayed by men and women of European, Asian, and African descent. The set contains expressions of happiness, sadness, anger, fear, disgust, and embarrassment in 5 levels of intensity as well as a neutral expression for each actor. Expressions were FCAS coded to assure identical expressions across actors.

Amsterdam Dynamic Facial Expression Set [VHFD11] is a set of 648 videos of nine emotional expressions: anger, disgust, fear, joy, sadness, surprise, contempt, pride and

⁴www.paulekman.com/product/pictures-of-facial-affect-pofa

⁵www.er.uqam.ca/nobel/r24700/Labo/Labo/MSEFE.html

embarrassment. Expressions are displayed by 22 non-FACS-certified North-European and Mediterranean models (10 females, 12 males).

SmartKom video database [SST02] consists of 448 multi-modal recordings of 224 persons showing joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, and neutral expressions.

Belfast Naturalistic video database⁶ contains 298 audio-visual clips from 125 speakers (31 males and 94 females) and the annotations of their affective facial expressions provided by 7 experts.

CK+ dataset [LCK⁺10] contains (1) video recordings of the facial behavior of 210 adults (18 to 50 years of age; 69% females; 81% Euro-Americans, 13% Afro-Americans, and 6% other groups). Individuals performed a series of 23 facial displays including single AUs and combinations of AUs; (2) FACS coding of the peak frames of 593 posed sequences; (3) image data from the pool of 593 sequences that had a nominal emotion label from the 7 basic emotion categories (i.e., anger, contempt, disgust, fear, happy, sadness and surprise).

MMI Facial Expression Database [VP10] consists of over 2900 videos and images of 75 subjects. The AUs presence (i.e., neutral, onset, apex or offset) is FACS coded in videos, and partially for the images.

Although these databases have been used successfully for facial expression recognition and synthesis, they have common limitations, such as: (1) only a limited number of facial movements are provided; (2) all the possible intensities of different expressions are not provided; (3) all combinations of AUs activation with different intensities are not provided for each face (i.e., it is difficult for a human actor to generate all combinations of different AUs); (4) because it is difficult for human posers to display exactly the same AU activation intensity, datasets are not always consistent across

⁶<http://sspnet.eu/2010/02/belfast-naturalistic/>

subjects [KTRS12, SE07]; (5) most of the provided emotional expressions are static (images), which can limit their usefulness to study gradual changes of facial muscle movements; and/or (6) although a few databases provide images and videos of individual AUs and of combination of AUs, most of the databases provide data on six to nine human facial expressions of emotions only.

In this dissertation, I posit that realistic facial expressions generated and validated based on FACS, can provide valuable additional data on facial expression generation and a platform for researchers to experience with. HapFACS (discussed in Section 3.1) aims at addressing some limitations of the current datasets by increasing the number of facial movements that can be activated and manipulated according to FACS on a 3D virtual character’s face. HapFACS can act as an infinite database of expressions, and can be used to create any possible combination of facial muscle movements on different virtual character models. HapFACS enables researchers to create both custom images and videos of facial expressions.

2.5.3 Virtual Character Animation

FACS is currently being used extensively in virtual characters that need to portray facial expressions. For example, Smartbody [TRM⁺08, Sha11], the virtual character animation system developed by researchers at the Institute for Creative Technologies (ICT), provides important aspects of realistic character modeling, such as locomotion, facial animation (11 AUs), speech synthesis, reaching/grabbing, and various automated non-verbal behaviors. SmartBody is a very powerful, but heavy-weight, software requiring programming experience and extensive effort to be integrated with researchers’ systems and applications.

FACSGen [KTRS12] is a software developed on top of the FaceGen⁷ virtual character software and it simulates 35 AUs. Whereas FaceGen faces are quite realistic and well rendered, some aspects of the FACSGen software can limit its extended use for the type of research we mentioned above: (1) it only implements 35 out of 46 AUs (30 facial AUs, 14 eye and head direction AUs, and 2 head movements); (2) it cannot activate the bilateral AUs asymmetrically; (3) characters do not have lip-synchronization abilities (which limits its appeal for researchers interested in speaking characters); (4) strong graphics expertise is needed to combine FaceGen faces with a character's body, to embed the character in another software, and to add lip-synchronization abilities; and (5) the system is not freely available to the research community.

In a research by Helmut and Leon [Beu11], the authors implemented 13 AUs on faces with soft-looking skins, which can simulate wrinkles on skin. They generated the AU expressions by using manually generated morphs for a virtual character. Although they do not provide any evaluations for the accuracy of the generated AUs, resulting expressions seem promising. However, in addition to the similar limitations enumerated for the FACSGen, Helmut and Leon's product neither supports embedding the character in other applications, nor exporting the generated facial expressions as images or videos.

Villagrasa and Sanchez [VS09] presented a 3D facial animation system named FAcE!, which is able to generate different facial expressions throughout punctual and combined activation of AUs. The FAcE! system is implemented on the 3DStudio Max platform. The resulting virtual model is able to activate single or combined AUs, express emotions, and display phonemes on lips. The AUs are generated using manually generated morphs and rigs for the virtual character. The FAcE! simulates a total of 66 movements and AUs, which can be activated both unilaterally and

⁷<http://www.facegen.com/>

bilaterally. It also can express four emotions namely, happy, fear, sad, and anger. However, making changes to their character’s skin, age, gender, ethnicity, lighting, and enabling lip-synchronized speech requires graphics expertise that IVA researchers might want to avoid acquiring. Moreover, it is not freely available to the research community either.

Alfred [BFA09] is a facial animation system, which uses a slider-based GUI, a game-pad, and a data glove for user input. Alfred uses the 23 AUs of the FACS for the description and creation of the facial expressions. The AUs are generated using manually generated morphs. This character also supports lip synchronization. However, the AU expression accuracy is not evaluated by certified FACS-coders, and the technical experience needed to setup the system and integrate it with other applications deter users from using it.

Wojdel and Rothkrantz [WR05] presented a parametric approach to generation of FACS-based facial expressions. They used 38 control markers on one side of a human subject’s face as well as positions of facial features, such as mouth-contour, eye-contour and eye-brows, and took frontal pictures of the face when single AUs were activated. The authors found mathematical functions of the marker movements for each single AU activation, and implemented 32 symmetric AUs. They used fuzzy logic to generate some rules to enable AU co-occurrence and prevent oppositions in AU activation (e.g., activating AU 51 and AU 52 at the same time are not anatomically possible). They evaluated the AU generation accuracy of their system with 25 subjects on Ekman’s 6 universal expressions mentioned above (anger, disgust, fear, happiness, sadness, surprise) and reported a 64% recognition rate. This software does not enable the user to activate the bilateral AUs asymmetrically; it do not have lip-synchronization abilities; integration of the head model to characters in real ap-

plications needs graphics expertise; and finally, the system is not freely available to the research community.

Digital Emily [ARL⁺09] was generated by filming an actress while she spoke and by capturing the motion of the actress' facial expressions showing different emotions, mouth movements, and eye movements. The actress was posed for thirty-three different facial expressions based loosely on FACS. Motion capture techniques were used to map marked point on the actress's face to the vertices of her face 3D model. A semi-automatic video-based facial animation system was then used to animate the 3D face rig. However, Digital Emily was rendered offline, involved just the front of the face, and was never seen in a tight closeup. Although the facial movements are not generated exactly based on FACS, no evaluation of the accuracy of the facial expression are provided.

More recently, Alexander et al. generated a new virtual character called Digital Ira [AFB⁺13] using a similar approach to the one they used in animating Digital Emily. Digital Ira is a real-time (Digital Emily was rendered offline), photoreal digital human character, which can be seen from any viewpoint, in any lighting, and can perform realistically from video performance capture even in a tight closeup. However, this virtual character is not available to the research community, also, integration of the head model to a real application needs graphical expertise.

Greta [PP01, BPN⁺10] and iFACE [ADJE06] are other virtual characters being used in facial expression research. These character systems develop, display and animate facial expressions based on FAPs and MPEG-4 standard [BCL00]. Greta can also simulate the skin wrinkles. However, since they are not based on FACS, it does not provide any simulations of the AUs and asymmetric animations.

Therefore, a simple and user-friendly system as well as an easy-to-use programming API that allows easy manipulation of virtual character's facial expressions based

on FACS can cover the limitations of the above systems and needs of the research community.

2.5.4 Haptek Avatar System

The Haptek⁸ software, developed by Chris Shaw, is a light weight avatar system, which is a popular software for many research groups currently working on embodied virtual characters [HJR10, AWL11, BT11, BTTS12, BST10, BR11, DMNP10, FMS⁺12, IY10, LSN10, NI10, NDRM10, PBCS10, SCB⁺10, SD12, THSh⁺11, VF10, CVG⁺10, SLN11, CVG⁺10]. Its popularity among academic groups working on ECAs is due to the fact that, unlike the high-resolution 3D characters seen in video games and digital animation movies, which require pre-scripting movements and many expensive graphic artists and animators to draw and render each facial expression, Haptek offers low cost programmable 3D-characters that have the best lip synchronization on the market. In addition, characters portraying features of different ethnicities can easily be designed.

Haptek has great accessibility, and provides an excellent solution for researchers who may not necessarily be interested in the most photo-realistic virtual face possible, but who do want to see the effects of facial movements on an anthropomorphic face. Haptek characters have three versions of head-only, torso, and full-body, which provide the ability to use them in a variety of systems or applications, in which non-verbal gestures, other than facial expressions, are needed as well.

Haptek characters can be integrated easily to applications and enable them to have a real-time talking character with lip synchronization based on speech synthesizers or pre-recorded sound files. However, Haptek suffers from not having an accessible programming API and interface for FACS-based facial expression generation, which

⁸<http://www.haptek.com>

is the crucial part of emulating face-to-face interaction with non-verbal behaviors. Therefore, a solution is needed, which addresses this issue by providing an API to control Haptek characters' facial expressions in real-time, based on FACS, and to control facial expressions of emotions based on EmFACS [FE83].

2.6 Summary of Literature Review

As I discussed in this chapter, in the state of the art of rapport and empathy modeling, generally, non-verbal rapport and empathy are modeled different ways: rapport, motor empathy (mimicry), emotional (affective) empathy, cognitive empathy, verbal empathy, and different combinations of these empathy types.

However, there are multiple limitations in these approaches that are addressed in this dissertation: **(1)** non-realtime recognition of the affective state of the subjects, which is addressed using realtime facial/head/body gesture recognizers; **(2)** not using the facial expressions as the most important modality in human behavioral judgment [AR92] in both recognition and expression phases, which is addressed by using highly expressive characters that are capable of expressing different facial expressions based on FACS, also as mentioned before, realtime facial expression recognizers are used to perceive the emotional state of the users; **(3)** using rapport and empathy in non-emotional contexts, which is addressed by using the rapport and empathizing ability in health counseling context, which can be highly emotional for people; **(4)** unclear mapping of the subject's recognized features to the character's reactions, which is addressed by creating individual models for each non-verbal behavior; and **(5)** using characters with low expressivity, which limits conveying the non-verbal behaviors, that is addressed by using highly expressive 3D characters.

In the state of the art for character animation, especially in facial expression animation, the need of an accessible and easy-to-use solution is highly sensed, which

enables us to (1) control characters' facial expressions in real-time based on a standard, such as FACS (and EmFACS); and to (2) generate standard, believable, and reproducible facial expressions (images and videos) on characters of different ethnicities, genders, and ages. I addressed this need with implementing the HapFACS software and API (discussed in Section 3.1).

In the state of the art of automatic gesture generation, generally, two approaches are taken for modeling the non-verbal gestures and automating the gesture generation: **(1)** using hand-crafted rule-based models, and **(2)** using machine learning techniques.

Hand-crafting the rules requires some social science expertise in human communication and is a time consuming process to map multiple input features to the output gestures (time complexity increases exponentially when the number of input features increases). Machine learning techniques enable us to either generate these rules automatically, or generate probabilistic models for different gestures which map the input features to the output gestures. Whereas Machine Learning (ML) techniques have recently began to address the limitations of early rule-based techniques, current research using the ML approach has a number of limitations of its own.

Therefore, in this dissertation, I aim to use machine learning to address the fore-mentioned limitations of the hand-crafted rule-based models, and to address the current limitations of the machine learning approaches taken to model non-verbal behaviors including: **(1)** not taking into account interactive features, such as user's facial expressions, gaze, and head movements, for decision making, which I addressed by using the outputs of a realtime facial expression recognizer and eye gaze tracker; **(2)** modeling very few non-verbal modalities, namely head nod, gaze, and smile, while other non-verbal are as important and can help generating believable characters, such as *head shake*, *head nod-shake*, *body lean* (e.g., left/right, forward), *eyebrow movements* (e.g., up and down), *emotional facial expressions* (e.g., happy, sad, sur-

prised, angry, and disgust), and *hand gestures* (e.g., formless-flick, point, contrast, iconic, close, and open); **(3)** using either visual or textual features for learning the non-verbal behaviors, while a combination of both types is used in this dissertation in order to increase the amount of information perceived from the input data and improve the performance of the modeled gestures; and finally **(4)** modeling either speaker or listener non-verbal gestures, while both roles are important in expressing gestures. People express different non-verbal behaviors when they are speaking and listening with different patterns. So, in this dissertation, I addressed this limitation by generating individual models for speaker and listener roles. For example, I generated a head nod model for the speaker and another head nod model for the listener.

Haptek Character Animation**3.1 HapFACS**

One of the most important media in human social communication is the face [AR92], and the integration of its signals with other non-verbal and verbal messages is crucial for successful social exchanges [All02]. However, while much is known about the appearance and human perception of facial expressions, especially emotional facial expressions, emotion researchers still have open questions about the *dynamics* of human facial expression generation and their perception [Sch01].

There are therefore advantages to have software, such as HapFACS, that can emulate facial expression generation via 3D animated characters: 1) to animate intelligent virtual agents (IVA) that can naturally portray human-like expressions when they interact with humans; and 2) to develop and test emotion theories of (human) dynamic facial expression generation.

Indeed, IVAs – which simulate humans’ innate communication modalities, such as facial expressions, body language, speech, and natural language understanding, to engage their human counterparts – have emerged as a new type of computer interfaces for a wide range of applications, e.g., interactive learning [SAD⁺06], e-commerce [BST10], virtual patients [KPG⁺07], virtual health coaches [LAYR13], video games [ADJE06] and virtual worlds [MGR03].

IVAs therefore need to be able to portray appropriate levels of social realism, e.g., portray believable facial expressions, and gestures. However, many IVA researchers interested in modeling social intelligence do not have the graphics expertise for the difficult task of generating and animating 3D models, in order to showcase the embodiment of their models of social intelligence. Similarly, psychologists working on

facial expression generation rely on the analysis of large corpora of videos and images of human expressions, but do not have the means to test their theories in a systematic fashion.

Because it is difficult to animate 3D characters, many researchers buy and use third-party software. To date, there are a handful of systems that provide the ability to manipulate facial expressions on 3D characters in terms of Action Units (see Section 2.5.3), most of them are not freely available to the research community and/or require computer graphics expertise, with characters often difficult to integrate with one's system.

I developed HapFACS, a free software and API, based on the Haptek 3D-character platform running on the free Haptek Player¹, to address the needs of IVA researchers working on 3D speaking characters, and the needs of psychologists working on facial expression generation. HapFACS provides the ability to manipulate the activation – in parallel or sequentially – of combinations of the smallest groups of facial muscles capable of moving independently (referred to as Action Units), which can help the development of facial expression generation theories, as well as the creation of believable speaking IVAs. HapFACS has been very well received by affective computing researchers [LAYR13, HDC15, AN14, ALY14, MNP13, MPNP13, ALYR13].

HapFACS² is a free stand-alone software and API implemented in the C# language. It uses virtual characters rendered in the free HaptekPlayer software, and simulates FACS AUs on these characters. Currently, HapFACS includes more than

¹<http://www.haptek.com/>

²Affective Social Computing Lab has an agreement with Haptek to be able to distribute HapFACS source code under a free non-exclusive license only for academic, research or non-profit centers and only for personal and noncommercial purposes as per the license terms provided at <http://ascl.cis.fiu.edu/hapfacs.html>. HapFACS requires Haptek Player available at www.haptek.com, which is also free for non-commercial use, per its license agreement.

165 characters and hair styles, and users can use Haptek PeoplePutty software to create new characters and import them to HapFACS.

3.1.1 HapFACS Functionalities

HapFACS enables users to: **(1)** control 49 AUs (12 upper-face, 21 lower-face, and 16 head/eye position) of characters' faces; **(2)** activate individual AUs and AU combinations with different intensities; **(3)** activate AUs bilaterally and unilaterally. Users can activate 13 AUs unilaterally (namely AU2, AU13, AU14, AU15, AU38, AU39, AU46, AU61, AU62, AU63, AU64, AU65, AU66); **(4)** generate EmFACS emotions with different intensities; **(5)** generate reproducible, realistic, 3D, static and dynamic (video) outputs; **(6)** generate Haptek hyper-texts provided by a C# API to enable reproduction of the HapFACS facial expressions in other applications with embedded Haptek avatars³.

For image generation, when a HapFACS user activates an AU with a specific intensity, its corresponding set of registers is activated with the selected intensities. In addition, users can select to activate the AU unilaterally for 19 AUs. When the AUs are activated, users can take and save a photo of the generated face. A snapshot of the HapFACS software interface is shown in Figure 3.1.

For video generation, users need to provide: the AU, side (i.e., left/right or bilateral), starting intensity, ending intensity, starting time, and ending time of the AU activation. HapFACS changes the intensity linearly from the start intensity to the end intensity and generates a video of the resulting expression (non-linear activation

³Features, such as modifiable background, lighting, and skin texture, are provided by the Haptek API, and HapFACS user-friendly interface enables non-experts to utilize all these functionalities without having to learn Haptek C++ or JavaScript APIs.

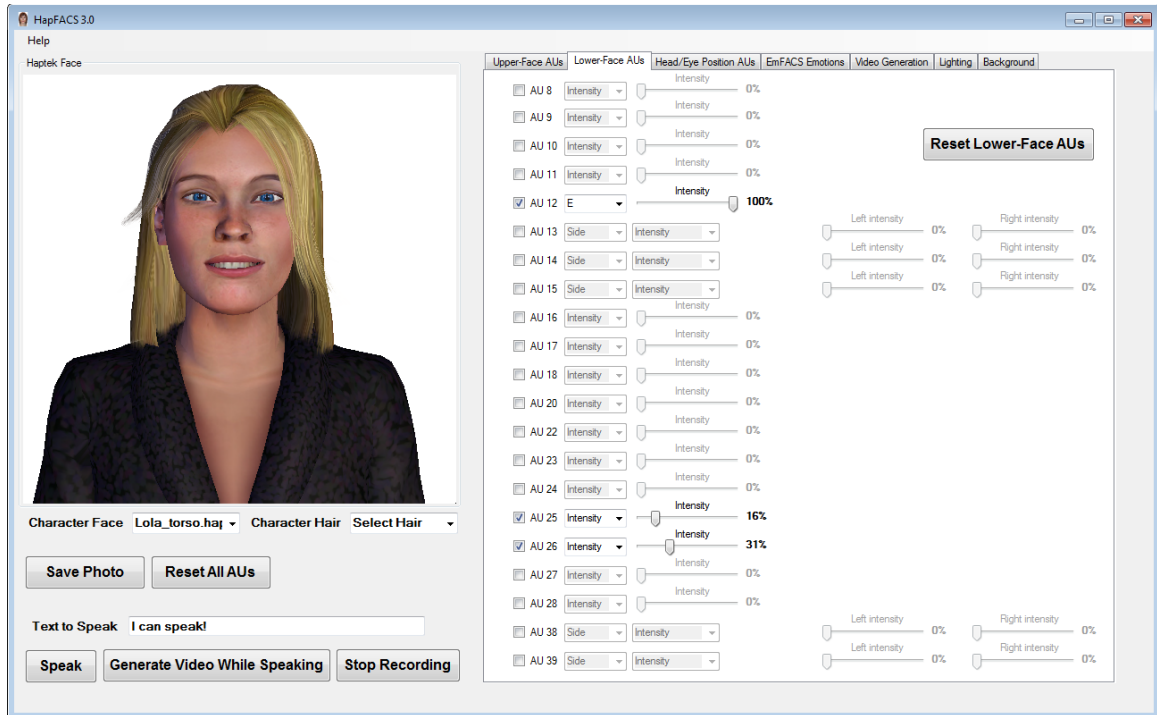


Figure 3.1: HapFACS interface snapshot.

is planned for a future version). Users can activate different AUs in parallel during the video generation and select overlapping activation times for the AUs.

In addition, users can generate EmFACS-based emotional facial expressions with different intensities for 9 emotions: happiness, sadness, surprise, anger, fear, disgust, contempt, embarrassment, and pride [FE83].

The users can also change the avatar and its hair style from the provided character and hair styles, or import any other Hapttek characters and hair styles by simply copying them into the corresponding folders in the HapFACS code. HapFACS can import characters of different ages, ethnicities, and genders generated in the PeoplePutty software. I have provided introduction and tutorial videos^{4,5} for the HapFACS, in which I show all the functionalities of the HapFACS.

⁴<http://youtu.be/aep5PRk2r6U>

⁵<http://youtu.be/0y13Xo9-uaI>

3.1.2 Registers, Switches and Action Units

For simulating the AUs on the characters' faces, I match each AU introduced in the FACS manual [EFH02] to a combination of the Haptek registers and switches. If we think of a Haptek character as an object in an object-oriented programming language, registers would be the character's (i.e., object's) properties, which can be adjusted to change the facial appearance of the character, and switches would be the methods (or functions) that simulate some gestures on the character (e.g., head nod).

Haptek provides a total of 71 registers for head characters (152 and 136 registers for torso and full-body characters, respectively). However expertise is needed to learn how registers work⁶. Haptek original registers and switches are not based on FACS, therefore, to generate facial expressions based on the FACS, I followed five steps:

1. I explored all the Haptek facial, head, and eye registers and switches, which manipulate the facial, head, and eye movements and gestures. Haptek registers/switches used in my implementation include 6 for head movements, 4 for eye movements, 8 for upper-face movements, and 21 for lower-face. A complete list of the Haptek registers is available in Table 3.1. Also, Figure 3.2 shows different Haptek registers used in HapFACS on a character's face.

⁶For users interested in registers involved for each AU, HapFACS source code and documentation are freely available upon request at <http://ascl.cis.fiu.edu>.

Table 3.1: Haptek registers for a full-body character.

Body Parts Affected	Haptek Register	Function	Body Parts Affected	Haptek Register	Function
Head	HeadSideBend HeadForward HeadTwist	Bend head left/right Move head forward/backward Twist head left/right		MidBrowUD RBrowUD LBrowUD REyeBallUD REyeBallLR LEyeBallUD LEyeBallLR L.lidUp R.lidup L.lidL R.lidL eyesdia blink winkL winkR trust distrust eyes_sad eyes_mad	Move mid-brow up/down Move right brow up/down Move left brow up/down Move right eye ball up/down Move right eye ball left/right Move left eye ball up/down Move left eye ball left/right Move left eye lid up/down Move right eye lid up/down Move left eye lid left/right Move right eye lid left/right Change eye pupil size Close/open eyes Close/open left eye Close/open right eye Close/open eye lids Narrow eyes Express sadness with eyes & brows Express anger with eyes & brows
Hand	RWristFlop RWristWave LWristFlop LWristWave RFingerThumbOut RFingerThumb RFingerThumbMid RFingerThumbTip RFingerIndexOut RFingerIndex RFingerIndexMid RFingerIndexTip RFingerMiddleOut RFingerMiddle RFingerMiddleMid RFingerMiddleTip RFingerRingOut RFingerRing RFingerRingMid RFingerRingTip RFingerPinkyOut RFingerPinky RFingerPinkyMid RFingerPinkyTip LFingerThumbOut LFingerThumb LFingerThumbMid LFingerThumbTip LFingerIndexOut LFingerIndex LFingerIndexMid LFingerIndexTip LFingerMiddleOut LFingerMiddle LFingerMiddleMid LFingerMiddleTip LFingerRingOut LFingerRing LFingerRingMid LFingerRingTip LFingerPinkyOut LFingerPinky LFingerPinkyMid LFingerPinkyTip	Flop right wrist Wave right wrist Flop left wrist Wave left wrist Move right thumb out Move right thumb Move right thumb from middle Move right thumb from tip Move right index finger out Move right index finger in Move right index finger in from middle Move right index finger in from tip Move right middle finger out Move right middle finger in Move right middle finger in from middle Move right middle finger in from tip Move right ring finger out Move right ring finger in Move right ring finger in from middle Move right ring finger in from tip Move right pinky finger out Move right pinky finger in Move right pinky finger in from middle Move right pinky finger in from tip Move left thumb out Move left thumb in Move left thumb in from middle Move left thumb in from tip Move left index finger out Move left index finger in Move left index finger in from middle Move left index finger in from tip Move left middle finger out Move left middle finger in Move left middle finger in from middle Move left middle finger in from tip Move left ring finger out Move left ring finger in Move left ring finger in from middle Move left ring finger in from tip Move left pinky finger out Move left pinky finger in Move left pinky finger in from middle Move left pinky finger in from tip	Upper face	TorsoSideBend TorsoBow TorsoTwist LumbarSideBend LumbarTwist RClavicalForward RClavicalUp LClavicalForward LClavicalUp RShoForward RShoOut RShoTwist LShoForward LShoOut LShoTwist RElbowBendJoint RElbowTwist LElbowBendJoint LElbowTwist	Bend torso left/right Bow torso Twist torso left/right Bend lumbar left/right Twist lumbar left/right Move right arm clavical forward/backward Move right arm clavical up/down Move left arm clavical forward/backward Move left arm clavical up/down Move right arm forward/backward Move right arm right/left Twist right arm left and right (out and in) Move left arm forward/backward Move left arm right/left Twist left arm left/right Bend right elbow up/down Twist right elbow Bend left elbow up/down Twist left elbow
			Neck	NeckSideBend NeckForward NeckTwist	Bend neck left/right Move neck forward/backward Twist neck left/right
			Morph	her_e	Change morph to male/female
				NostrilR3tx NostrilL3tx NostrilR3ty NostrilL3ty CheekR2sy CheekL2sy smile3 smile4 smirk smirkL smile_asymmetric sneer underbite sidepurse kiss lipcornerL3ty lipcornerR3ty mouth2ty aa ey uh b ch d iy ow ih th s g f eh uw	Move right nostril right/left Move left nostril right/left Move right nostril up/down Move left nostril up/down Move right cheek up/down Move left cheek up/down Smile with closed mouth Smile with opened mouth Smirk with right side of mouth Smirk with left side of mouth Extreme smile Sneer Underbite Sidepurse lips to left/right Kiss Move left lip corner up/down Move right lip corner up/down Move mouth up/down 'a' viseme 'ey' viseme 'uh' viseme 'b' viseme 'ch' viseme 'd' viseme 'iy' viseme 'ow' viseme 'ih' viseme 'th' viseme 's' viseme 'g' viseme 'f' viseme 'eh' viseme 'uw' viseme
Translate and rotate	LocalRotateX LocalRotateY LocalRotateZ LocalTranslateX LocalTranslateY LocalTranslateZ RotateZ RotateY RotateX	Rotate figure around X axis Rotate figure around Y axis Rotate figure around Z axis Move figure along X axis Move figure along Y axis Move figure along Z axis Rotate figure around X axis Rotate figure around Y axis Rotate figure around Z axis			
Leg and foot	RThighForward RThighOut RThighTwist LThighForward LThighOut LThighTwist RKneeBack RKneeTwist LKneeBack LKneeTwist RAnkleFlop RAnkleWave LAnkleFlop LAnkleWave RToeBend LToeBend	Move right leg forward Move right leg out Twist right leg from hip Move left leg forward Move left leg out Twist left leg from hip Bend right knee Twist right knee Bend left knee Twist left knee Flop right ankle Wave right ankle Flop left ankle Wave left ankle Bend right toe Bend left toe	Lower face and viseme		
emotions	sad_closed mad_open mad_closed	Express sadness with closed mouth Express anger with opened mouth Express anger with closed mouth			

2. For each AU, I found a subset of $\langle r, v \rangle$ tuples, where r is a register/switch whose activation in a combination simulates the same movements in the face as an actual AU activation; and v is the maximum intensity value of this specific register/switch. The HapFACS designer, who is a FACS-certified coder, found the maximum intensity of the registers experimentally based on the maximum possible activation of the AUs on a human’s face. For example, for AU4, the set of tuples used is:

$$\{\langle \text{MidBrowUD}, 2 \rangle, \langle \text{LBrowUD}, 0.6 \rangle, \langle \text{RBrowUD}, 0.6 \rangle, \langle \text{eyes_sad}, 1.25 \rangle\}.$$

3. The FACS manual [EFH02] introduces 6 intensity levels for each AU: (0) not active, (A) trace, (B) slight, (C) marked or pronounced, (D) severe or extreme, and (E) maximum. Assuming that E is activating the AU with 100% intensity, based on the FACS manual, I assigned 85% to D , 55% to C , 30% to B , 15% to A , and 0% to θ .
4. For each AU intensity level, I applied the same percentages to the intensity range of the Haptek registers. Although the intensity values in the FACS are represented as ranges, in HapFACS I represented them as discrete values based on our empirical approximations. For example, in AU4, the maximum value of the *MidBrowUD* register is 2.00, so its value for different intensities are:

$$A = 0.15 \times 2.0 = 0.3, B = 0.6, C = 1.1, D = 1.7, E = 2.$$

5. In addition to the discrete intensity levels (i.e., 0, A, B, C, D, and E), I enabled users to change AU intensities continuously from neutral to maximum intensity, by mapping the [0%, 100%] intensity range to the $[0, v]$ of each register r .

When I combine two or more AUs that share a common register, I accumulate their shared register intensities. However, I limit the intensity to the maximum limit, so

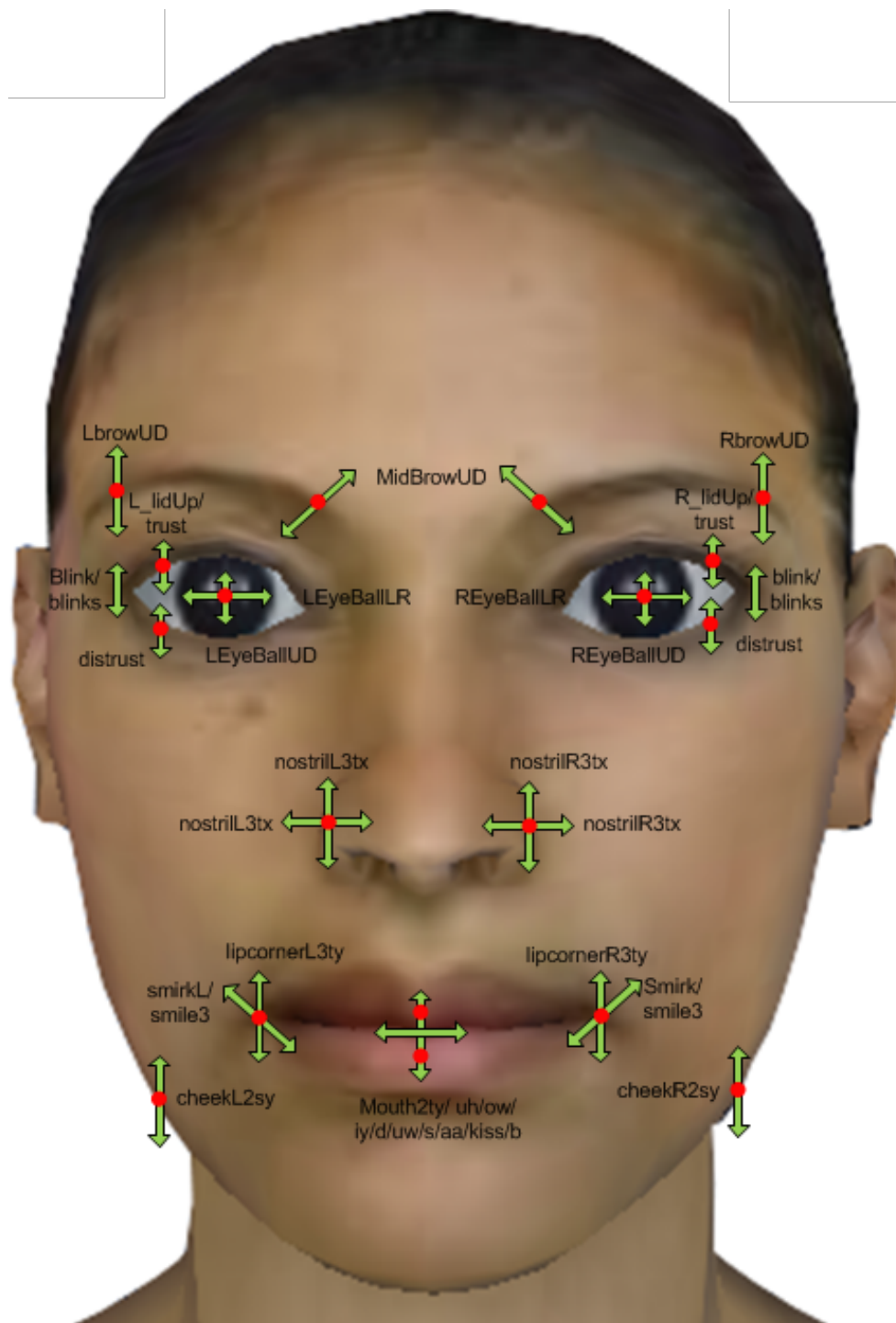


Figure 3.2: Hapttek facial and head registers.

that the summation of intensities does not deform the face beyond its physiologically possible appearance.

Table 3.2 shows the final mapping between the FACS action units and the Haptek facial variables (registers and switches) including the maximum value of the Haptek variable to show the highest FACS intensity (i.e., E intensity).

3.1.3 Integration with Other Software

As shown in Figure 3.3 (with circled numbers indicating the order of events), HapFACS users can re-produce expressions generated using HapFACS on any Haptek characters. When a facial expression animation is generated using HapFACS, a hypertext file with the *.hap* extension (the animation file type for Haptek characters including the relevant registers for the animation, its intensities and activation time) is exported automatically, as well as the video file.

Users only need to load the exported hypertext *.hap* file in their own Haptek character, who will then portray the same previously generated expression(s). Animation files can be loaded to the characters by dragging and dropping the files to the character, or passing the following hypertext to the Haptek character: `\load[file = [fileName.hap]]`⁷.

When using the HapFACS API for generating facial expressions and animations, the hypertext to re-produce the expression is returned as a string by the API. The hypertext can be passed to any Haptek character embedded in a software, in order to show the same expressions.

In order to embed HapFACS in another application and use the complete set of HapFACS functionalities, first, users need to embed a Haptek character into their

⁷I have created introductory and tutorial videos on HapFACS functionalities with animation demos, available at <http://ascl.cis.fiu.edu>.

Table 3.2: Mapping between FACS action units and Haptek facial variables.

AU	Haptek registers/switches	Max intensity	AU	Haptek registers/switches	Max intensity
1	MidBrowUD	-2.50	38	nostrilL3ty nostrilR3ty nostrilL3tx nostrilR3tx	0.30 0.30 0.60 -0.60
2	LBrowUD RBrowUD	-1.50 -1.50	39	nostrilL3ty nostrilR3ty nostrilL3tx nostrilR3tx	-0.30 -0.30 -0.60 0.60
4	MidBrowUD LBrowUD RBrowUD eyes_sad	2.00 0.60 0.60 1.25	41	trust	1.00
5	trust	-0.85	42	MidBrowUD	1.50
6	lipcornerL3ty lipcornerR3ty smile3 kiss	-0.70 -0.70 0.50 0.60	43	blink	1.40
7	distrust	1.20	44	MidBrowUD LBrowUD RBrowUD eyes_sad	1.00 -0.80 -0.90 1.40
8	b	0.75	45	blinks	CloseEye
9	nostrilL3ty nostrilR3ty nostrilL3tx nostrilR3tx MidBrowUD	0.60 0.60 0.10 -0.10 1.00	46	blinks	winkleleftfast/ winkrightfast
10	uh ow d iy nostrilL3ty nostrilR3ty nostrilL3tx nostrilR3tx	-1.10 0.61 1.40 -0.50 0.45 0.45 0.30 -0.30	51	HeadTwist	-1.00
11	nostrilL3tx nostrilR3tx uh d	0.20 -0.20 -0.40 0.31	52	HeadTwist	0.62
12	smile3	0.50	53	HeadForward	-0.65
13	lipcornerL3ty lipcornerR3ty uw	1.30 1.30 -0.25	54	HeadForward	0.60
14	smirk smirkL	0.30 0.40	5 ⁵	HeadSideBend	0.60
15	lipcornerL3ty lipcornerR3ty	-1.30 -1.30	56	HeadSideBend	-0.60
16	th	0.20	57	HeadForward NeckForward	-1.00 0.80
17	lipcornerL3ty lipcornerR3ty aa ow mouth2ty	-0.60 -0.60 -0.50 -0.40 0.20	58	HeadForward NeckForward	1.00 -0.80
18	kiss	1.30	M59	gestures	nod
20	lipcornerL3ty lipcornerR3ty b ow mouth2ty	-0.60 -0.60 0.50 -1.30 -0.40	M60	gestures	shake
22	ch	1.00	61	LEyeBallLR REyeBallLR	-0.45 -0.45
23	b kiss	1.00 0.65	62	LEyeBallLR REyeBallLR	0.45 0.45
24	b kiss	0.90 0.25	63	LEyeBallUD REyeBallUD	-0.45 -0.45
25	ey	0.80	64	LEyeBallUD REyeBallUD	0.40 0.40
26	aa	1.10	65	LEyeBallLR REyeBallLR	0.45 -0.45
27	aa ey	1.30 1.20	66	LEyeBallLR REyeBallLR	-0.45 0.45
28	b	1.30			

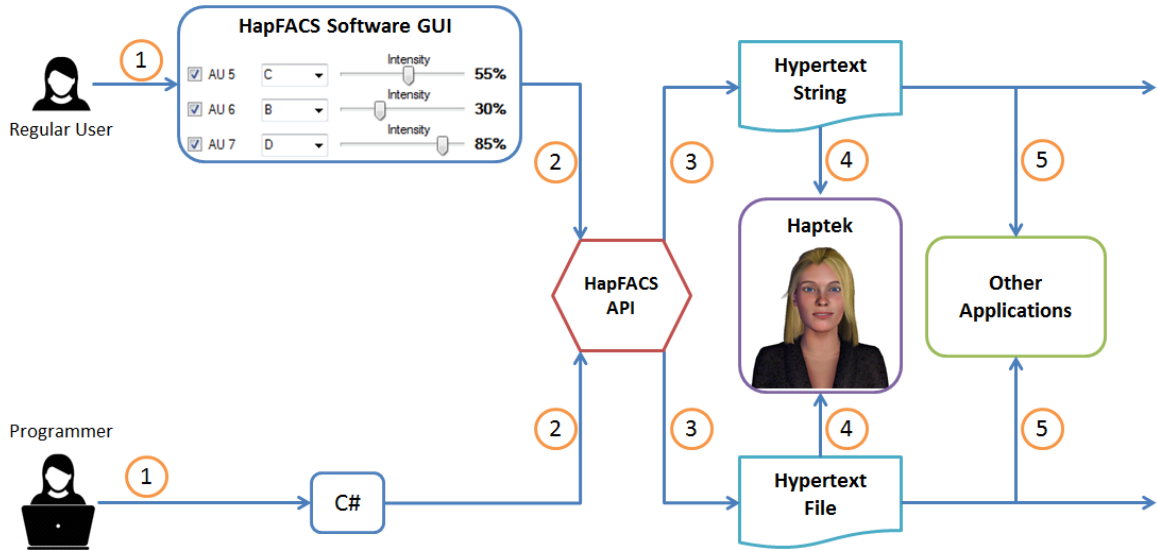


Figure 3.3: HapFACS work flow.

Windows Frame or HTML file, then they either: (1) include the Dynamic-Link Library (DLL) file provided for HapFACS API in their software, and call HapFACS methods; or (2) modify and include HapFACS C# code in their system.

3.1.4 HapFACS Validation

I performed 8 experiments to validate the HapFACS-generated AUs expressions, emotional facial expressions, and lip-synchronized speaking animations.

One set of experiments in Section 3.1.4 (experiments 1 and 2 described in Sections 3.1.4 and 3.1.4) were conducted with *FACS-certified coders*, in order to attempt to *objectively assess the validity of action unit activations* (singly or in combinations). FACS-certification is obtained by studying the manual, doing practice coding with video images, and taking a final test for the certification. Typically, this process takes 50 to 100 hours.

Another set of experiments (experiments 3-8 described in Sections 3.1.4 to 3.1.4) were conducted with lay participants to evaluate *subjective perceptions*. This set of

experiments evaluate participants' perception of the portrayed expressions in terms of the *9 EmFACS expressions (plus neutral)* [FE83], both statically from still images and dynamically from videos.

Since I am also interested in providing support to researchers working on speaking intelligent virtual agents, in Section 3.1.4, I evaluated the *characters' performance while they speak*. Experiment 7 (described in Section 3.1.4) focuses on evaluating the accuracy of the displayed expressions while the characters speak. Experiment 8 (described in Section 3.1.4) evaluates participants' perception of the quality of the characters' lip-synchronization while they speak.

In experiments 1 and 2, I will use accuracy, precision, recall, and F1-measure as the evaluation metrics. I define these terms as follow:

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (3.1)$$

$$Precision = \frac{t_p}{t_p + f_p} \quad (3.2)$$

$$Recall = \frac{t_p}{t_p + f_n} \quad (3.3)$$

$$F1-Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.4)$$

where t_p , f_p , t_n , and f_n are defined as follow:

- true positive (t_p): number of AUs correctly coded as present in a given combination.
- false positive (f_p): number of AUs incorrectly coded as present in a given combination.

- true negative (t_n): number of AUs correctly coded as absent in a given combination.
- false negative (f_n): number of AUs incorrectly coded as absent in a given combination.

In short, **accuracy** is the ratio of the correct codings, including (1) the AUs that are correctly coded as present in the video and (2) the AUs correctly coded as absent in the video. **Precision** is the percentage of correctly recognized AUs in the video over the total number recognized AUs in that video. **Recall** is the ratio of correctly recognized AUs in the video over the total number of AUs actually present in that video. Finally, **F1-Measure** is the harmonic mean of precision and recall, and is a measure of the coding accuracy. These measures are float numbers between 0 to 1. The higher they are, the better is the performance.

FACS-Certified Coders Evaluations

Participants: To perform this category of experiments, I asked three FACS-certified coders (36 years old White female; 26 years old Hispanic female; and 31 years old White male) to rate our individual and combinations of AUs.

Stimuli and Design: For each of the 49 individual AUs, and each of the 54 AU combinations (most common combinations indicated in the FACS manual) one 3-second video was generated. Each of the videos were performed by one of the 8 created character models (2 white males, 2 black males, 2 white females, and 2 black females) shown in Figure 3.4.

Almost equal number of videos were created using each model. Each video displayed a linear change of the AU intensity starting from 0% (i.e., 0 intensity in the FACS manual) to 100% (i.e., E intensity in the FACS manual) in 1 second (i.e., onset duration = 1000 ms), constant at peak for 1 second (i.e., apex duration = 1000 ms),

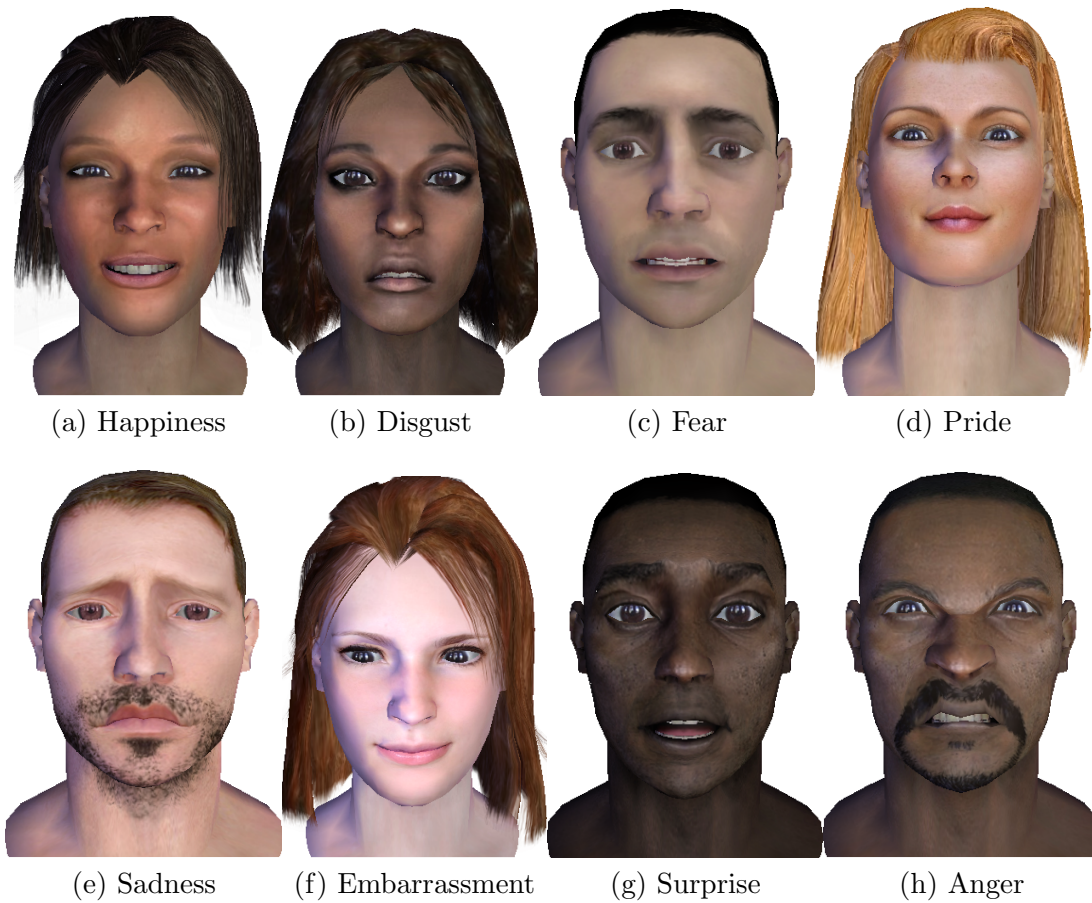


Figure 3.4: Models used for evaluation.

and lowered from peak to 0% intensity in 1 second. All videos have a frame rate of 30 frames/sec and a constant size of 485×480 pixels.

Videos of the individual AUs were also accompanied by three still 485×480 pixel images of the AU in 3 different intensities (30%, 60%, 90%) performed by the same model as in the video to help the coders see the changes in the face in different intensities.

Videos of AU combinations were accompanied by one still 485×480 pixel image of the same AU combination in 70% intensity performed by the same model as in the video to help the coders see the activated combination in a single frame.

Procedure: Videos and images were both named randomly to prevent subjects' tagging bias. Videos were hosted on YouTube⁸, and images were hosted on the survey website. In Experiment 1, I asked the FACS-certified coders to identify which AU was activated in each video (each accompanied with 3 images). In Experiment 2, FACS-certified coders were asked to identify all active AUs in each combination video. The coders were not asked to rate the intensity, because it is shown that judgments over the intensity show poor inter-rater agreement [SCW01].

Experiment 1: Validation of Individual AUs In this experiment, I examined the accuracy of the AUs generated by HapFACS in terms of AUs described in the FACS manual [EFH02].

Results and Discussion: As shown in Table 3.3, 41 out of the 49 simulated AUs were recognized by all the three coders with 100% recognition rate. The other 8 AUs, namely AU11 (Nasolabial Deepener), AU13 (Sharp Lip Puller), AU14 (Dimpler), AU16 (Lower Lip Depressor), AU20 (Lip Stretcher), AU41 (Glabella Lowerer), AU42 (Inner Eyebrow Lowerer), and AU44 (Eyebrow Gatherer), were recognized correctly

⁸<http://www.youtube.com>

by only two of the coders (i.e., 66.67% precision). The main reason for not being recognized perfectly is that these AUs are very similar to other AUs (listed in Table 3.4) and are mistakenly identified as other AUs. The **average recognition rate of all 49 AUs was 94.6%**.

Table 3.3: Individual AU recognition rates in Experiment 1.

Action Unit	Accuracy	Precision	Recall	F1-Measure
11, 13, 14, 16, 20, 41, 42, 44	0.99	0.67	0.67	0.67
Other AUs	1.00	1.00	1.00	1.00

The Cronbach α values (intra-class correlation) calculated in Table 3.4 show that, for all the AUs, intra-class correlation of the recognized AUs are greater than 0.75, which means that the ratings are statistically reliable. For AUs 4, 12, 24, 25, 38, and 43 the α score is less than one, because they are mistakenly recognized.

Table 3.4: Inter-rater correlation (Cronbach α) for individual AU recognitions.

AU	α	AUs Mistaken with	AU	α	AUs Mistaken with
4	0.922	-	25	0.750	-
11	0.750	38	38	0.899	-
12	0.922	-	41	0.750	43
13	0.750	14	42	0.750	4
14	0.750	12	43	0.899	-
16	0.750	25	44	0.750	4
20	0.750	24	All Other AUs	1.0	-
24	0.899	-			

Experiment 2: Validation of AU Combinations This experiment was designed to examine the accuracy of AU combinations (documented in FACS as the most common ones) generated by HapFACS.

Results and Discussion: Table 3.5 shows the accuracy, precision, recall, and F1-measure, defined in Equations 3.1 to 3.4, of each individual AU recognition in all the AU combinations. Individual AUs used in the combinations are recognized with

an average accuracy of **0.96**, average precision of **0.81**, average recall of **0.83**, and average F1-measure of **0.81**.

Table 3.5: Individual AU recognition results in AU combinations of Experiment 2.

Action Unit	Accuracy	Precision	Recall	F1-Measure
1 (Inner Brow Raiser)	0.99	1.00	0.95	0.98
2 (Outer Brow Raiser)	0.98	1.00	0.8	0.89
4 (Brow Lowerer)	0.96	0.80	0.8	0.8
5 (Upper Lid Raiser)	0.98	1.00	0.78	0.88
6 (Cheek Raiser)	0.90	0.53	0.94	0.68
7 (Lid Tightener)	0.91	0.52	0.93	0.67
9 (Nose Wrinkler)	0.98	0.71	0.83	0.77
10 (Upper Lip Raiser)	0.90	0.76	0.77	0.76
12 (Lip Corner Puller)	0.90	0.83	0.84	0.83
14 (Dimpler)	0.98	0.90	0.75	0.82
15 (Lip Corner Depressor)	0.96	0.88	0.85	0.87
16 (Lower Lip Depressor)	0.91	0.53	0.6	0.56
17 (Chin Raiser)	0.93	0.83	0.83	0.83
18 (Lip Pucker)	0.99	0.75	1.00	0.86
20 (Lip Stretcher)	0.97	1.00	0.67	0.80
22 (Lip Funneler)	0.99	0.67	1.00	0.80
23 (Lip Tightener)	0.88	0.79	0.61	0.69
24 (Lip Pressor)	0.95	0.56	0.56	0.56
25 (Lips Part)	0.96	0.92	0.96	0.94
26 (Jaw Drop)	0.93	0.62	0.53	0.57
27 (Mouth Stretch)	0.98	0.69	1.00	0.82
43 (Eyes Closed)	0.99	1.00	0.67	0.8
53 (Head Up)	1.00	1.00	1.00	1.00
54 (Head Down)	1.00	1.00	1.00	1.00
62 (Eyes Turn Right)	1.00	1.00	1.00	1.00
64 (Eyes Down)	0.98	1.00	0.33	0.50

Table 3.6 shows the recognition accuracy, precision, recall, and F1-measure for different combinations. The generated AU combinations are recognized with an average accuracy of **0.98**, average precision of **0.80**, average recall of **0.81**, and average F1-measure of **0.80**.

Also, the Cronbach α was calculated to evaluate the intra-class correlation in recognizing the AU combinations. The α value was equal to 0.779, which shows a high correlation between the coders and reliability of the ratings.

Table 3.6: AU combination recognition rates. AUs in parentheses show false-positive recognitions.

HapFACS-Generated AU Combination	Recognized AUs by FACS-Certified Coders	Accuracy	Precision	Recall	F1-Measure
1 + 2	1 + 2	1.00	1.00	1.00	1.00
1 + 4	1 + 4	1.00	1.00	1.00	1.00
1 + 2 + 4	1 + 2 + 4	0.99	1.00	0.78	0.88
1 + 2 + 5	1 + 2 + 5	1.00	1.00	1.00	1.00
4 + 5	4 + 5 + (41 + 42)	0.98	0.71	0.83	0.77
5 + 7	5 + 7	0.99	1.00	0.83	0.91
6 + 43	6 + 43	0.97	0.63	0.83	0.71
6 + 7 + 12	6 + 7 + 12	0.99	1.00	0.89	0.94
6 + 12 + 15	6 + 12 + 15 + (13 + 14)	0.9	0.78	0.78	0.78
6 + 12 + 15 + 17	6 + 12 + 15 + 17 + (7 + 10)	0.96	0.75	0.75	0.75
6 + 12 + 17 + 23	6 + 12 + 17 + 23 + (7 + 10)	0.97	0.82	0.75	0.78
7 + 12	7 + 12 + (6)	0.99	0.86	1.00	0.92
7 + 43	7 + 43	0.99	1.00	0.83	0.91
9 + 17	9 + 17 + (10 + 13 + 24)	0.98	0.67	1.00	0.80
9 + 16 + 25	9 + 16 + 25 + (22 + 41)	0.98	0.80	0.89	0.84
10 + 14	10 + 14 + (12 + 25)	0.97	0.63	0.83	0.71
10 + 15	10 + 15 + (25)	0.99	0.86	1.00	0.92
10 + 17	10 + 17 + (11 + 24 + 25)	0.97	0.63	0.83	0.71
10 + 12 + 25	10 + 12 + 25 + (6 + 7 + 9)	0.97	0.70	0.78	0.74
10 + 15 + 17	10 + 15 + 17 + (25 + 38)	0.97	0.78	0.78	0.78
10 + 16 + 25	10 + 16 + 25 + (26)	0.97	0.78	0.78	0.78
10 + 17 + 23	10 + 17 + 23 + (9 + 15 + 24)	0.97	0.70	0.78	0.74
10 + 20 + 25	10 + 20 + 25 + (16)	0.98	0.88	0.78	0.82
10 + 23 + 25	10 + 23 + 25 + (11 + 16 + 26)	0.97	0.70	0.78	0.74
10 + 12 + 16 + 25	10 + 12 + 16 + 25 + (6 + 7 + 26)	0.95	0.69	0.75	0.72
12 + 15	12 + 15 + (6 + 17 + 23)	0.97	0.56	0.83	0.67
12 + 17	12 + 17 + (6 + 7 + 23)	0.97	0.63	0.83	0.71
12 + 23	12 + 23 + (6 + 13)	0.98	0.71	0.83	0.77
12 + 24	12 + 24 + (6 + 23)	0.97	0.56	0.83	0.67
12 + 25 + 26	12 + 25 + 26 + (6 + 7 + 10 + 27)	0.95	0.58	0.78	0.67
12 + 25 + 27	12 + 25 + 27 + (7 + 10 + 16)	0.96	0.64	0.78	0.70
12 + 15 + 17	12 + 15 + 17 + (6 + 7)	0.97	0.78	0.78	0.78
12 + 16 + 25	12 + 16 + 25 + (6)	0.98	0.80	0.89	0.84
12 + 17 + 23	12 + 17 + 23 + (7 + 28)	0.97	0.78	0.78	0.78
20 + 23 + 25	20 + 23 + 25 + (12 + 15)	0.97	0.75	0.67	0.71
22 + 23 + 25	22 + 23 + 25 + (16 + 38)	0.97	0.78	0.78	0.78
23 + 25 + 26	23 + 25 + 26	0.98	1.00	0.67	0.80
14 + 17	14 + 17 + (6 + 7 + 12 + 23)	0.97	0.60	1.00	0.75
14 + 23	14 + 23 (12 + 18)	0.97	0.67	0.67	0.67
15 + 17	15 + 17	0.99	1.00	0.83	0.91
15 + 23	15 + 23 + (17)	0.99	0.83	0.83	0.83
17 + 23	17 + 23 + (24)	0.98	0.80	0.67	0.73
17 + 24	17 + 24 + (10)	0.98	0.80	0.67	0.73
18 + 23	18 + 23 + (26)	0.98	0.80	0.67	0.73
20 + 25 + 26	20 + 25 + 26 + (10 + 16)	0.97	0.75	0.67	0.71
20 + 25 + 27	20 + 25 + 27 + (12 + 16)	0.98	0.80	0.89	0.84
4 + 5 + 7 + 24	4 + 5 + 7 + 24 + (23 + 41 + 42)	0.96	0.75	0.75	0.75
10 + 16 + 25 + 26	10 + 16 + 25 + 26 + (15 + 27)	0.95	0.73	0.67	0.70
14 + 54 + 62 + 64	14 + 54 + 62 + 64 + (12)	0.97	0.89	0.67	0.76
1 + 2 + 4 + 5 + 20 + 25 + 26	1 + 2 + 4 + 5 + 20 + 25 + 26 + (13 + 16)	0.94	0.83	0.71	0.77
6 + 12	6 + 12	1.00	1.00	1.00	1.00
12 + 53 + 64	12 + 53 + 64	0.99	1.00	0.78	0.88
1 + 4 + 15	1 + 4 + 15 + (44)	0.99	0.89	0.89	0.89
1 + 2 + 5 + 25 + 27	1 + 2 + 5 + 25 + 27	0.99	1.00	0.93	0.97

Validation of Emotion Inferences from EmFACS Stimuli

Stimuli and Design: For each emotion, two videos were generated, one with 50% intensity and one with 100 % intensity, hence a total of 20 videos were used for this experiment. Lengthwise, videos were 3 seconds and their size was 485×485 pixels. The videos showed the activation of the emotion from 0% to 100%-intensity (or 50% for lower intensity version) in one second (i.e., onset duration = 1000 ms), constant at peak intensity for one second (i.e., apex duration = 1000 ms), and decreasing from peak to 0% intensity in one second. The same 8 models of Experiment 1 were used to animate the emotions. For Experiment 4, I converted the videos to gray-scaled to study the effects of color in recognition rates. Also, for experiment 5, I created 10 still images (one for each emotion) of size 485×485 pixels, using the same models at each of the two intensity levels (i.e., 100% and 50%). For Experiment 6, I converted the images to gray-scale.

The 10 used emotions were *neutral*, *happiness*, *sadness*, *surprise*, *anger*, *fear*, *disgust*, *contempt*, *embarrassment*, and *pride*, with the AUs shown in Table 3.7.

Table 3.7: AUs involved in emotional expressions.

Emotion	Action Units	Emotion	Action Units
<i>Happiness</i>	6, 12, 25	<i>Disgust</i>	9, 15, 16
<i>Sadness</i>	1, 4, 15	<i>Contempt</i>	12, 14R
<i>Surprise</i>	1, 2, 5, 26	<i>Embarrassment</i>	12, 52, 62, 64
<i>Anger</i>	5, 7, 9, 10, 15, 17, 42	<i>Pride</i>	12, 53, 58, 64
<i>Fear</i>	1, 2, 4, 5, 20, 26	<i>Neutral</i>	0

Procedure: In Experiments 3, 4, 5, and 6, subjects were asked to choose which emotion was portrayed by each video, as well as to rate how believable each emotion was being portrayed by the character on a 5-level Lickert scale (0: not believable at all, 5: very believable). Participants could select the same emotion for multiple

videos, or select ‘None’ if they believed some other emotion (not included in the above emotions) was portrayed in a video.

Experiment 3: Colored Video Stimuli In order to see how well HapFACS characters portray EmFACS [FE83] standard emotions, the AUs that are involved in creating these emotions were manipulated in two different intensities of 50% and 100%.

Participants: I recruited 66 students on Florida International University (FIU) campus as well as 80 participants on Amazon Mechanical Turk (AMT). Subjects were compensated with \$5 vendor gift cards. I randomly assigned 82 subjects to the 100% intensity experiment and 64 subjects to the 50% intensity experiment. Table 3.8 shows the demographic information of the subjects.

Table 3.8: Subjects’ demographic data in each group of Experiment 3.

Experiment Group	Female (Avg. age)	Male (Avg. age)	White	Black	Asian	Hispanic
100% Intensity	42.7% (31.3)	57.3% (28.3)	57.3%	7.3%	11%	24.4%
50% Intensity	45.3% (32.1)	54.7% (28.1)	54.7%	9.4%	14.1%	21.9%

Results and Discussion: Tables 3.9 and 3.10 show the recognition rates of the emotions in 100% and 50%-intensity versions respectively. The average recognition rate for 100% and 50%-intensity videos were **83.8%** and **72.3%**, respectively. Therefore, dynamic emotional facial expressions (i.e., videos) of both extreme and subtle intensities were perceived correctly from the HapFACS animations. Figures 3.5 and 3.6 depict these results in diagrams.

The believability (i.e., how natural and believable is the character when expressing the emotion) of the characters was reported as 3.8 (standard deviation = 0.97) for the 100%-intensity videos and 3.84 (standard deviation = 0.92) for 50%-intensity videos. Table 3.11 shows the believability for each individual video.

Table 3.9: Emotion recognition percentages of the 100%-intensity colored videos.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Videos	<i>Anger</i>	79.3	2.4	6.1	0.0	1.2	1.2	0.0	7.3	0.0	1.2	1.2
	<i>Contempt</i>	0	68.3	0.0	3.7	0.0	18.3	2.4	0.0	1.2	2.4	3.7
	<i>Disgust</i>	7.3	2.4	75.6	0.0	7.3	0.0	0.0	4.9	0.0	1.2	1.2
	<i>Embarrass</i>	0.0	6.1	0.0	75.6	0.0	3.7	1.2	1.2	0.0	3.7	8.5
	<i>Fear</i>	0.0	1.2	2.4	6.1	86.6	0.0	0.0	1.2	2.4	0.0	0.0
	<i>Happiness</i>	0.0	1.2	0.0	4.9	0.0	91.5	1.2	0.0	1.2	0.0	0.0
	<i>Pride</i>	1.2	9.8	1.2	1.2	1.2	2.4	74.4	0.0	2.4	1.2	4.9
	<i>Sadness</i>	0.0	1.2	3.7	1.2	0.0	0.0	0.0	87.8	1.2	0.0	4.9
	<i>Surprise</i>	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	98.8	0.0	0.0
	<i>Neutral</i>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0

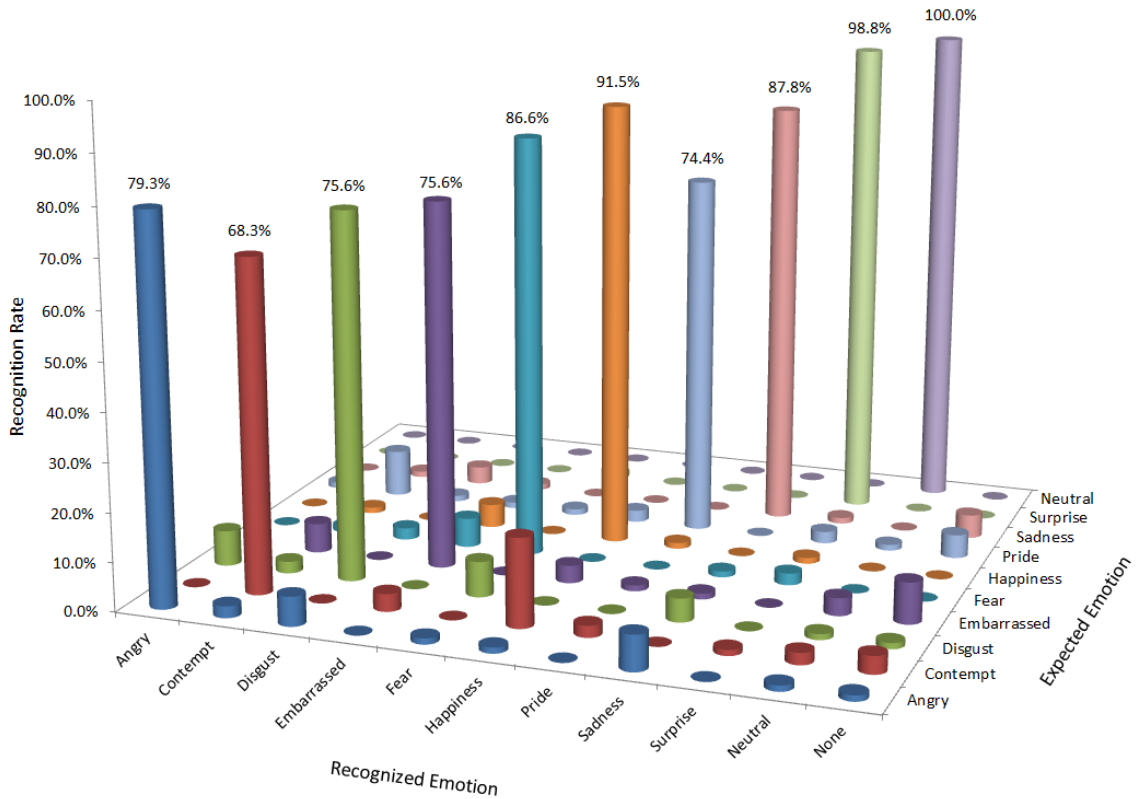


Figure 3.5: Emotion recognition percentages of the 100%-intensity colored videos.

Table 3.10: Emotion recognition ratings of the 50%-intensity colored videos.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Videos	<i>Anger</i>	54.7	0.0	6.3	0.0	6.3	0.0	3.1	26.6	0.0	1.6	1.6
	<i>Contempt</i>	0.0	65.6	0.0	0.0	1.6	6.3	6.3	0.0	0.0	7.8	12.5
	<i>Disgust</i>	10.9	4.7	62.5	3.1	0.0	0.0	0.0	7.8	0.0	4.7	6.3
	<i>Embarrass</i>	1.6	4.7	0.0	67.2	0.0	4.7	10.9	0.0	0.0	3.1	7.8
	<i>Fear</i>	1.6	0.0	9.4	3.1	76.6	1.6	0.0	1.6	1.6	0.0	4.7
	<i>Happiness</i>	0.0	4.7	0.0	0.0	0.0	84.4	4.7	0.0	1.6	1.6	3.1
	<i>Pride</i>	1.6	7.8	3.1	1.6	4.7	4.7	56.3	0.0	6.3	4.7	9.4
	<i>Sadness</i>	0.0	4.7	0.0	4.7	3.1	1.6	1.6	78.1	3.1	1.6	1.6
	<i>Surprise</i>	0.0	0.0	0.0	1.6	0.0	1.6	1.6	1.6	92.2	1.6	0.0
	<i>Neutral</i>	1.6	3.1	0.0	0.0	0.0	3.1	0.0	3.1	0.0	85.9	3.1

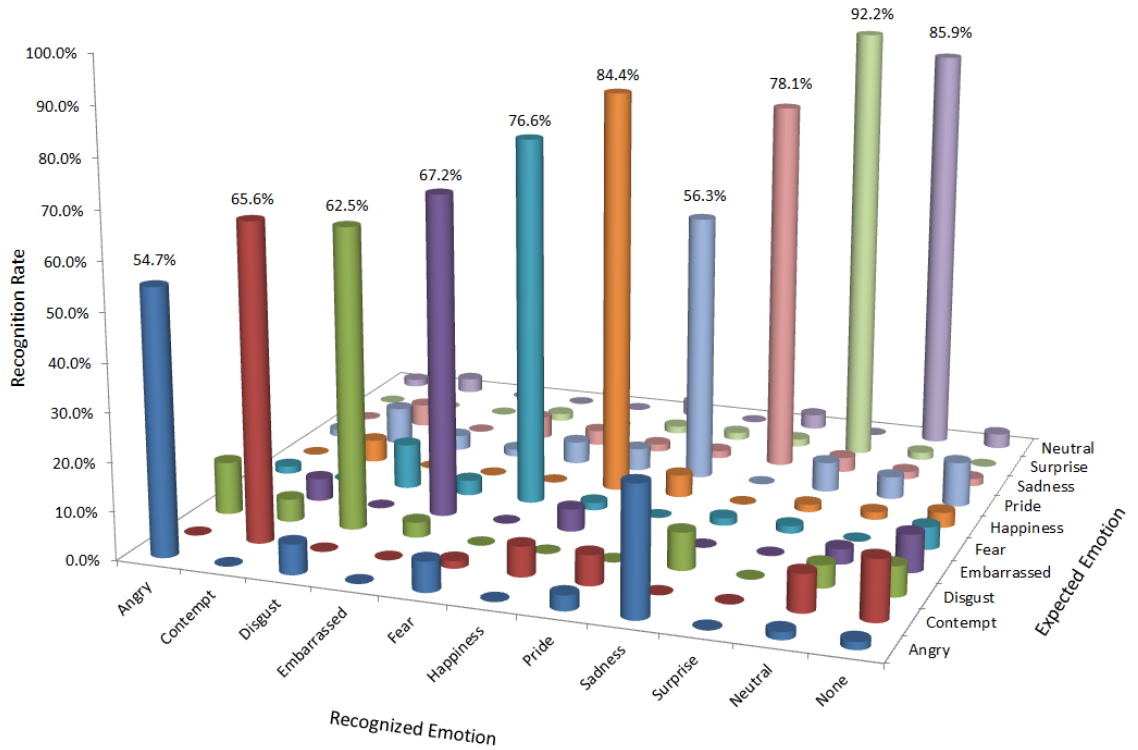


Figure 3.6: Emotion recognition percentages of the 50%-intensity colored videos.

In order to find out the effects of intensity on emotion recognition and believability, I performed two ANOVA analyses, for 100% and 50% intensity videos. Each ANOVA analysis was a 10×2 (i.e., Emotion \times Intensity) analysis ($df = 1$). Results revealed no significant effect of intensity on emotion recognitions ($p > 0.05$) or believability ($p > 0.05$).

Cronbach α values (intra-class correlation) were computed for each of the expressions, using participants' ratings as columns (items) and the 10 videos as rows (cases). Results shown in table 3.11 indicate that for all expressions $\alpha > 0.7$, which means the faces were rated *reliably*.

Table 3.11: Believability (sd = std. deviation) and Cronbach α for Experiment 3.

Measure Emotion	100%-Intensity Group		50%-Intensity Group	
	Believability (sd)	Cronbach α	Believability (sd)	Cronbach α
Anger	3.9 (0.97)	0.992	3.84 (0.86)	0.966
Contempt	3.5 (1.00)	0.989	4.22 (0.7)	0.982
Disgust	3.83 (0.93)	0.990	3.38 (1.05)	0.975
Embarrassment	3.72 (0.93)	0.992	3.88 (0.77)	0.982
Fear	3.73 (0.9)	0.995	3.78 (0.79)	0.988
Happiness	3.1 (1.10)	0.997	3.52 (1.11)	0.994
Pride	3.7 (0.86)	0.990	3.66 (0.98)	0.965
Sadness	3.7 (0.93)	0.997	3.88 (0.98)	0.989
Surprise	4.13 (0.77)	1.00	4.27 (0.7)	0.997
Neutral	4.3 (0.82)	1.00	4.02 (0.77)	0.995

In order to make sure that there is no major difference between the annotations of the AMT subjects and the FIU student subjects, I performed two T-tests between these two populations for each individual emotion (one for the 50% intensity and one for the 100% intensity). The goal is to know if I can combine all the AMT and local users in our analyses. Having the null hypothesis as “two ratings are coming from the same population”, the T-test results show that with $\alpha = 0.05$, $df = 20$, for all emotions, the t value obtained is less than the critical value of 2.085. Therefore, I fail to reject the null hypothesis, or in other words, I cannot say that the two samples

are coming from two different populations. Based on the T-test results, in all the validation experiments, I combined the AMT and FIU subjects.

Experiment 4: Gray-Scaled Video Stimuli This experiment aims to determine if the *color of each video* brings any changes over the emotion recognition of the participants. Similar to Experiment 3, AUs involved in facial expressions of emotions were manipulated in two different intensities of 50% and 100%.

Participants: I recruited 80 AMT subjects to recognize the emotion portrayed in each video. Forty subjects performed the 100%-intensity experiment and 40 subjects performed the 50%-intensity experiment. Table 3.12 shows the demographic information of the subjects.

Table 3.12: Subjects’ demographic data in each group of Experiment 4.

Experiment Group)	Female (Avg. age)	Male (Avg. age)	White	Black	Asian	Hispanic
100% Intensity	60% (35.7)	40% (32.2)	77.5%	2.5%	10%	10%
50% Intensity	55% (32.7)	45% (30.8)	90%	2.5%	5%	2.5%

Results and Discussion: Tables 3.13 and 3.14 show the recognition rate of the emotions in 100% and 50%-intensity respectively. The average recognition rate for 100% and 50%-intensity videos were **80.3%** and **71.8%** respectively. Figures 3.7 and 3.8 depict these results in diagrams. Therefore, the colorfulness of the videos do not bias the subjects’ emotion recognition.

Believability of the characters in 100%-intensity videos was reported as 3.7 (sd = 0.96) and in 50%-intensity videos reported as 3.7 (sd = 0.91). Table 3.15 shows the believability for each individual video. In order to find out the effects of intensity on emotion recognition and believability, I performed two ANOVA analyses, for 100% and 50% intensity videos. Each ANOVA analysis was a 10×2 (i.e., Emotion \times Intensity) analysis ($df = 1$). Results revealed no significant effect of intensity on emotion recognitions ($p > 0.05$) or believability ($p > 0.05$).

Table 3.13: Emotion recognition ratings of the 100%-intensity gray-scaled videos.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Videos	<i>Anger</i>	72.5	5.0	5.0	2.5	0.0	2.5	0.0	7.5	2.5	0.0	2.5
	<i>Contempt</i>	0.0	65.0	0.0	0.0	0.0	17.5	15.0	0.0	2.5	0.0	0.0
	<i>Disgust</i>	2.5	2.5	72.5	2.5	2.5	2.5	2.5	2.5	2.5	5.0	2.5
	<i>Embarrass</i>	0.0	2.5	0.0	80	2.5	7.5	0.0	0.0	2.5	5.0	0.0
	<i>Fear</i>	2.5	0.0	2.5	5	82.5	0.0	0.0	0.0	7.5	0.0	0.0
	<i>Happiness</i>	0.0	0.0	0.0	2.5	0.0	95	0.0	0.0	2.5	0.0	0.0
	<i>Pride</i>	0.0	0.0	0.0	0.0	0.0	15	62.5	0.0	10	5	0.0
	<i>Sadness</i>	2.5	0.0	5.0	2.5	0.0	2.5	2.5	80	2.5	2.5	0.0
	<i>Surprise</i>	0.0	0.0	0.0	0.0	7.5	0.0	0.0	0.0	92.5	0.0	0.0
	<i>Neutral</i>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0

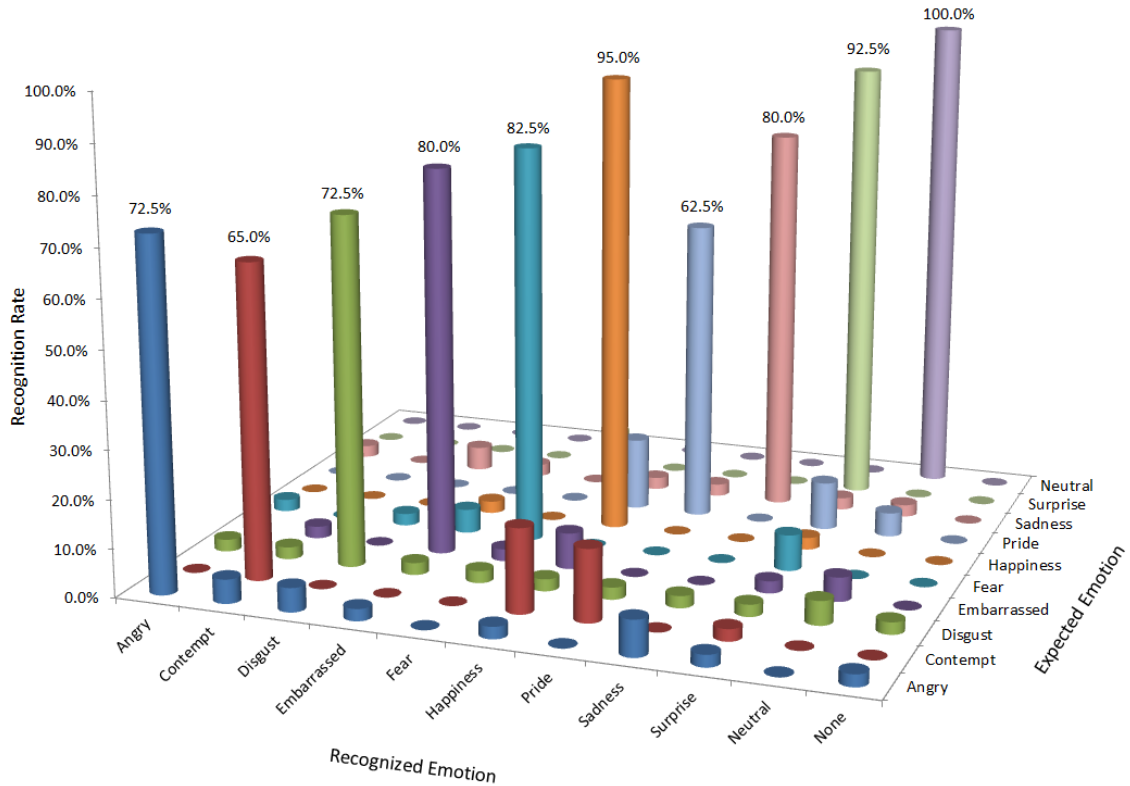


Figure 3.7: Emotion recognition percentages of the 100%-intensity gray-scaled videos.

Table 3.14: Emotion recognition ratings of the 50%-intensity gray-scaled videos.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Videos	<i>Anger</i>	60	0.0	0.0	0.0	10.0	2.5	7.5	7.5	0.0	7.5	5.0
	<i>Contempt</i>	0.0	57.5	0.0	0.0	0.0	10.0	10.0	0.0	10.0	5.0	7.5
	<i>Disgust</i>	2.5	0.0	67.5	0.0	2.5	0.0	2.5	2.5	0.0	5.0	17.5
	<i>Embarrass</i>	0.0	7.5	2.5	75.0	2.5	5.0	0.0	0.0	0.0	0.0	7.5
	<i>Fear</i>	7.5	0.0	0.0	0.0	72.5	0.0	2.5	0.0	7.5	5.0	5.0
	<i>Happiness</i>	0.0	7.5	0.0	2.5	0.0	85.0	0.0	0.0	0.0	2.5	2.5
	<i>Pride</i>	0.0	0.0	0.0	0.0	0.0	7.5	45.0	2.5	2.5	20.0	10.0
	<i>Sadness</i>	2.5	0.0	0.0	0.0	7.5	0.0	0.0	77.5	0.0	5.0	7.5
	<i>Surprise</i>	0.0	0.0	0.0	0.0	7.5	5.0	0.0	0.0	82.5	2.5	2.5
	<i>Neutral</i>	0.0	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	95.0	2.5

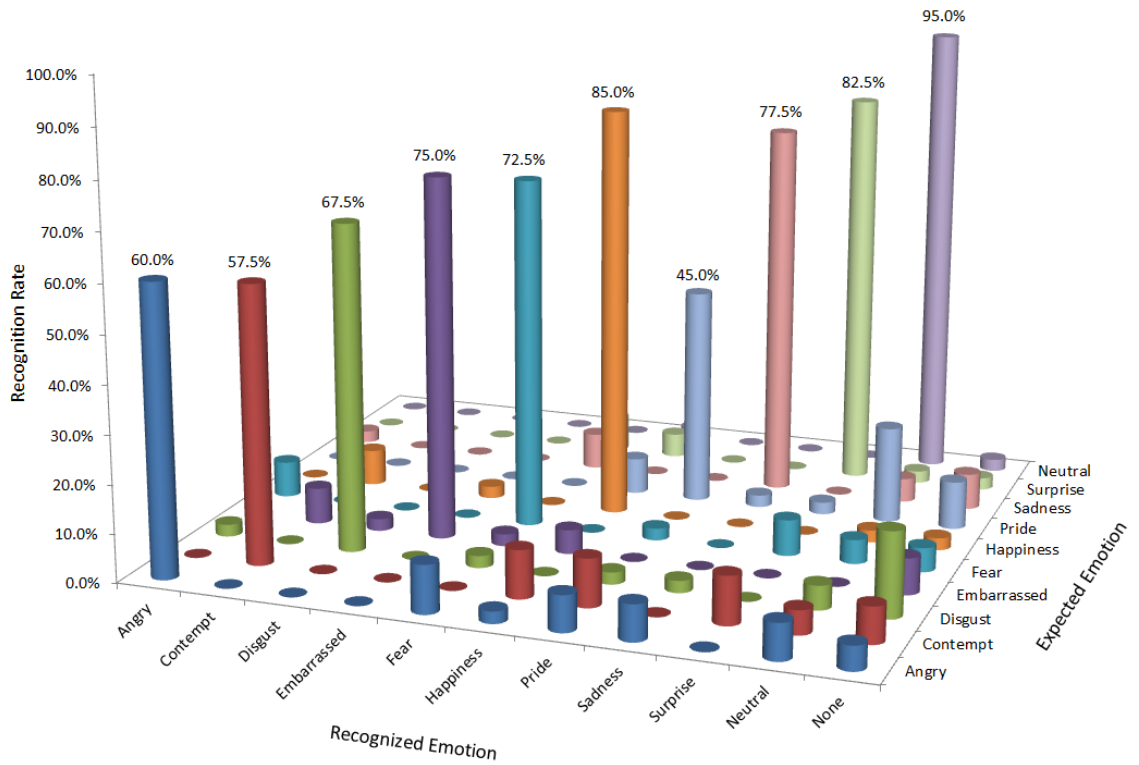


Figure 3.8: Emotion recognition percentages of the 50%-intensity gray-scaled videos.

Cronbach α values (intra-class correlation) were computed for each of the expressions, using participants' ratings as columns (items) and the 10 videos as rows (cases). Results shown in table 3.15 indicate that for all expressions $\alpha > 0.7$, which means the faces were rated *reliably*.

Table 3.15: Believability (sd = std. deviation) and Cronbach α for Experiment 4.

Measure Emotion	100%-Intensity Group		50%-Intensity Group	
	Believability (sd)	Cronbach α	Believability (sd)	Cronbach α
Anger	3.6 (0.93)	0.982	3.7 (0.79)	0.97
Contempt	3.9 (0.89)	0.978	3.8 (0.87)	0.967
Disgust	3.5 (0.96)	0.943	3.4 (0.87)	0.985
Embarrassment	3.4 (0.98)	0.986	3.8 (0.78)	0.989
Fear	3.6 (0.84)	0.989	3.7 (0.86)	0.972
Happiness	3.4 (1.22)	0.987	3.3 (1.1)	0.981
Pride	3.5 (0.82)	0.975	3.5 (1.0)	0.928
Sadness	3.8 (0.78)	0.988	3.6 (1.0)	0.986
Surprise	3.8 (1.1)	0.986	3.9 (0.76)	0.986
Neutral	4.2 (0.79)	0.995	4.0 (0.8)	0.984

Experiment 5: Colored Image Stimuli This experiment was designed to test how well emotions were portrayed by HapFACS in static colored images.

Participants: I recruited 66 FIU students, and 80 AMT workers for this experiment. I randomly assigned 70 subjects to the 100% intensity experiment and 76 subjects to the 50% intensity experiment. Table 3.16 shows the demographic information of the subjects.

Table 3.16: Subjects' demographic data in each group of Experiment 5.

Experiment Group	Female (Avg. age)	Male (Avg. age)	White	Black	Asian	Hispanic
100% Int.	42.9% (33.5)	57.1% (27.7)	52.9%	8.6%	15.7%	22.9%
50% Int.	43.4% (30.6)	56.6% (28)	61.8%	6.6%	10.5%	21.1%

Results and Discussion: Tables 3.17 and 3.18 show the recognition rate of the emotions in 100% and 50%-intensity respectively. The average recognition rate for 100% and 50%-intensity images were **82.1%** and **72.5%** respectively. Figures 3.9 and

3.10 depict these results in diagrams. Therefore, static emotional facial expressions of both extreme and subtle intensities were perceived correctly from the HapFACS images.

Table 3.17: Emotion recognition ratings of the 100%-intensity colored images.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Images	<i>Anger</i>	82.9	1.4	2.9	0.0	2.9	0.0	0.0	7.1	0.0	1.4	1.4
	<i>Contempt</i>	0.0	65.7	0.0	5.7	0.0	15.7	8.6	0.0	1.4	0.0	2.9
	<i>Disgust</i>	10.0	2.9	78.6	0.0	2.9	0.0	0.0	5.7	0.0	0.0	0.0
	<i>Embarrass</i>	0.0	7.1	0.0	77.1	0.0	5.7	0.0	0.0	0.0	2.9	7.1
	<i>Fear</i>	0.0	0.0	4.3	0.0	82.9	0.0	0.0	5.7	4.3	0.0	2.9
	<i>Happiness</i>	0.0	4.3	0.0	2.9	0.0	91.4	0.0	0.0	1.4	0.0	0.0
	<i>Pride</i>	0.0	5.7	0.0	0.0	1.4	2.9	70.0	0.0	2.9	7.1	10.0
	<i>Sadness</i>	1.4	0.0	2.9	0.0	4.3	0.0	0.0	88.6	0.0	1.4	1.4
	<i>Surprise</i>	0.0	0.0	0.0	1.4	4.3	1.4	0.0	0.0	91.4	1.4	0.0
	<i>Neutral</i>	0.0	1.4	0.0	0.0	1.4	0.0	1.4	0.0	0.0	92.9	2.9

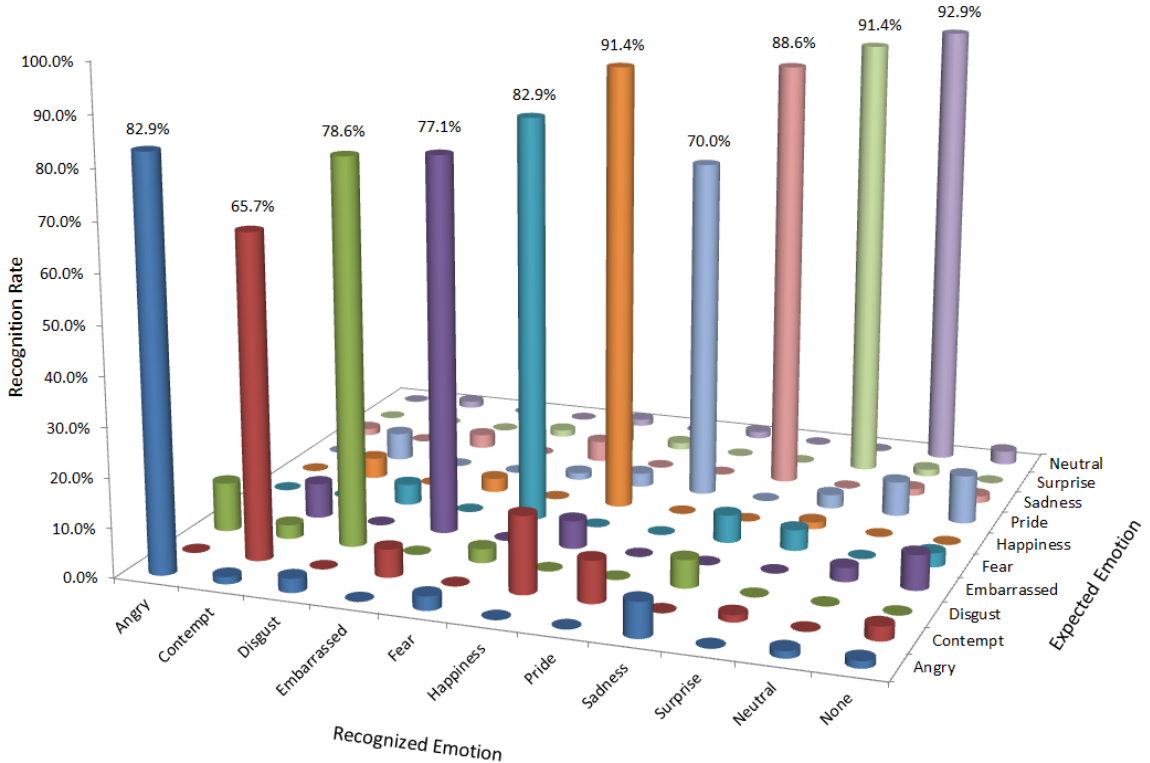


Figure 3.9: Emotion recognition percentages of the 100%-intensity colored images.

Table 3.18: Emotion recognition ratings of the 50%-intensity colored images.

		Recognized Emotions										
		Anger	Contempt	Disgust	Embarrass.	Fear	Happiness	Pride	Sadness	Surprise	Neutral	None
Images	Anger	68.4	1.3	11.8	0.0	2.6	0.0	2.6	2.6	0.0	6.6	3.9
	Contempt	0.0	56.6	0.0	2.6	0.0	11.8	7.9	0.0	1.3	13.2	6.6
	Disgust	6.6	2.6	64.5	2.6	5.3	0.0	0.0	0.0	3.9	5.3	9.2
	Embarrass	0.0	3.9	0.0	63.2	0.0	7.9	0.0	0.0	0.0	15.8	9.2
	Fear	0.0	0.0	1.3	5.3	72.4	0.0	0.0	10.5	3.9	0.0	6.6
	Happiness	0.0	7.9	1.3	1.3	0.0	81.6	1.3	0.0	1.3	1.3	3.9
	Pride	1.3	0.0	0.0	3.9	2.6	0.0	53.9	0.0	3.9	18.4	15.8
	Sadness	2.6	0.0	1.3	3.9	1.3	0.0	0.0	86.8	0.0	0.0	3.9
	Surprise	0.0	1.3	0.0	2.6	5.3	1.3	0.0	0.0	89.5	0.0	0.0
	Neutral	1.3	0.0	0.0	3.9	2.6	1.3	0.0	1.3	0.0	88.2	1.3

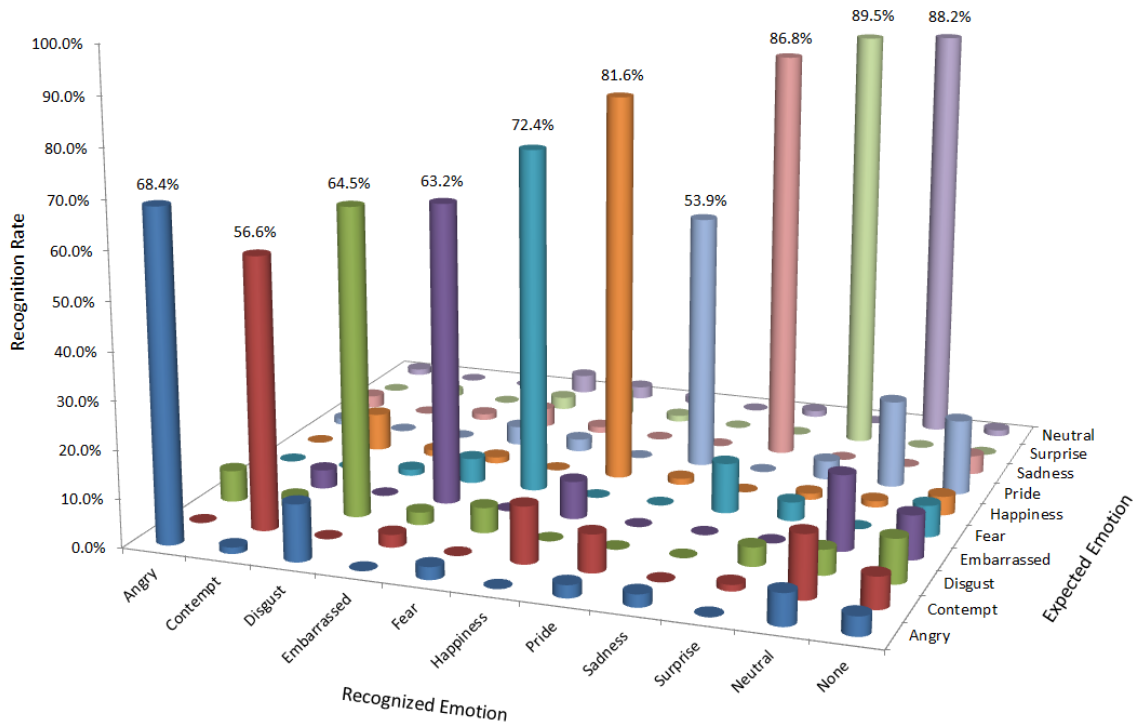


Figure 3.10: Emotion recognition percentages of the 50%-intensity colored images.

The believability of the characters in 100%-intensity images was reported as 3.9 (stdev = 0.80) and in 50%-intensity images reported as 3.7 (stdev = 0.91). Table 3.19 shows the believability for each individual image. In order to find out the effects of intensity on emotion recognition and believability, I performed two ANOVA analyses, for 100% and 50% intensity images. Each ANOVA analysis was a 10×2 (i.e., Emotion \times Intensity) analysis ($df = 1$). Results revealed no significant effect of intensity on emotion recognitions ($p > 0.05$). However, a significant effect of intensity on believability ($F = 9.55, p < 0.05$) was found. The increased intensity may increase the salience of the emotion in images, which causes an increase in believability, which explains the effect of intensity on believability in images. I did not have the similar effect of intensity on believability in video expressions, which can be because videos are more expressive than images, therefore, dynamic feature of the videos dominates the intensity.

Cronbach α values (intra-class correlation) were computed for each of the expressions, using participants' ratings as columns (items) and the 10 videos as rows (cases). Results shown in table 3.19 indicate that for all expressions $\alpha > 0.7$, which means the faces were rated *reliably*.

Table 3.19: Believability (sd = std. deviation) and Cronbach α for Experiment 5.

Measure Emotion	100%-Intensity Group		50%-Intensity Group	
	Believability (sd)	Cronbach α	Believability (sd)	Cronbach α
Anger	4.3 (0.84)	0.993	3.8 (0.85)	0.984
Contempt	4.1 (0.64)	0.981	3.8 (0.78)	0.972
Disgust	4 (0.76)	0.990	3.3 (0.89)	0.982
Embarrassment	3.7 (0.94)	0.991	3.8 (0.84)	0.982
Fear	3.9 (0.79)	0.993	3.5 (0.89)	0.989
Happiness	3.9 (0.79)	0.997	3.5 (0.99)	0.993
Pride	3.4 (0.79)	0.986	3.5 (0.97)	0.974
Sadness	4.1 (0.89)	0.996	3.8 (0.9)	0.996
Surprise	3.9 (0.92)	0.997	3.7 (0.91)	0.996
Neutral	4.1 (0.80)	0.998	3.7 (0.91)	0.996

Experiment 6: Gray-Scaled Image Stimuli Similar to Experiment 5, this experiment was designed to evaluate the correctness of the emotions portrayed by HapFACS in still black and white pictures, which was done mainly to find out if colorfulness influences the subjects’ responses.

Participants: The total of 80 AMT workers were recruited to partake in this survey, and each were asked to select the emotion being shown in each image. Forty subjects were randomly assigned to the 100%-intensity experiment and the other 40 were assigned to the 50%-intensity experiment. Table 3.20 shows the demographic information of the subjects.

Table 3.20: Subjects’ demographic data in each group of Experiment 6.

Experiment Group	Female (Avg. age)	Male (Avg. age)	White	Black	Asian	Hispanic
100% Intensity	57.5% (33.7)	42.5% (31.6)	82.5%	7.5%	7.5%	2.5%
50% Intensity	62.5% (34.7)	37.5% (32.1)	81.3%	0%	9.4%	9.4%

Results and Discussion: Tables 3.21 and 3.22 show the recognition rates of the emotions in 100% and 50%-intensity images respectively. The average recognition rate for 100% and 50%-intensity images were **82.5%** and **77.8%** respectively. Figures 3.11 and 3.12 depict these results in diagrams. Results show that the quality and colorfulness of the images do not bias the subjects’ emotion recognition.

Table 3.21: Emotion recognition ratings of the 100% intensity gray-scaled images.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Images	<i>Anger</i>	75.0	0.0	7.5	0.0	2.5	0.0	0.0	10.0	0.0	0.0	5.0
	<i>Contempt</i>	0.0	62.5	0.0	0.0	0.0	25.0	5.0	0.0	0.0	7.5	0.0
	<i>Disgust</i>	10.0	7.5	75.0	0.0	2.5	0.0	0.0	2.5	0.0	0.0	2.5
	<i>Embarrass</i>	0.0	5.0	0.0	80.0	0.0	0.0	2.5	0.0	0.0	10.0	2.5
	<i>Fear</i>	0.0	0.0	0.0	0.0	87.5	0.0	5.0	7.5	0.0	0.0	0.0
	<i>Happiness</i>	0.0	0.0	0.0	0.0	0.0	92.5	0.0	0.0	7.5	0.0	0.0
	<i>Pride</i>	0.0	0.0	0.0	0.0	0.0	5.0	67.5	0.0	5.0	20.0	2.5
	<i>Sadness</i>	5	0.0	2.5	0.0	0.0	0.0	0.0	90.0	2.5	0.0	0.0
	<i>Surprise</i>	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	97.5	0.0	0.0
	<i>Neutral</i>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	97.5	2.5

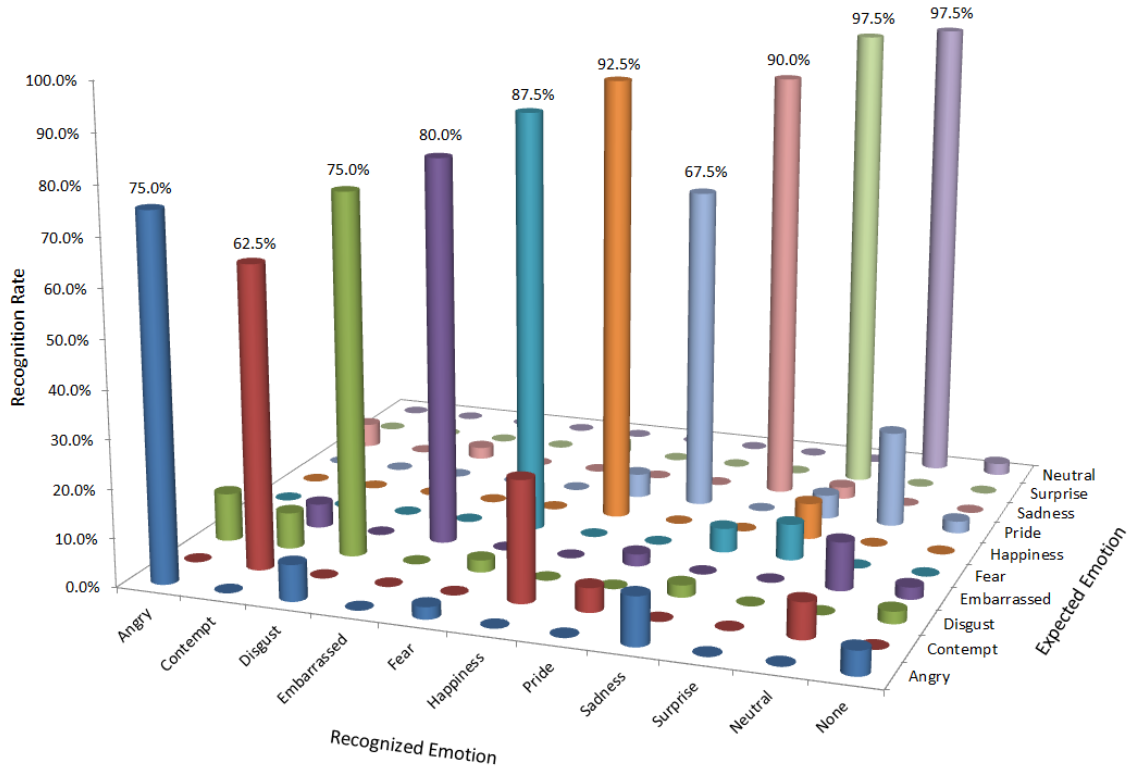


Figure 3.11: Emotion recognition percentages of the 100%-intensity gray-scaled images.

Table 3.22: Emotion recognition ratings of the 50% intensity gray-scaled images.

		Recognized Emotions										
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>	<i>None</i>
Images	<i>Anger</i>	67.5	0.0	0.0	0.0	0.0	0.0	0.0	12.5	0.0	7.5	12.5
	<i>Contempt</i>	0.0	55.0	0.0	7.5	0.0	7.5	0.0	0.0	2.5	17.5	10.0
	<i>Disgust</i>	2.5	0.0	65.0	0.0	0.0	0.0	0.0	2.5	0.0	12.5	17.5
	<i>Embarrass</i>	0.0	5.0	0.0	80.0	0.0	0.0	2.5	0.0	0.0	10.0	2.5
	<i>Fear</i>	0.0	0.0	0.0	0.0	82.5	2.5	0.0	2.5	2.5	7.5	2.5
	<i>Happiness</i>	0.0	5.0	0.0	0.0	0.0	90.0	0.0	0.0	0.0	0.0	5.0
	<i>Pride</i>	0.0	0.0	0.0	0.0	0.0	0.0	62.5	0.0	0.0	25.0	12.5
	<i>Sadness</i>	2.5	0.0	0.0	0.0	2.5	0.0	0.0	87.5	2.5	2.5	2.5
	<i>Surprise</i>	0.0	2.5	0.0	2.5	5.0	2.5	0.0	0.0	87.5	0.0	0.0
	<i>Neutral</i>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0

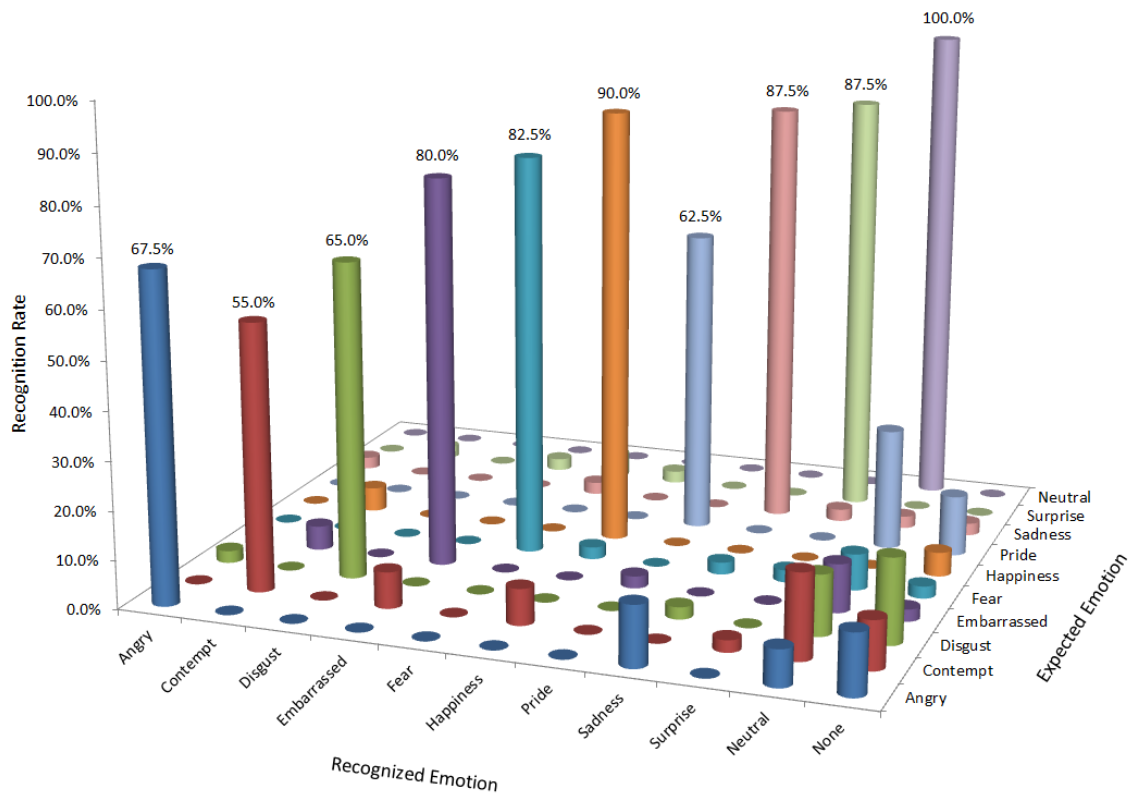


Figure 3.12: Emotion recognition percentages of the 50%-intensity gray-scaled images.

The believability of the characters in 100%-intensity images was reported as 3.8 (stdev = 0.86) and in 50%-intensity images reported as 3.6 (stdev = 0.81). Table 3.23 shows the believability for each individual image. In order to find out the effects of intensity on emotion recognition and believability, I performed two ANOVA analyses, for 100% and 50% intensity images. Each ANOVA analysis was a 10×2 (i.e., Emotion \times Intensity) analysis ($df = 1$). Results revealed no significant effect of intensity on emotion recognitions ($p > 0.05$). However, a significant effect of intensity on believability ($F = 5.88, p < 0.05$) was found, which confirms results of Experiment 5.

Table 3.23: Believability (sd = std. deviation) and Cronbach α for Experiment 6.

Measure Emotion	100%-Intensity Group		50%-Intensity Group	
	Believability (sd)	Cronbach α	Believability (sd)	Cronbach α
Anger	4.0 (0.68)	0.987	3.7 (0.73)	0.979
Contempt	3.9 (0.89)	0.977	3.5 (0.77)	0.959
Disgust	3.4 (0.98)	0.986	3.8 (0.96)	0.978
Embarrassment	3.9 (0.72)	0.994	3.5 (0.9)	0.802
Fear	3.8 (0.83)	0.994	3.7 (0.74)	0.974
Happiness	3.7 (0.78)	0.991	3.4 (0.92)	0.941
Pride	3.6 (1.0)	0.983	3.5 (0.8)	0.753
Sadness	3.9 (0.83)	0.992	3.8 (0.75)	0.943
Surprise	3.9 (0.92)	0.993	3.7 (0.72)	0.969
Neutral	3.8 (0.84)	0.991	3.7 (0.73)	0.916

Cronbach α values (intra-class correlation) were computed for each of the expressions, using participants' ratings as columns (items) and the 10 videos as rows (cases). Results shown in Table 3.23 indicate that for all expressions $\alpha > 0.7$, which means the faces were rated *reliably*.

Evaluating Speaking Characters

Experiment 7: Validation of Emotional Speech For IVA researchers interested in speaking characters, I validated how well HapFACS simulated characters can show emotions while they speak.

Participants: I recruited 20 FIU students and 40 AMT workers as participants for this experiment. Table 3.24 shows the demographic information of the subjects.

Table 3.24: Subjects’ demographics in Experiment 7.

Female (Avg. age)	Male (Avg. age)	White	Black	Asian	Hispanic	Caucasian
61.7% (30.7)	38.3% (28.6)	63.3%	6.7%	8.3%	20%	1.7%

Stimuli and Design: Ten videos were generated (8.3 seconds long in average) with random sentences. The emotions portrayed in the videos were the ones shown in Table 3.7 with 100% intensity. The same 8 models of Experiment 1 were used to portray each video. In all videos, a neutral voice and utterance was used, in order to focus the study on emotional facial expressions (rather than voice or utterance).

Procedure: Each subject was asked to recognize the expressed emotion while the character was speaking. Also, participants were asked to rate the believability of the character while speaking on a 5-level Likert scale (0: not believable at all, 5: very believable).

Results and Discussion: Table 3.25 shows the emotion recognition rates in speaking characters with the average recognition rate of **74.3%**. Figure 3.13 depicts these results in a diagram.

Table 3.25: Emotion recognition ratings of the videos in Experiment 7.

		Recognized Emotions									
		<i>Anger</i>	<i>Contempt</i>	<i>Disgust</i>	<i>Embarrass.</i>	<i>Fear</i>	<i>Happiness</i>	<i>Pride</i>	<i>Sadness</i>	<i>Surprise</i>	<i>Neutral</i>
Videos	<i>Anger</i>	66.7	0.0	5.0	0.0	3.3	0.0	16.7	0.0	5.0	3.3
	<i>Contempt</i>	0.0	45.0	0.0	5.0	0.0	23.3	13.3	0.0	6.7	6.7
	<i>Disgust</i>	11.7	3.3	80.0	0.0	1.7	0.0	1.7	1.7	0.0	0.0
	<i>Embarrass</i>	0.0	6.7	1.7	70.0	0.0	5.0	0.0	3.3	0.0	6.7
	<i>Fear</i>	3.3	1.7	1.7	1.7	76.7	0.0	15.0	0.0	0.0	0.0
	<i>Happiness</i>	0.0	6.7	0.0	0.0	0.0	85.0	0.0	1.7	0.0	3.3
	<i>Pride</i>	0.0	1.7	0.0	1.7	0.0	1.7	48.3	0.0	25.0	1.7
	<i>Sadness</i>	1.7	1.7	3.3	0.0	0.0	0.0	1.7	88.3	0.0	0.0
	<i>Surprise</i>	0.0	0.0	0.0	0.0	1.7	0.0	6.7	0.0	90.0	0.0
	<i>Neutral</i>	0.0	1.7	0.0	1.7	0.0	0.0	0.0	0.0	0.0	93.3

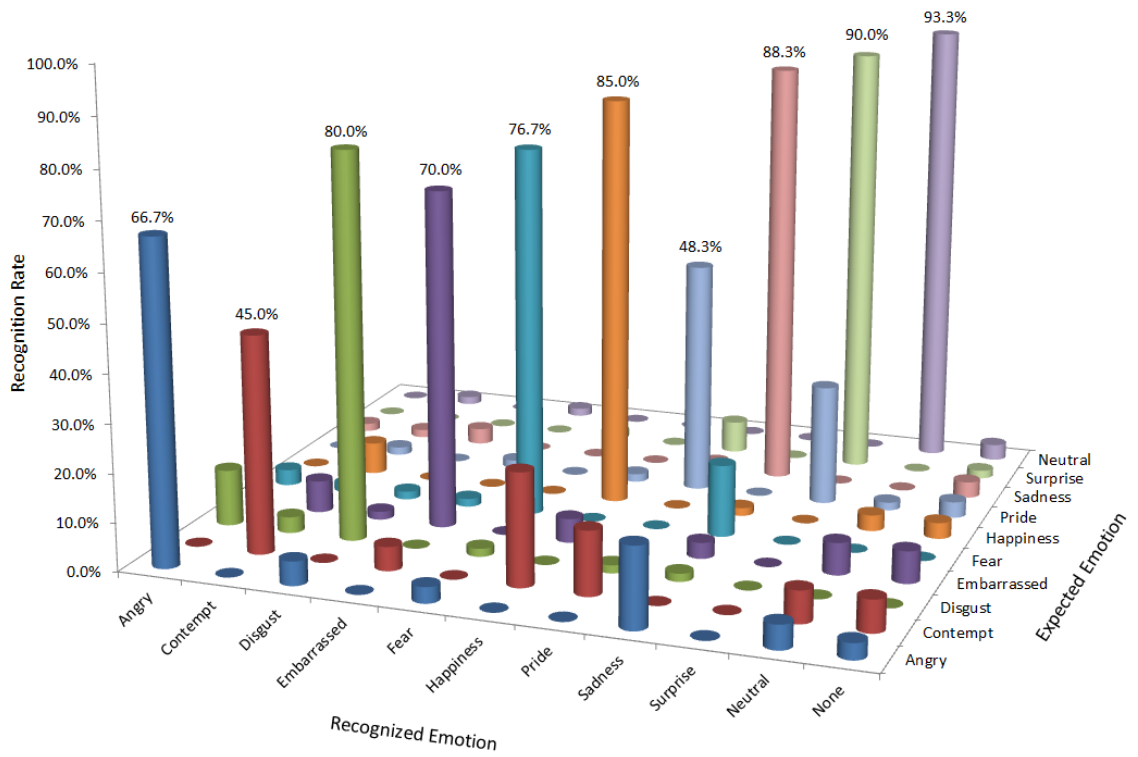


Figure 3.13: Validation of speaking characters showing emotional faces.

While the character is speaking and expressing an emotion, many of the lower-face AUs are activated for speaking. For emotions that are expressed only with lower-face AUs, such as *contempt*, speaking can lower the recognition rates. Also, as confirmed by other studies [LDB⁺10, vdSHFD11], recognition rates for contempt, sadness, disgust, and fear may be lower than happiness, surprise, and anger. Reasons for that could be that (1) they are less universally recognized expressions across cultures [EF86, RE95], or/and (2) they are subject to dialect-like variations in muscle activations [EBLvH07]. Taken together, the low recognition rates on these emotions may be a general feature of the emotion expressions, and not of the presented images/videos.

In average, the believability of the characters in the speaking videos was rated as 3.3 (stdev = 1.01). Table 3.26 shows the believability for each individual video. Cronbach α values (intra-class correlation) were computed for each of the expressions, using participants' ratings as columns (items) and the 10 videos as rows (cases). Results shown in Table 3.26 indicate that the faces were *reliably* rated.

Table 3.26: Believability (sd = std. deviation) and Cronbach α for Experiment 7.

Measure Emotion	Believability (sd)	Cronbach α
Anger	3.3 (0.91)	0.979
Contempt	3.4 (0.91)	0.944
Disgust	2.9 (0.95)	0.989
Embarrassment	3.3 (0.94)	0.983
Fear	3.2 (1.08)	0.987
Happiness	3.6 (0.98)	0.994
Pride	3.2 (1.16)	0.949
Sadness	3.2 (0.95)	0.996
Surprise	3.1 (1.07)	0.996
Neutral	3.6 (0.98)	0.998

Experiment 8: Validation of Lip-Synchronization The purpose of this experiment was to test how well HapFACS characters are able to speak words and sentences in a lip-synchronized manner.

participants: I recruited 18 FIU students and 40 AMT subjects for this experiment. Table 3.27 shows the demographic information of the subjects.

Table 3.27: Subjects’ demographics in Experiment 8.

Female (Avg. age)	Male (Avg. age)	White	Black	Asian	Hispanic
60.3% (33.45)	39.7% (29.9)	72.4%	6.9%	3.4%	17.2%

Stimuli and Design: The same 8 models of Experiment 1 were used to generate the videos, in each of which a random sentence was pronounced. A neutral face was used within all the videos. The videos were of size 485×485 pixels, and on average each one is 5 seconds long.

Procedure: Each subject was asked to rate how well the character’s lips were synchronized with the words it spoke in the video (i.e., how well were the visemes timed and aligned with the phonemes) in a 5-level Likert scale (0: lips are not synchronized at all, 5: lips are completely synchronized). Also, participants were asked to rate the believability of the character while speaking in a 5-level Likert scale (0: not believable at all, 5: very believable).

Results and Discussion: Results show that, subjects reported the lip-synchronization of the characters as 3.28 (stdev = 1.1) accurate with a believability of 3.11 (stdev = 1.1). These results show that the lip-synchronization performance and character-believability while speaking are rated positively, which adds to the abilities of the characters.

Cronbach α values (intra-class correlation) were computed for each of the videos, using participants’ ratings as columns (items) and the 5 videos as rows (cases): $\alpha = 0.928$.

HapFACS and FACSGen

I compared HapFACS with FACSGen, which is the most similar software to HapFACS [KTRS12]. Although FACSGen characters do not have lip-synchronization nor bodies (i.e., only heads) and therefore currently have limited appeal for researchers interested in speaking IVAs, FACSGen is powerful to generate realistic 3D faces. As discussed earlier, we matched our experiment settings (including stimuli and process) to those performed with FACSGen, in order to be able to compare it with HapFACS.

Results: HapFACS expresses 49 individual AUs with average recognition rate of 94.6%, whereas FACSGen expresses 35 individual AUs with average recognition rate of 98.6%. HapFACS is evaluated as 98% accurate in expressing 54 AU combinations while FACSGen is evaluated as 80.1% accurate in expressing the same 54 combinations.

Table 3.28 compares the recognition rates and expression believability of HapFACS and FACSGen for static emotional expressions.

Table 3.28: Comparison between recognition rates and believability of the static emotional expressions generated in FACSGen and HapFACS. FACSGen believability rates are scaled from 7 to 5-scale for comparison.

Emotion	HapFACS		FACSGen	
	100% (Bel.)	50% (Bel.)	100% (Bel.)	50% (Bel.)
<i>Angry</i>	75 (4.0)	67.5 (3.7)	87.82 (3.6)	71.79 (3.3)
<i>Contempt</i>	62.5 (3.9)	55 (3.5)	56.41 (3.1)	48.08 (3.1)
<i>Disgust</i>	75 (3.4)	65 (3.8)	68.59 (3.0)	61.54 (2.9)
<i>Embarrassment</i>	80 (3.9)	80 (3.5)	69.23 (3.4)	60.26 (3.1)
<i>Fear</i>	87.5 (3.8)	82.5 (3.7)	72.44 (3.2)	67.31 (3.1)
<i>Happiness</i>	92.5 (3.7)	90 (3.4)	88.46 (3.7)	77.56 (3.4)
<i>Pride</i>	67.5 (3.6)	62.5 (3.5)	74.36 (3.7)	71.79 (3.5)
<i>Sadness</i>	90 (3.9)	87.5 (3.8)	83.97 (3.4)	76.28 (3.3)
<i>Surprise</i>	97.5 (3.9)	87.5 (3.7)	87.82 (3.7)	87.82 (3.5)
<i>Neutral</i>	97.5 (3.8)	100 (3.7)	-	-

General Discussion

I studied HapFACS validity for creating AUs defined in FACS, and the emotional meaning conveyed by HapFACS expressions. Experiment 1 reported validation data for 49 single AUs. Experiment 2 reported validation data for 54 AU combinations and validation of individual AUs used in those combinations. All the expressions were implemented in faces of different sexes and ethnicity.

The recognition rates of the AUs were high and the AUs interacted predictably in combination with each other (to generate facial expressions of emotions). For all AUs, validity of the AU appearance was scored satisfactorily by FACS-certified coders. Based on the reported high recognition rates for combinations and for individual AUs, obtained with the good intra-rater reliability scores, results suggest that the AUs synthesized by HapFACS are valid with respect to the FACS.

Overall, when performed with high intensity, *surprise*, *fear*, *happiness*, *sadness*, and *neutral* were the most easily recognizable emotions, whereas *contempt* and *pride* were the most difficult to detect. When performed with low intensity, *surprise*, *happiness*, and *neutral* were the easiest to recognize, while *anger*, *contempt*, *pride*, and *disgust* were more difficult to recognize.

Contempt expression is sometimes perceived as *pride* (or even *happiness*), which we hypothesize is due to the subtle asymmetric smile that can be displayed in pride (but in happiness as well). The low recognition rate of *contempt* also confirms the findings by Langner et al. [LDB⁺10] and Van der Schalk et al. [vdSHFD11], who discuss the reason as a general feature of the contempt expression, which is not as expressive nor visual as other expressions.

Experiments 3 to 6 showed that participants recognized the expected affective meanings conveyed by emotional expressions generated with HapFACS. The reported

recognition rates were high and comparable to previous research [BH05, GdRLV08, LDB⁺10, TRS09b, vdSHFD11].

3.2 HapGest: Haptek Gesture Synchronizer

When I pass a sentence to the Haptek virtual character to be uttered, I expect the character to (1) speak out the sentence, and (2) perform the appropriate non-verbal gestures at appropriate times based on the output of the gesture models. However, the Haptek does not provide any synchronized control on the non-verbal behaviors of the character while speaking. In other words, Haptek does not enable us to know when the audio corresponding to a specific text has been reached. Therefore, the character does not have enough control over the spoken words, in order to express specific gestures while specific words are being pronounced. I implemented a module for the Haptek characters, called HapGest, which synchronizes the verbal and non-verbal expressions and provides the fore-mentioned capability.

For example, let's say you would like a Haptek character to speak out the sentence "OK, I understand your situation". At the same time, you would like the character to perform a head nod when it is pronouncing the "OK" and perform a hand point gesture while it is pronouncing the "your" word. As mentioned before, original Haptek API does not provide us with this ability. HapGest, as I will explain next, (1) works as a Markup Language, which enables us to include XML tags in the sentence, in order to indicate where the gestures should be expressed, and (2) interprets the XML tags and sends appropriate hypertexts to the Haptek character, in order to express the desired gestures at the right time.

As mentioned before in Section 2.4.1, in general, there are two ways to synchronize between a character animation and a the character's speech (either through a TTS engine or from recorded audio samples): (1) estimate word and phoneme timings and

construct an animation schedule prior to execution; (2) use real-time events from a TTS engine and compile a set of event-triggered rules to govern the generation of the non-verbal behaviors. In HapGest, I used the second approach using the Microsoft Speech API's (SAPI) events.

The SAPI has an XML Markup Language, which provides different control handles on its output audio. For example, I can manipulate the output audio's volume, speed, rate, and pitch, as well as words' spelling, emphasis, silence, and pronunciation, in addition to the voice type, and language. One of the provided controls by SAPI is *throwing events* when the TTS reaches a specific word in the sentence. This control is called *bookmarking tag*, and performed using the *Bookmark* tag. The Bookmark tag inserts an event into the output audio stream. I used this event to signal the application when the audio corresponding to a specific word is reached. The Bookmark tag has one attribute, called *Mark*, whose value is a string. This value can be used to differentiate between bookmark events. For example, in the above example sentence, HapGest adds two Bookmark tags to the sentence as follow: "OK <Bookmark Mark='head-nod'>, I understand your <Bookmark Mark='hand-point'> situation."

As depicted in Figure 3.14, HapGest architecture includes the following three modules:

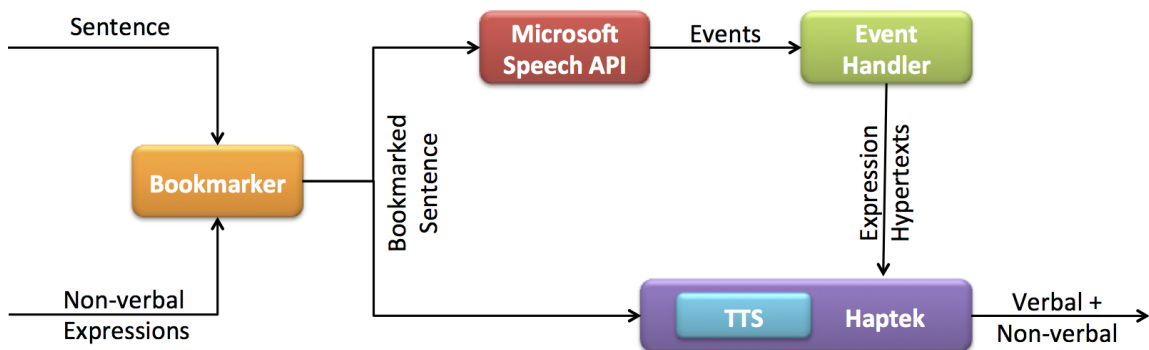


Figure 3.14: HapGest architecture.

1. **Bookmarker**, receives a sentence (e.g., “Ok, I understand your situation”) and a list of non-verbal expressions that are supposed to be animated while each word in the sentence is being pronounced (e.g., head-nod, neutral, neutral, hand-point, neutral). For each non-neutral non-verbal expression assigned to a word (by the non-verbal behavior models), the Bookmarker attaches a Bookmark tag to the word with a Mark value equal to the name of the expression. Therefore, the output of the Bookmarker is the input sentence with the Bookmark tags inserted into it. The bookmarked sentence is sent at the same time to the Microsoft SAPI and the Haptek character. Haptek has an internal Text-To-Speech (TTS) to speak out the sentence.
2. **Microsoft Speech API**, receives the bookmarked sentence from the Bookmarker module. Both the Microsoft SAPI and the Haptek TTS process the bookmarked sentence at the same time. The reason of using a second Microsoft SAPI instance, while there is an internal one in the Haptek, is that Haptek does not provide control over its internal SAPI, in order to handle events. I mute the volume of Microsoft SAPI module, in order to hear only the audio output from Haptek TTS. Microsoft SAPI module and Haptek TTS start reading the sentence at the same time. While Haptek TTS is speaking out the words, the Microsoft SAPI throws appropriate events defined by the Bookmark tags.
3. **Event Handler**, catches the events thrown by the Microsoft SAPI module at each Bookmark tag, and handles the events. Therefore, *Event Handler* (1) catches each Bookmark event, (2) reads the Mark attribute value of the event, in order to understand its corresponding expression, (3) generates a Haptek hypertext to call an appropriate Haptek switch, which expresses the required expression, and (4) sends the hypertext to the Haptek character to be animated at the same time that its corresponding word is being pronounced.

Following the above process, I was able to synchronize the verbal and non-verbal modalities of the HapteK character. I called this synchronization composite the *HapGest*.

CHAPTER 4

On-Demand Virtual Counselor (ODVIC)

I followed three main reasons for implementing the On-Demand Virtual Counselor (ODVIC): (1) finding out whether delivering the counseling material via virtual characters can enhance the user acceptance, (2) developing some early insights into the potential impact of an empathy model on the user's acceptance of the character with enabling the virtual character to build rapport with the clients, even with a simple rule-based approach, and (3) developing a framework and test-bed for evaluating our empathy model.

I based my intervention content on an existing computer-based and evidenced-based Adaption of Motivational Interviewing (AMI) intervention named the Drinker's Check Up (DCU), which has been implemented as a text-only web-based system¹ [MR10, HSD05]. The DCU specifically targets excessive drinking behaviors. It is claimed that people can reduce their drinking by an average of 50% using this AMI. The DCU is the most widely used brief AMI, where the client is given feedback, in an MI "style" based on individual answers from standardized assessment measures [BAD02].

Because the ODVIC is aimed at being a test-bed to evaluate my non-verbal model of rapport for generating an expressive character - rather than to test drinking health outcomes of patients - my system differs from the current text-only web-based DCU in that (1) although all the five psychometric assessment instruments used in DCU are implemented, I used only one of them, namely the AUDIT, in the evaluation process, in order to keep the evaluation time shorter; and (2) a multimodal sensing and expressing character delivers the assessment instruments.

¹<http://www.drinkerscheckup.com/>

In the following sections, I will discuss my approach to develop a novel modality for the computer-delivery of Brief Motivational Interventions (BMIs) for behavior-change in the form of a personalized **On-Demand Virtual Counselor (ODVIC)**, accessed over the internet. ODVIC is a multimodal Embodied Conversational Agent (ECA) who empathically delivers an evidence-based behavior-change intervention by adapting, in real-time, its verbal and non-verbal communication messages to those of the user's during their interaction. The current focus of this work is on excessive alcohol consumption as a target behavior, but the approach is adaptable to other target behaviors (e.g., overeating, lack of exercise, drug use). As mentioned earlier, I based my current approach on the successful existing patient-centered brief motivational intervention called DCU [Mil88, HSD05], whose computer-delivery with a text-only interface has been found effective to reduce alcohol consumption in problem drinkers.

4.1 Introduction

There is a growing societal need to address the increasing prevalence of behavioral health issues, such as obesity, alcohol or drug use, and general lack of treatment adherence for a variety of health problems. The statistics, worldwide and in the USA, are daunting. Excessive alcohol use is the third leading preventable cause of death in the United States [Nat11] (with 79,000 deaths annually), and is responsible for a wide range of health and social problems (e.g., risky sexual behavior, domestic violence, loss of job). Alcoholism is estimated to affect 10-20% of US males, and 5-10% females sometime in their lifetimes.

Similar risks exist with other forms of substance abuse. In 2010, the *World Health Organization (WHO)* reported that obesity - worldwide - has more than doubled since 1980. In 2011, 1.5 billion adults in the world were overweight, of which 500 million were obese, and 43 million children under the age of five were overweight [WHO11].

In the USA alone, obesity affects 33.8% of adults, 17% (or 12.5 million) of children and teens, and it has tripled in one generation. These behavioral issues place people at risk of serious diseases; e.g., obesity can lead to diabetes, alcoholism to cirrhosis, physical inactivity to heart disease.

On the positive side though, these behavioral health issues (and associated possible diseases) can often be prevented with relatively simple lifestyle changes, such as losing weight with a diet and/or physical exercise, and learning how to reduce alcohol consumption. Medicine has therefore started to move toward finding ways of preventively promoting wellness rather than solely treating already established illness. In order to address this new focus on wellbeing, health promotion interventions aimed at helping people to change their lifestyle have been designed and deployed successfully in the past few years.

Evidence-based patient-centered *Brief Motivational Interviewing (BMI)* interventions have been found particularly effective in helping people find intrinsic motivation to change problem behaviors (e.g., excessive drinking and overeating) after short counseling sessions, and to maintain healthy lifestyles over the long-term [ER01, DDR01]. A methodological review of clinical trials of 361 treatments showed that out of 87 treatment methods, the top two ranked treatment styles were: 1) Brief Interventions and 2) Motivational enhancement therapies [MR02]. It is reported that 5 minutes of advice and discussion about behavioral problems (e.g., alcohol or drug use) following a screening can be as effective as more extended counseling, and that a single session can be as effective as multiple sessions [BG92].

Lack of locally available personnel well-trained in BMI, however, often limits access to successful interventions for people in need. Yet, the current epidemic nature of these problems calls for drastic measures to rapidly increase access to effective behavior change interventions for diverse populations. To fill this accessibility gap,

evidence has accumulated about the general efficacy of *Computer-Based Interventions (CBIs)* [Hes97, BTB⁺08, Ski94, Cun99, PSSJC08].

The success of CBIs, however, critically relies on insuring engagement and retention of CBI users so that they remain motivated to use these systems and come back to use them over the long term as necessary (e.g., for booster sessions, follow-ups, and lifestyle maintenance sessions). Whereas current BMI interventions delivered by computers have been found effective, high drop-out rates due to their users' low level of engagement during the interaction limit their long-term adoption and potential impact [PSSJC08, Ver10].

One crucial aspect positively affecting the health outcomes of BMIs (and most counseling techniques for that matter), involves the ability of the therapist to establish rapport and to express empathy [MR02]. As discussed in Section 2.2, *Empathy* is a complex phenomenon with different types of definitions. However, there is a general consensus that empathy can involve cognitive attributes or affective attributes, which can also be combined during full-blown empathy [GM85].

Because of their text-based only interfaces, current CBIs can therefore only express limited empathy (mostly reflected in the choice of textual wording of the intervention). Fortunately, in the last decade, at the same time as CBIs are being developed and studied in healthcare, computer science research has progressed in the design of simulated human characters and avatars with anthropomorphic communicative abilities [CSPC00]. Expressive virtual characters have become increasingly common elements of user interfaces for a wide range of applications, such as interactive learning environments, e-commerce, digital entertainment, and virtual worlds.

Virtual characters who specifically focus on dialog-based interactions are called *Embodied Conversational Agents (ECAs)*, also known as Intelligent Virtual Agents (IVA). ECAs are digital systems created with an anthropomorphic embodiment (be

it graphical or robotic), and are capable of having a conversation (albeit still limited) with a human counterpart, using some artificial intelligence broadly referred to as an “agent”. With their anthropomorphic features and capabilities, they interact using humans’ innate communication modalities, such as facial expressions, body language, speech, and natural language understanding, and can also contribute to bridging the digital divide for low reading and low health literacy populations, as well as for technophobic individuals [NK11, BPJ09].

Therefore, I posit that (1) using well-designed virtual empathic and rapport-enabled characters (i.e., ECAs) for the delivery of BMIs has the potential to increase users’ engagement and users’ motivation to *continue* to interact with them, and that as a result (2) users’ increased exposure to engaging evidence-based BMIs will increase their effectiveness for behavior change.



Figure 4.1: ODVIC Amy in her office.

In the rest of this chapter, I first review the current research on BMIs, I then discuss my approach to develop a novel modality for the computer-delivery of BMIs for behavior change in the form of a 3D personalized On-Demand Virtual Counselor (ODVIC), accessed anytime anywhere over the internet (see Figure 4.1). I then discuss how I designed the ODVIC to partially simulate both aspects of empathic communication (affective and cognitive), using a scheme for the agent’s dynamic dis-

play of facial expressions and verbal reflective listening based on the user's perceived expressions and answers.

Without claiming that my virtual character can fully empathize with the user, which would require the ability to subjectively experience and understand the user's feelings, I then show, with results of user studies, that the ODVIC has enough expressive abilities to provide the user with a better experience than when interacting with the DCU delivered with a text-only interface, or with a non-expressive character.

4.2 Motivational Interviewing (MI) and Brief MI

Motivational Interviewing (MI) has been defined by Miller and Rollnick [MR02] as a directive client-centered counseling style for eliciting behavior-change by helping clients to explore and resolve ambivalence. One of MI central goals is to *magnify discrepancies that exist between someone's goals and current behavior*. MI basic tenets are that (1) if there is no discrepancy, there is no motivation; (2) one way to develop discrepancy is to become ambivalent; (3) as discrepancy increases, ambivalence first intensifies; if discrepancy continues to grow, ambivalence can be resolved toward change.

In the past few years, adaptations of MI have mushroomed with the purpose to meet the need for motivational interventions within medical and healthcare settings [BAD02] where sessions can be as short as 20-40 minutes.

Furthermore, whereas initially used with addictive behavior problems, such interventions have been adapted and implemented with great success for a variety of behaviors, ranging from diabetes self-management [Doh00] to treatment adherence among psychiatric patients [Swa99] to fruit and vegetable intake among African Americans [Res00], among other target behaviors.

BMIs combine MI style of communication with the common underlying elements of effective brief interventions characterized by the acronym FRAMES [Bie93, Mil94]: **F**eedback about client's individual status is personalized, **R**esponsibility for changing is left with the individual, **A**dvice is provided in a supportive manner, **M**enus of different options for changing that respect individual's readiness to change are offered, **E**mpathic style of communication is central to the individual-clinician relationship; and **S**elf-efficacy is nurtured and emphasized.

Because BMIs are highly structured – first, *assessment* of target behavior patterns, then normative *feedback*, then *menu* of change options depending on client's readiness – they lend themselves well to computer-delivery [LW08b], while remaining effective [Hes97, BTB⁺08] and well-accepted by people [Ski94, Cun99].

Internet-Delivered Interventions, in particular, present a number of advantages over traditional modes of delivery [PSSJC08]: they are able to reach a large audience in a cost effective manner (possibly in remote locations) with 24-hour access; they offer participants privacy and anonymity (users tend to disclose more information about risky behaviors to them than to human counselors [SS86]); they can automatically tailor information derived from individual assessment to an individual's specific needs [BSV⁺96, NBH07]; they can diminish variability between different counselors, which accounts for 25% to 100% changes in rates of improvement among clients [MR02]; and they demonstrate infinite patience to respect the individual's readiness to change (sometimes very slow coming) [PV97]. It is also interesting to note that internet-delivered interventions for alcohol reduction are particularly useful for people less likely to access traditional alcohol-related services, such as women and young people [WKS⁺10].

The BMI intervention called the *Drinker's Check-Up (DCU)* [Mil88] is the focus of my current work. DCU has been computerized as a menu-based text-only intervention

delivered over the internet, that specifically targets excessive drinking behaviors and with which heavy drinkers can reduce their drinking by an average of 50% at 12-month follow-ups [SH04, HSD05]. My ODVIC delivers the same intervention content as DCU, with additional empathic messages that I describe in the next sections.

4.3 Health Counselor System Architecture

In an effort to address the limitations of current computer-based interventions, namely users' loss of interest over the long term and drop-outs (which are also problematic in classical face-to-face interventions), my approach is to (1) use ECAs and also (2) leverage users' acceptance of ECAs by developing an expressive empathic 3D animated character. My virtual agent is able to perceive the user's (i.e., client's) facial expressions and text entries as it delivers the adapted content of the DCU [SH04, HSD05] in an empathetic style. It combines (1) partial non-verbal mimicry (head nods and facial expressions), essential in building rapport and expressing empathy [BBLM86], and (2) verbal reflective listening (RL) considered as one of the main ways of conveying empathy in patient-centered interventions [Rog59].

4.3.1 System Overview

The ODVIC system delivers personalized and tailored behavior-change interventions via multi-modal verbal and non-verbal channels. The system is developed in the .NET framework as a three-tier architecture. The system architecture is composed of the main modules shown in Figure 4.2, which I describe in details in this section.

During the interaction, the user's utterances are processed by the *Dialog Module*, which directs the MI sessions and elicits information from the user. *Non-verbal Communication Module*, captures and processes user's facial expressions in real-time

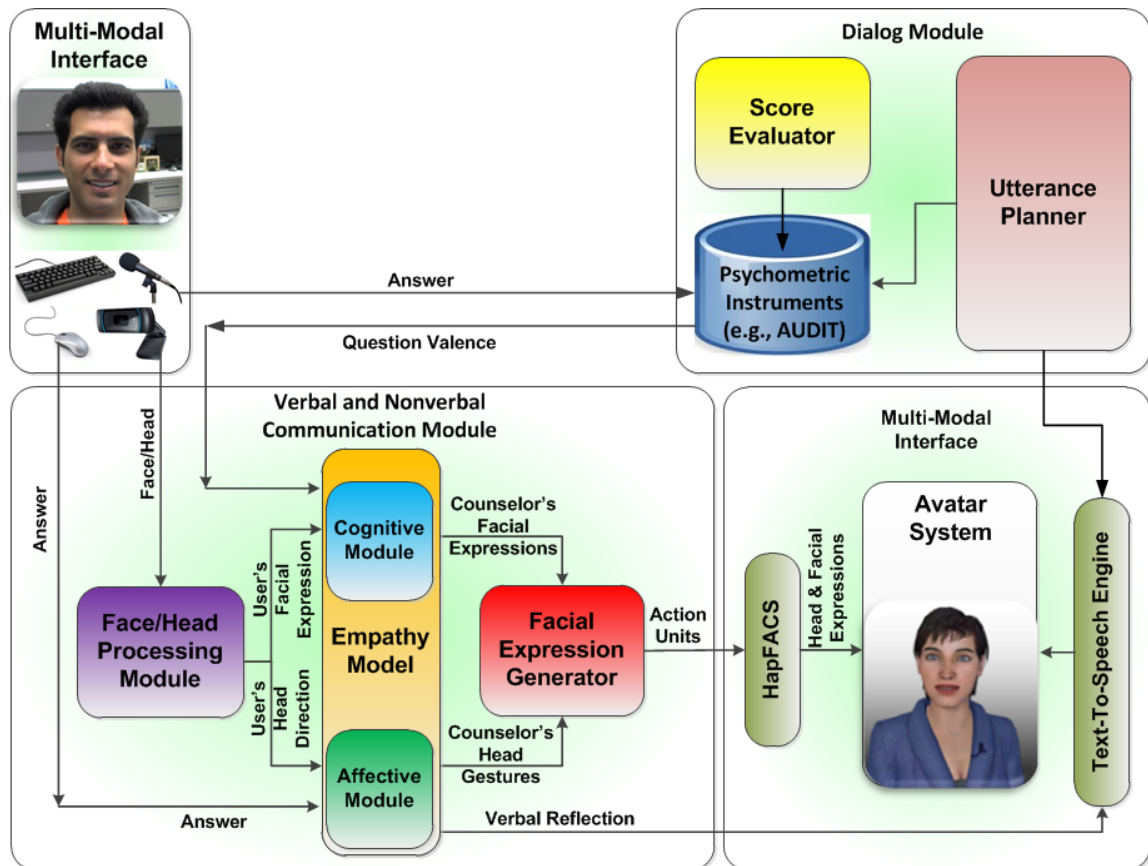


Figure 4.2: System Architecture.

to assess the user’s most probable affective states, then combines with affect related information elicited from utterances to decide about the counselor’s empathic responses. In this way, I can convey an ongoing sense of empathy and rapport via a *Multi-modal Avatar-based Interface* (using a 3D animated virtual character with verbal and non-verbal communication). The *Score Evaluator* performs the required psychometric analysis based on the information collected by the *Dialog Module*, and its results are maintained in a database over multiple sessions to offer a dynamically tailored intervention in the form of normative feedback or specific behavior change plans.



Figure 4.3: Ethnicity and gender concordance.

4.3.2 Avatar-Based Multi-Modal User Interface

The web-based user interface of the system uses an embedded anthropomorphic ECA, which can deliver verbal communication with automatic lip synchronization, as well as non-verbal communication cues (e.g., facial expressions, head nods, mutual gaze, and head movements).

I integrated a set of features that has been considered necessary for health promotion interventions [LYL⁺12] (using the Haptেক avatar system [WS96, SW97, SW00]):

1. A 3D graphical avatar whose appearance is well-accepted by users as documented in earlier studies [LBAM04, LYL⁺12].
2. A selection of different avatars with different ethnicities (e.g., skin color, and facial proportions) in both genders shown in Figure 4.3.
3. A subset of the facial expressions implemented by HapFACS (see Section 3.1).
4. A Text-To-Speech (TTS) engine able to read text in a lip synchronized manner.
5. Lip-synchronized pre-recorded voices for the text provided in the interventions.

4.3.3 Dialog Module

The *Dialog Module* evaluates and generates dialog utterances using three components: *Utterance Planner*, a collection of *Psychometric Instruments*, and *Score Evaluator*.

The interaction of the client with the system is based on a series of dialog sessions, each of which having a specific assessment goal to identify different aspects of the user’s drinking behavior problem (if any). Each session is based on a psychometric instrument.

4.3.4 Psychometric Instruments

I used different well-validated Psychometric Instruments (i.e., questionnaires) used in the DCU [MR10, HSD05] (which are also commonly used by therapists to assess an individual’s alcohol use in the assessment sessions [MR02]). These questionnaires are kept inside a database. Each psychometric instrument contains a set of questions representing the plan for that assessment session, and a set of response options for each question.

Although the full-fledged DCU intervention is implemented in my system to deliver tailored interventions and behavior-change plans, I focused my testing on the psychometric analysis portion using one instrument called AUDIT.

Alcohol Use Disorders Identification Test (AUDIT)

AUDIT [BHBSM01] is a 10-item questionnaire that I use to identify people whose alcohol consumption has become hazardous or harmful to their health. The “amount and frequency of drinking, alcohol dependence, and problems caused by alcohol” are queried using this instrument. Questions are scored using a 5-point Likert scale. The total score is the summation of all the answers’ scores. Table 4.1 shows the way AUDIT scores are interpreted. The cut-off numbers may be different based on average body weight, gender, race, and cultural standards.

Table 4.1: AUDIT score interpretation.

AUDIT Score	Interpretation
score < 4	No drinking problems
$4 \leq \text{score} \leq 8$	Harmful for ages under 18 and females
score > 8	Alcohol dependence.
$8 < \text{score} \leq 15$	Should be advised to reduce drinking
$16 \leq \text{score} \leq 19$	Should be suggested counseling
score ≥ 20	Should be warranted further diagnose

4.3.5 Utterance Planner

This component of the *Dialog Module* decides about the next utterance based on the previous interactions with the client. To measure the client’s score in each context, there are sets of questions for that context, which are conceptually related with each other. The utterance planner aims at detecting discrepancies between client’s answers for questions in the same context.

This module follows a set of well-documented MI techniques known as OARS to generate the dialogs: Open-ended questions, Affirmations, Reflective listening, and Summaries. These techniques are applied toward goals concerning specific behaviors (e.g., excessive alcohol use, drug use, overeating). The engine is not covering the open-ended questions in the current implementation, instead it uses predefined questions and answers within each static psychometric instrument. Therefore, in the current implementation of the Utterance Planner, utterances (i.e., questions) are followed in a predefined order.

Reflective listening is a client-centered communication strategy involving two key steps: seeking to understand a speaker’s thoughts or feelings; and conveying the idea back to the speaker to confirm that the idea has been understood correctly [Rog59].

4.3.6 Score Evaluator

The *Score Evaluator* is responsible for processing the psychometric data collected in the *Dialog Module*. Based on instructions in each instrument, the *Score Evaluator* module calculates the score of the client for a particular measuring instrument. The result is used to identify specific aspects of the drinking problem, such as the amount and frequency of drinking, alcohol dependence, and problems caused by alcohol. Also, the score is used in the Empathy Model (discussed next) to empathize with the user based on his/her history of answers.

4.3.7 Empathy Model

This Module is responsible for empathizing with the user. In order to adapt the non-verbal behaviors of the character with the non-verbal behaviors of the user, I implemented a *Face/Head Processing Module*, which captures the client's face/head images through a camera and recognizes his/her facial expressions and head movements using a face recognition engine (algorithm published in [TW11, WHT10]).

Outputs of the *Empathy Model* include (1) the counselor's facial expressions, (2) counselor's head movement, (3) and counselor's head gestures. This preliminary empathy model does not use the textual transcripts, however, in the rapport model discussed in Chapter 5, the dialog contents are also used as the input, and also, more gestures and expressions are added to the non-verbal behaviors of the character.

The decisions made by the *Empathy Model* are sent to the *Facial Expression Generator*. This module returns the face, head, and eye AUs to be activated (with their activation intensities). The AUs and their intensities are then sent to the HapFACS API, which can map the AUs to the virtual character's face and head.

Discussing issues about at-risk behaviors, such as heavy drinking, are emotional for people to talk about (e.g., shame, discouragement, anger, hopefulness, satisfaction,

and pride). Empathy and positive regard toward the client are therefore critical therapeutic conditions to create an atmosphere of safety and acceptance, where clients feel free to explore and change [MR09] at-risk behaviors. As discussed earlier, in MI and BMI sessions, what is crucial is the ability of the therapist to *express accurate empathy* by applying “a skillful reflective listening to clarify and amplify the [user’s] own experiencing and meaning” [MR02].

The *Empathy Model* emulates two kinds of empathy: *affective empathy* and *cognitive empathy*. Affective empathy refers to the ability to react emotionally when one perceives that another is experiencing, or about to experience, an emotion [Wis87]. Cognitive empathy involves an understanding (rather than a feeling) of another’s experiences and concerns, combined with the capacity to communicate that understanding [Hoj07].

Whereas my system does not understand the subjective experience of the user’s emotions, it does perceive the user’s emotions (with computer vision) and reacts emotionally (with 3D realtime animations) to convey affective empathy and a sense of rapport.

The *Empathy Model* captures and processes user’s facial expressions in real-time to assess the user’s most probable affective states, then combines it with affect related information elicited from utterances to decide about the counselor’s empathic responses. It is responsible for simple verbal reflection of user’s answers, and for other feedbacks, such as facial expressions, and head nods.

This model uses a set of inputs to decide about the counselor’s empathic behaviors:

1. Emotional facial expressions: facial photos are taken using the camera through the JPEG-Cam Flash/Javascript library, saved as an image file on the server, and sent to the face recognizer server. The system recognizes the client’s emotional facial expressions and categorizes them into five categories of happy, sad,

angry, surprised, and neutral. The *Face/Head Processing Engine* uses the facial expression recognition algorithm proposed in [TW11, WHT10].

2. Head movements: the *Face/Head Processing Engine* returns degrees of the three possible head movements: *head yaw* (up and down), *head pitch* (left and right), and *head roll* (left and right rolls);
3. Smile: the *Face/Head Processing Engine* returns the user's smiling status as one of its outputs (smile with an open mouth). The smile status is slightly different from happy facial expression. The happiness is recognized from different movements of the face, such as eyes, cheeks, and lips. But, smile is only the state of the lips.
4. Counselor's question valence: the counselor can expect whether her/his question will be pleasant or unpleasant for the client. So, the counselor simulates the role-taking mode of empathy and puts herself/himself in the client's shoes to guess her/his emotion in response to asking each question. After asking a question, the client appraises it based on her/his goal and situation, and reacts emotionally to it. For each *Psychometric Instrument*, the *Empathy Model* uses the OCC [OCC88] cognitive structure of emotion to predict the client's emotions. Based on the OCC, one feels *joyful* if she/he is *pleased* about a *desirable* event, and feels *distressed* if she/he is *displeased* about an *undesirable* event. I manually assigned a valence value (i.e., pleasurable or unpleasurable) to each question in the database.
5. Client's answer to the counselor's question: for any counselor's question, the client provides an answer through a menu-based interface using mouse/keyboard. The client's answers are passed to the *Score Evaluator* module to evaluate her/his answer's risk level (and returned as a score).

6. History of the client's previous answers: after receiving each answer from the client and scoring it using the *Score Evaluator*, a cumulative score is calculated for the client based on her/his history of answers until then. This cumulative score shows the alcohol consumption risk level of the user. Based on the user-model [YAL12], in different assessment sessions, this score can represent the strength of the client's dependence to alcohol, drinking risk factors, motivation to change, frequency of drinking, and drinking consequences.

Given the above parameters, the empathy model decides which affective/cognitive empathic responses to express. The *Empathy Model* contains a rule-based system, which uses a set of pre-defined rules in a *Decision Tree*, to decide about the next counselor's empathic reaction to the client, both verbally and non-verbally. This system decides "what facial expression to express", "when to show head nods", "what eyebrow expressions to show", "when to express subtle/large smile", and "what verbal reflections to express".

For each user's answer to the questionnaire items, the empathy model returns a *simple* verbal feedback from a pool of appropriate verbal feedbacks for that answer (saved in database). *Verbal reflection* is commonly used by counselors to create a stronger connection with the clients and create closeness and rapport. Verbal reflections can be simple or complex. Simple reflections can be a repetition or rephrasing of the client's responses. For example, counselor asks "How often do you have a drink containing alcohol?", the client selects the answer "Two to three times a week", then the counselor reflects back "So, you drink at least twice a week".

When the counselor's question valence is positive or the user's emotional facial expression is positive (e.g., happy, surprised), the decision tree tends to express a positive facial expression and vice versa. Also, lower risk level of the user's answer and lower overall score (history of answers) cause more positive expressions, and vice

versa. For example, if (1) the counselor asks “Does drinking help you to relax?”, which is a positive valence question, (2) the user expresses a “happy” facial expression, (3) the user answers “No” (i.e., low risk answer), and (4) user has the AUDIT score of 5, which is a low score, then the decision tree returns a happy face with a large smile and raised eyebrows. A small sample part of the decision tree (located the cognitive module) is shown in Figure 4.4.

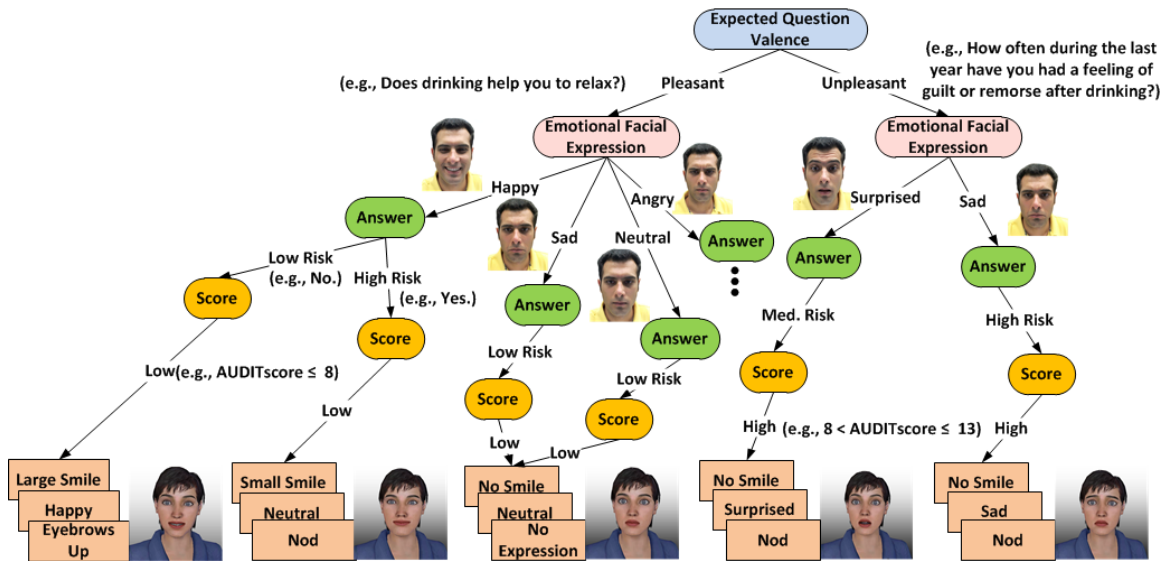


Figure 4.4: A sample piece of the decision tree used in cognitive module.

Facial Expression Generator

The *Facial Expression Generator* generates the virtual character’s facial expressions and head movements based on the Facial Action Coding System (FACS) [EF78], and Emotional FACS (EmFACS) [FE83] (described in Section 2.5.1).

The *Facial Expression Generator* module uses *HapFACS* (described in Section 3.1) to generate facial expressions based on FACS and EmFACS. It accepts the outputs of the *Empathy Model*, and maps them to their appropriate AUs. The AUs are then passed to the HapFACS API to generate the emotional facial expressions on the character face.

In the next section, I describe the evaluation of the system, conducted to compare three delivery modes, text-only, non-empathic, and empathic character.

4.4 ODVIC Evaluation

Since the performance criteria of the conversational agents, such as virtual health counselors, is dependent upon the satisfaction of their users, it is necessary to measure the users' perception of the agent's action.

I designed an evaluation scheme to evaluate the user's acceptance of the virtual counselor and to evaluate the character's properties (e.g., likability, animacy, and anthropomorphism). I combined two questionnaires developed by Heernik et al. [HKEW09] and Bartneck et al. [BKC08] and adapted them to my health counseling application.

4.5 Hypotheses

I hypothesize that counselors with different delivery modalities (i.e., virtual character vs. text) and different levels of empathizing abilities (e.g., with or without non-verbal expressions displayed appropriately at specific times based on the content of the interaction) will have different effects on the quality of the interaction with users.

I expect the character with empathizing abilities (e.g., appropriate facial expressions, head nod, verbal reflective listening) to have more positive effects than a neutral character and a text-only system, in terms of the users' acceptance of the system, among other measures.

Whereas it may seem intuitive that the system with empathizing abilities would outperform both other systems, studies found that neutral characters and text-only systems can at times be perceived better than empathic ones with respect to some

features [BP05], such as the interaction time per day/week and steps followed by the users (maybe because interaction with the text-only and neutral systems goes on faster than the empathic one and users can follow more steps in an equal amount of time). These findings motivated my choice to compare all three versions of the system: text-only, non-empathic character, and empathic character. And indeed my results also indicate that perceived ease of use (see Section 4.6.1) and anxiety (see Section 4.6.1) are not positively affected by the use of the empathic agent.

4.5.1 Procedure

I asked the participants to attend the first session of an interview with the virtual counselor, which includes the AUDIT [BHBSM01] psychometric instrument, to assess the client's dependence to alcohol and frequency of drinking. The clients sat in front of a computer with a camera connected to it. I gave them oral instructions about the way the system works. They had access to a computer mouse and keyboard to select their answers to the counselor's questions from multiple choice menus. Users had the option to choose their preferred counselor's gender and ethnicity among the available characters (Hispanic, Caucasian, African American), some of which are shown in Figure 4.3. The default counselor was a Caucasian female (named Amy) who speaks in English. I have implemented three conditions for the experiment:

1. **Text-only** Drinker's Check-Up (DCU): during the session, the exact same content of the DCU [HSD05] is delivered to the user using text-only web pages.
2. **Empathic** counselor: during the counseling session, Amy reacts to the client with verbal and non-verbal empathic reactions. She expresses different emotional facial expressions (happy, sad, concerned, surprised, and neutral); head gesture (nod); large and subtle smile; head movement mimicry (pitch, yaw,

roll); eyebrow movement; mutual gaze; and lip synchronized verbal reflections. Being polite and getting permission for pursuing the interview is an empathic technique, so, at the beginning and end of the interview Amy requests for user's permission to continue.

An interview session begins with a verbal introduction of the system by Amy. Following the introduction, Amy asks for permission from the client to go to the next step. Then, she gives an overview of what will happen during the interview and asks for permission again to start the interview. During the introduction, Amy shows a neutral face and does not provide any empathic responses to the participant. The interview session involves a set of questions about the user's drinking behaviors. For each question, the client selects an answer from a list of 3-5 answers. At the end of the interview, Amy asks for permission to give a normative feedback about the user's drinking behavior in her/his age group and gender. During the interview (excluding the introduction), Amy empathizes with the client using the Empathy Model (described in Section 4.3.7). After the feedback, the user is directed automatically to an online questionnaire (see Section 4.5.2), which debriefs her/him about the performance of the virtual counselor.

3. **Non-empathic** (neutral) counselor: Amy shows a neutral facial expression during the introduction and interview, does not empathize with the user at all, and ignores the user's changes of emotional state. At the beginning and end of the interview Amy does not request for user's permission to continue.

After getting the approval from the Institutional Review Board (IRB), participants were recruited from volunteer university students (through fliers and emails) and Mechanical Turk workers. They were randomly assigned to each of the three experiment conditions. From the total number of 81 users (45 females with average age

of 24.6 years old and 36 males with average age of 26.3 years old), 26 were assigned to the empathic counselor, 25 to the neutral counselor, and 30 to the text-only version. From the local subjects (i.e., subjects interviewed in our lab), 32 of them were males and 19 of them were females. The ethnicity distribution of these participants was as 55% White, 27% Hispanic, 16% African American, and 2% Asian.

In the next section I describe the after-experiment questionnaire used to debrief the clients about the acceptance and performance of the counselor.

4.5.2 Questionnaire

I designed an online after-experiment questionnaire to evaluate the counselor's *empathy, anthropomorphism, animacy, likability, perceived intelligence, perceived safety, subjective performance, and user's acceptance*. It is based on a combination of the model presented by Heerink et al. [HKEW09] and the "Godspeed questionnaire" [BKC08].

Heerink's model evaluates the **users' acceptance** of assisting social artificial agents. This model involves different constructs, each of which is represented by multiple statements. Users reply to these statements on a 5-point Likert scale (-2 to +2). For positive statements (e.g., "I enjoyed the health counselor talking to me"), "-2" means "strongly disagree" and "+2" means "strongly agree". For negative statements (e.g., "I found the health counselor boring"), "-2" means "strongly agree" and "+2" means "strongly disagree". I use the following 10 constructs with the given definitions:

- Attitude (ATT): positive or negative feelings about the technology. The statements used to evaluate the attitude of the clients toward the virtual counselor are: (1) I think it's a good idea to use the counselor; and (2) The counselor would make my life more interesting.

- Intention to Use (ITU): outspoken intention to use the system over a longer period in time. I use the following statement to evaluate the clients' intention to use the system: (3) I think I'll use the system again.
- Perceived Enjoyment (PENJ): feelings of joy or pleasure associated by the user with the use of the system. The following statements are used in this category: (4) I enjoyed the counselor talking to me; (5) I enjoyed participating in this session with the counselor; (6) I found the counselor enjoyable; (7) I found the counselor fascinating; and (8) I found the counselor boring.
- Perceived Ease of Use (PEOU): degree to which the user believes using the system would be free of effort. I used five statements to evaluate the clients' perception about the system's ease of use: (9) I think I learned quickly how to use the health counselor; (10) I found the counselor easy to use; (11) I think I can use the counselor without any help; (12) I think I can use the counselor, if there is someone around to help me; and (13) I think I can use the counselor, if I have a good manual.
- Perceived Sociability (PS): perceived ability of the system to perform sociable behavior. The following statements are used in this category: (14) I consider the counselor a pleasant conversational partner; (15) I feel the counselor understands me; (16) I think the counselor is nice; and (17) I think the counselor is empathizing with me.
- Perceived Usefulness (PU): degree to which a person believes using the system would enhance his or her daily activities. The statements used for evaluating the perceived usefulness of the virtual counselor are: (18) I think the counselor is useful to me; and (19) I think the counselor can help me.
- Social Presence (SP): experience of sensing a social entity when interacting with the system. The four statements used in this category are: (20) When

interacting with the counselor, I felt like I'm talking to a real person; (21) I sometimes felt as if the counselor was really looking at me; (22) I can imagine the counselor to be a living creature; and (23) Sometimes the counselor seems to have real feelings.

- Trust (TRUST): belief that the system performs with personal integrity and reliability. I used the following statements to evaluate the clients' trust toward the virtual counselor: (24) I would trust the counselor, if it gave me advice; (25) I would follow the advice the counselor gives me; (26) I feel better interacting with the virtual counselor than with a human counselor in terms of privacy; and (27) I disclose more information about my drinking to the virtual counselor than a human counselor.
- Anxiety (ANX): evoking anxious or emotional reactions when using the system. The statements used in this category are: (28) I was afraid to make mistakes during the interview; (29) I was afraid to break something; (30) I found the counselor scary; and (31) I found the counselor intimidating.
- Social Influence (SI): user's perception of how people who are important to him think about him using the system. I used the following statements for evaluating the social influence of the virtual counselor: (32) It would give a good impression, if I should use the counselor later; and (33) I am comfortable to disclose information about my drinking to the counselor.

Bartneck [BKC08] have defined another questionnaire called "Godspeed" including five key concepts of HCI: **anthropomorphism**, **animacy**, **likability**, **perceived intelligence**, and **perceived safety** with the following definitions:

- Anthropomorphism (ANT): attribution of a human form, characteristics, or behavior to non-human concepts, such as robots, computers, and animals. In

this category, I asked the clients to rate the following statements for evaluating the anthropomorphism of the virtual counselor: (34) I rate the counselor as Fake/Natural; (35) I rate the counselor as Machine-like/Human-like; (36) I rate the counselor as Unconscious/Conscious; (37) I rate the counselor as Artificial/Lifelike; and (38) I rate the counselor's moves as Rigid/Elegant.

- Likability (LIKE): degree to which the agent evokes empathic or sympathetic feelings of the user. To evaluate the likability of the counselor, clients rated the following statements: (39) I rate my impression as Dislike/Like; (40) I rate the counselor as Unfriendly/Friendly; (41) I rate the counselor as Unkind/Kind; (42) I rate the counselor as Unpleasant/Pleasant; and (43) I rate the counselor as Awful/Nice.
- Animacy (ANIM): degree to which a computer agent is lifelike and can involve users emotionally. The animacy of the virtual counselor is evaluated by rating the following statements: (44) I rate the counselor as Dead/Alive; (45) I rate the counselor as Stagnant/Lively; (46) I rate the counselor as Mechanical/Organic; (47) I rate the counselor as Inert/Interactive; and (48) I rate the counselor as Apathetic/Responsive.
- Perceived Intelligence (PI): user's perception of the intelligence level of the agent. The statements rated in this category are: (49) I rate the counselor as Incompetent/Competent; (50) I rate the counselor as Ignorant/Knowledgeable; (51) I rate the counselor as Irresponsible/Responsible; (52) I rate the counselor as Unintelligent/Intelligent; and (53) I rate the counselor as Foolish/Moving Sensible.
- Perceived Safety (PSA): user's perception of the level of danger, and her/his level of comfort during the use. I evaluated the perceived safety of the virtual counselor using these statements: (54) During the interaction I was Anx-

ious/Relaxed; (55) During the interaction I was Agitated/Calm; and (56) During the interaction I was Quiescent/Surprised.

4.6 Results and Discussion

4.6.1 User Acceptance Results

I asked users to answer 56 questions, categorized in 15 classes. Clients answer each question in a 5-level Likert scale (-2 to +2). So, for each question, a 2×5 table is created which compares two of the experiment conditions (empathic vs. neutral, empathic vs. text, and neutral vs. text). The table rows are the experiment conditions, and the columns are the Likert scales (i.e., -2, -1, 0, +1, and +2). Clients' answers are analyzed using the Mantel-Haenszel-Chi-Square test (degree of freedom $df = 1$), which involves (1) assigning scores to the response levels, (2) forming means, and (3) examining location shifts of the means across the levels of the responses. The main difference between the regular Chi-Square and the Mantel-Haenszel test is that, in Chi-Square, clients' responses are compared with an expectation, while in Mantel-Haenszel test, there is no specific expectation and we compare the clients' responses in two conditions. More details of the Mantel-Haenszel test can be found in [MDK03].

I followed two null hypotheses: (1) text-only and avatar-based counselors have the same effects on the users; and (2) counselors with different levels of rapport abilities (empathic vs. neutral) have the same effects on the users. A common significance threshold value in the chi-square analysis is 5% (i.e., alpha). However, since I am performing three pairwise comparisons between the three different experimental conditions, to reduce the chance of false negative error (i.e, error type-I), I applied a Bonferroni correction on the alpha by dividing the alpha by 3 (i.e., $\alpha = \frac{5\%}{3} \approx 1.7\%$). Therefore, under the assumption of each null-hypothesis, a p value

of less than 0.017 rejects the null-hypothesis. Also, I compared the mean values of the same statements in the three experimental conditions to calculate the possible improvement/deterioration of them upon each other. The improvement/deterioration is calculated with the following formula:

$$\begin{aligned}
 \text{Improvement (or deterioration)} &= \frac{(\text{Mean}_1 - \text{Mean}_2)}{(\text{Likert Max Score} - \text{Likert Min Score})} \\
 &= \frac{(\text{Mean}_1 - \text{Mean}_2)}{2 - (-2)} \\
 &= \frac{(\text{Mean}_1 - \text{Mean}_2)}{4}
 \end{aligned}
 \tag{4.1}$$

Attitude (ATT)

Since interacting with an interface, which empathizes with the clients, is a new experience for the users, and provides a novel supportive way of interacting with the computer, I can expect that the clients show a more positive attitude to use the empathic counselor than the neutral and the text-only ones.

Results show significant differences in terms of attitude between the empathic and neutral conditions ($\chi^2 = 5.76, p = 0.016 < 0.017$); and between empathic and text-only conditions ($\chi^2 = 9.21, p = 0.002 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 0.081, p = 0.776 > 0.017$). These results indicate that a neutral avatar cannot improve the attitude to use a text-only counseling system. On the other hand, when an empathic avatar is used, significant differences appear. Therefore, the clients expect a human-like system to be empathic. This result confirms previous research by Nguyen and Masthoff [NM09].

The positive mean values of empathic ($mean = 0.78, stdev = 0.9$), neutral ($mean = 0.31, stdev = 1.05$), and text-only ($mean = 0.26, stdev = 0.86$) versions indicate that the clients have a positive attitude toward the system and found it a good idea to

use the virtual health counselor, regardless of the interface modality. However, the mean value comparison shows that the clients have 11.81% more positive attitude to use the empathic counselor than the neutral counselor and 13.06% more than the text-only version.

Intention to Use (ITU)

Results show significant differences in terms of intention to use between the empathic and neutral conditions ($\chi^2 = 6.41, p = 0.011 < 0.017$); and between the empathic and text-only conditions ($\chi^2 = 16.67, p \approx 0.000 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 4.60, p = 0.032 > 0.017$). These results support the previous result that the clients expect a human-like system to be empathic [NM09].

The positive mean values of empathic ($mean = 0.80, stdev = 0.89$) and neutral ($mean = 0.12, stdev = 0.89$) counselors show that the clients have positive intention to use the avatar-based counselors. This result confirms the results of a previous research [LYL⁺12], in which 74% of the clients reported a positive intention to use the avatar-based system. The negative mean value of text-only version ($mean = -0.45, stdev = 1.02$) indicates that the clients have negative intention to use the text-based system. The mean value comparison shows that the clients have 17.12% more intention to use the empathic counselor than the neutral counselor and 31.36% more than the text-only version. Also, they have 14.25% more intention to use the neutral counselor than the text-only one.

Perceived Enjoyment (PENJ)

Non-verbal mimicry increases rapport ([Laf79, LB76]), facilitates communication and may increase listeners' attention [LB76]. So, we can expect that the clients engage

more with the empathic counselor and find it more enjoyable than the neutral and the text-only ones.

Approving this hypothesis, results show significant differences in terms of perceived enjoyment between the empathic and neutral conditions ($\chi^2 = 24.40, p \approx 0.000 < 0.017$); and between the empathic and text-only conditions ($\chi^2 = 26.73, p \approx 0.000 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 0.013, p = 0.91 > 0.017$). Again, it shows that the clients expect a human-like system to be empathic.

The positive mean values of empathic (*mean* = 0.99, *stdev* = 0.63), neutral (*mean* = 0.31, *stdev* = 0.97), and text-only (*mean* = 0.39, *stdev* = 0.88) versions indicate that the clients perceived the system positively enjoyable, regardless of the interface modality. However, the mean value comparison shows that the clients enjoyed the empathic version 17.11% more than the neutral one, and 15.10% more than the text-only version. Therefore, the clients enjoy a text-only system more than a neutral human-like system.

Perceived Ease of Use (PEOU)

Results show no significant differences between any pairs of the experimental conditions: empathic and neutral conditions ($\chi^2 = 0.52, p = 0.471 > 0.017$); empathic and text-only conditions ($\chi^2 = 1.45, p = 0.228 > 0.017$); or neutral and text-only conditions ($\chi^2 = 0.07, p = 0.778 > 0.017$). This means that there is not enough statistical evidence to show that the different conditions have significant differences in terms of ease of use.

The positive mean values of empathic (*mean* = 0.84, *stdev* = 1.24), neutral (*mean* = 0.96, *stdev* = 1.27), and text-only (*mean* = 0.82, *stdev* = 1.24) versions indicate that the clients perceived all the version easy to use. However, the clients

prefer a character to help them during the interaction rather than a pure text-only intervention. It seems that enabling the character to build rapport with them complicates the use of the system. It is possible that users feel uneasy being watched or evaluated all the time with an intelligent ECA [CSX04]. Also, users feel that the counselor understands them (see Section 4.6.1), and they get the impression that a real person is talking to them, which may make it harder for them to use the system in presence of the counselor.

Perceived Sociability (PS)

Mimicking the facial expression and empathizing using the facial expressions of a speaker plays an important role in the perception of empathy [SbJS03]. So, we can expect that the empathic counselor reacts more appropriately to the clients' affective states and clients find it more understanding and empathizing than the neutral counselor and the text-only version.

Results show significant differences between all the three versions pairwise: empathic and neutral conditions ($\chi^2 = 36.57, p \approx 0.000 < 0.017$); empathic and text-only conditions ($\chi^2 = 17.58, p \approx 0.000 < 0.017$); neutral and text-only conditions ($\chi^2 = 6.22, p = 0.012 < 0.017$).

Statements in the *Perceived Sociability* category debrief the clients about the empathizing, understanding, and social abilities of the counselor. Therefore, the positive mean value of empathic counselor ($mean = 0.80, stdev = 0.87$) indicates that the clients perceived it empathizing, understanding, nice and sociable. On the other hand, negative mean value of the neutral version ($mean = -0.07, stdev = 0.97$) and small positive mean value of text-only version ($mean = 0.26, stdev = 0.98$) indicate that the clients perceived them respectively 21.68% and 13.56% less sociable than the empathic version.

The users perceived the empathic counselor a more pleasant conversational partner than the neutral one by 19.81% (statement 14). They reported that the empathic counselor understands the users 20.65% more than the neutral one (statement 15). The empathic counselor was rated 21.50% nicer than the neutral counselor (statement 16). Most importantly, the empathic counselor was perceived 24.77% more empathic than the neutral one (statement 17).

Perceived Usefulness (PU)

Results show significant differences between the empathic and neutral conditions ($\chi^2 = 10.13, p = 0.001 < 0.017$); and between the empathic and text-only conditions ($\chi^2 = 5.88, p = 0.015 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 1.36, p = 0.243 > 0.017$).

The positive mean values of empathic (*mean* = 0.68, *stdev* = 0.88), neutral (*mean* = 0.02, *stdev* = 1.08), and text-only (*mean* = 0.24, *stdev* = 0.97) versions indicate that the clients perceived the system positively useful regardless of the interface modality. However, the mean value comparison shows that the clients think that an empathic counselor is the most useful one (16.52% more than neutral and 10.94% more than text-only), but, if a counselor is not empathic it can be less useful than a pure text-only intervention system.

Social Presence (SP)

Non-verbal mirroring helps creating a smoother interpersonal interaction between partners [CB99], so, we can expect that the clients' engagement with the empathic system would be more than the neutral and the text-only ones.

Results show significant differences between the empathic and neutral conditions ($\chi^2 = 25.15, p \approx 0.000 < 0.017$); and between the empathic and text-only conditions

($\chi^2 = 46.20, p \approx 0.000 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 3.26, p = 0.071 > 0.017$). The not-significant difference between neutral and text-only versions and significant differences between the other two pairs support the same previous results.

The positive mean value of empathic ($mean = 0.21, stdev = 1.07$) indicates that the clients sense a social entity when interacting with the empathic counselor. But, negative mean values of neutral ($mean = -0.57, stdev = 0.99$), and text-only ($mean = -0.80, stdev = 0.93$) versions show that the clients do not have this sense when interacting with neutral and text-only versions. In terms of social presence, the empathic counselor makes 19.73% improvement over the neutral version and 25.14% improvement over the text-only version. On the one hand, negative mean values of the neutral version mean that the users did not feel that they are talking to a real person (statement 20), they did not imagine the counselor as a living creature (statement 22), and they did not feel that the counselor has real feelings (statement 23).

On the other hand, the positive mean values of the empathic version show that the users perceive the counselor as a real person who is looking at them and has real feelings. The mean value in statement 22 shows that although the empathic counselor is perceived more live than the neutral one, it is still not perceived as a living creature.

Trust (TRUST)

Since empathizing with the clients is known as a good way of building trust and receiving more information from the clients, we can expect that the clients can disclose more information to the empathic counselor than to the neutral one.

Results show significant differences between the empathic and neutral conditions ($\chi^2 = 13.01, p \approx 0.000 < 0.017$); and between the empathic and text-only conditions

($\chi^2 = 5.7, p = 0.0169 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 2.77, p = 0.096 > 0.017$).

Looking at the statements individually, the positive mean values in statement 24 indicate that the users would trust all the empathic ($mean = 0.88, stdev = 0.82$), neutral ($mean = 0.12, stdev = 0.93$), and text-only ($mean = 0.27, stdev = 0.99$) counselors, if they give them advice, however, they trust the empathic counselor 19.12% more than the neutral one and 15.18% more than the text-only version. Also, statement 25 shows that the users would follow the advice of the empathic ($mean = 0.68, stdev = 0.84$) counselor 6.42% more than the neutral ($mean = 0.42, stdev = 0.57$) one and 12.45% more than the text-only ($mean = 0.18, stdev = 0.97$) version.

In terms of privacy, users prefer to interact with a human counselor rather than a neutral ($mean = -0.38, stdev = 1.3$) virtual counselor (statement 26). But, they prefer to interact with an empathic ($mean = 0.56, stdev = 1.06$) counselor or a text-only system ($mean = 0.12, stdev = 1.15$) rather than a human counselor. Empathic counselor improved the neutral counselor by 23.62%, and improved the text-only version by 10.97%. As mean values show, users feel 12.65% more privacy when interacting with a pure text-only system than a neutral counselor.

Statement 27 shows that, in general, users believe that they can disclose more information to a human counselor than a virtual counselor delivered by a character. However, the empathic version ($mean = -0.13, stdev = 1.27$) has 7.45% improvement over the neutral ($mean = -0.42, stdev = 1.21$) one. More interestingly, the users believe that they can disclose more information about their drinking to a text-only system ($mean = 0.1, stdev = 1$) than a human.

Over all of the four statements in the *Trust* category, the empathic ($mean = 0.51, stdev = 1.07$) and text-only ($mean = 0.17, stdev = 1.03$) versions have positive mean values, and the neutral has a negative mean value ($mean = -0.07, stdev =$

1.01). Mean value comparisons show that the empathic counselor is 14.31% more trustful than the neutral one and 8.46% more than the text-only version.

Anxiety (ANX)

Affective virtual agents can support users through stressful tasks [GM04]. Therefore, my expectation is that the clients who use the empathic counselor feel less anxious than those who use the neutral version.

Results show no significant differences between any pairs of the experimental conditions: empathic and neutral conditions ($\chi^2 = 0.003, p = 0.954 > 0.017$); empathic and text-only conditions ($\chi^2 = 0.29, p = 0.591 > 0.017$); or neutral and text-only conditions ($\chi^2 = 0.32, p = 0.573 > 0.017$). This means that, there is not enough statistical evidence to show that the modality of delivering the intervention makes significant differences between the studied situations, in terms of privacy.

The positive mean values of empathic ($mean = 1.2, stdev = 0.87$), neutral ($mean = 1.19, stdev = 1.02$), and text-only ($mean = 1.26, stdev = 0.76$) versions indicate that none of the three counselor versions evoke anxiety while interacting with the clients and there are no significant improvements in the mean values. This means that the delivery modality (text-only vs. character-based) and the empathizing ability (empathic vs. neutral) did not reduce the **anxiety** level of the clients during the interaction, which does not support my expectation in the beginning of this section.

Social Influence (SI)

Results show no significant differences between any pairs of the experimental conditions: empathic and neutral conditions ($\chi^2 = 5.53, p = 0.018 > 0.017$); empathic and text-only conditions ($\chi^2 = 0.79, p = 0.373 > 0.017$); or neutral and text-only conditions ($\chi^2 = 2.81, p = 0.0935 > 0.017$). These results mean that there is not enough

statistical evidence to show that modality of delivering the intervention makes significant changes in the social influence of the system.

However, the positive mean values of empathic ($mean = 0.78, stdev = 1.03$), neutral ($mean = 0.27, stdev = 1.10$), and text-only ($mean = 0.61, stdev = 1.04$) versions show positive social influence on the clients regardless of the interaction modalities. The empathic counselor was reported to have 12.77% more social influence on the users than the neutral one and 4.35% more than the text-only version.

Figure 4.5 shows the mean value comparison of the three experimental conditions for the user acceptance features described above.

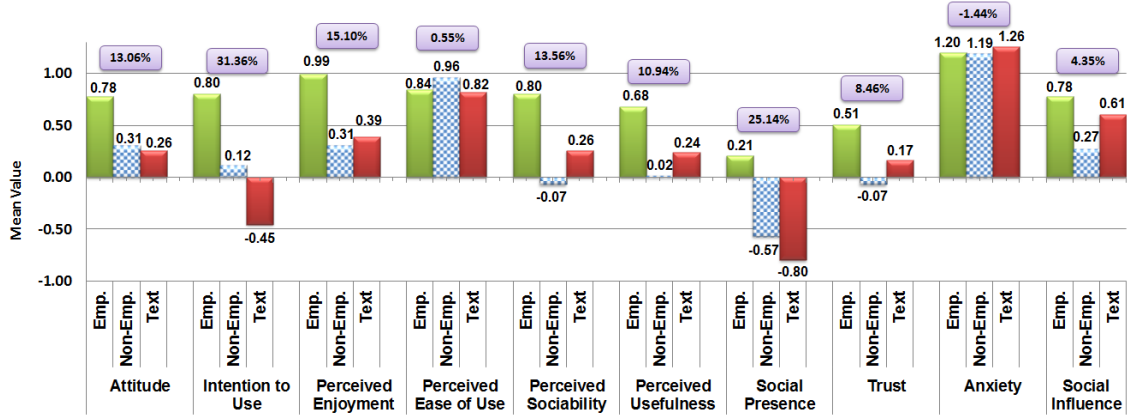


Figure 4.5: Mean value comparison of experimental conditions for user acceptance features. Percentages show the empathic character’s improvement over the text-only system.

4.6.2 Agent Evaluation Results

Anthropomorphism (ANT)

The visual channel facial expressions is deemed to be the most important in the human judgment of behavioral cues [AR92], because human observers seem to be mostly accurate in their judgment when looking at the face. This fact indicates that people

rely on displayed facial expressions to interpret someone's behavioral disposition. So, the empathic counselor is expected to be perceived more anthropomorphic and believable for the users than the neutral one.

Since no virtual character is used in the text-only version, I did not evaluate anthropomorphism for that version. However, I compared the empathic and neutral versions that include avatars. Results show that there are significant differences between the empathic and neutral counselors ($\chi^2 = 27.42, p \approx 0.000 < 0.017$) in terms of anthropomorphism.

The positive mean value of the empathic version ($mean = 0.28, stdev = 1.05$) indicates that the counselor was positively perceived anthropomorphic by the clients. On the other hand, the negative mean value of the neutral version ($mean = -0.47, stdev = 1.10$) indicates that the neutral version is perceived as not so anthropomorphic and it is perceived 18.73% less anthropomorphic than the empathic version.

Likability (LIKE)

Lakin et al. [LJC03] believe that mimicking the others' behavior causes feeling of closeness, liking, and smoother social interactions. So, we can expect that clients like the empathic counselor more than the neutral and text-only ones.

Results show significant differences between the empathic and neutral conditions ($\chi^2 = 21.51, p \approx 0.000 < 0.017$); and between the empathic and text-only conditions ($\chi^2 = 31.58, p \approx 0.000 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 0.93, p = 0.334 > 0.017$). This indicates that a neutral avatar does not affect the likability of the system but adding an empathic avatar affects the likability.

The positive mean values of empathic ($mean = 1.29, stdev = 0.64$), neutral ($mean = 0.85, stdev = 0.78$), and text-only ($mean = 0.76, stdev = 0.81$) versions

indicates that the clients liked all versions of the system. However, the empathic version is 10.85% more likable than the neutral and 13.11% more likable than the text-only version.

Animacy (ANIM)

Since no virtual character is used in the text-only version, I did not evaluate the animacy for that version. However, I compared the empathic and neutral versions, which include an avatar. Since the empathic counselor expresses different facial expressions and verbal reflections, it is expected to have a better animacy than the neutral one.

Results show that there are significant differences between the empathic and neutral counselors ($\chi^2 = 28.59, p \approx 0.000 < 0.017$). The positive mean value of the empathic version ($mean = 0.68, stdev = 0.98$) indicates that the counselor was perceived as well animated. On the other hand, the negative mean value of the neutral version ($mean = -0.11, stdev = 1.21$) indicates that the neutral version is not perceived so well animated and it is perceived 19.69% less animated than the empathic version.

Perceived Intelligence (PI)

Because the empathic feedbacks are provided based on the current most probable affective state of the client and her/his answers, the client may see the empathic counselor more intelligent than the neutral.

Results show significant differences between the empathic and neutral conditions ($\chi^2 = 18.76, p \approx 0.000 < 0.017$); and between the neutral and text-only conditions ($\chi^2 = 13.56, p \approx 0.000 < 0.017$); but no significant difference between the empathic and text-only conditions ($\chi^2 = 1.24, p = 0.266 > 0.017$).

The positive mean values of the empathic ($mean = 0.93, stdev = 0.74$), neutral ($mean = 0.42, stdev = 1.04$), and text-only ($mean = 0.82, stdev = 0.82$) versions indicates that the clients perceived all versions intelligent. However, comparison shows that, the empathic and text-only version are respectively 12.82% and 10.22% more intelligent than the neutral version. Therefore, adding a neutral avatar affects the perceived intelligence negatively, but an empathic avatar affects the perceived intelligence positively.

Perceived Safety (PSA)

Mimicry has been shown to influence the emotional state of an interaction partner positively [VbHKK04]. Also, affective virtual agents can increase client's abilities to recognize and regulate emotions and help motivating users [GM04]. So, we expect to see more positive emotions than negative ones during the interaction with the empathic counselor.

Results show significant differences between the empathic and neutral conditions ($\chi^2 = 11.44, p \approx 0.000 < 0.017$); and between the empathic and text-only conditions ($\chi^2 = 10.54, p = 0.001 < 0.017$); but no significant difference between the neutral and text-only conditions ($\chi^2 = 0.02, p = 0.895 > 0.017$). This indicates that a neutral avatar does not affect the level of perceived comfort/danger during the system use, but an empathic avatar does.

The positive mean values of empathic ($mean = 1.39, stdev = 0.95$), neutral ($mean = 0.79, stdev = 1.11$), and text-only ($mean = 0.82, stdev = 1.21$) versions indicate that the clients feel comfortable when using all versions of the system. However, the empathic version is perceived as 14.79% safer than the neutral one, and 14.21% safer than the text-only version.

Figure 4.6 shows the mean value comparison of the three experimental conditions for the character features described above.

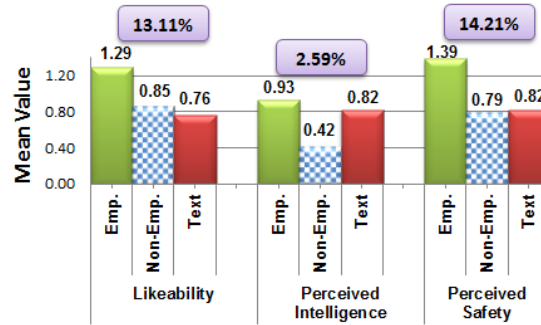


Figure 4.6: Mean value comparison of experimental conditions for the character features. Percentages show the empathic character’s improvement over the text-only system.

4.7 Summary

In this chapter I described the design, implementation, and evaluation of an empathic virtual character who can deliver an evidence-based Brief Motivational Intervention (BMI) for behavior change on excessive alcohol consumption - namely the Drinker’s Check-Up (DCU) [HSD05].

Although it may seem obvious that an empathic counselor is always perceived better than a neutral counselor, my results indicate that it is not the case in all aspects of the interaction, e.g., in my study, user’s anxiety to use the system was not improved by the ECA-delivery, nor was ease of use. I did not, however, test my approach with technophobic populations and it would be interesting to find out whether ECA research can specifically help such users reduce their anxiety while using technology.

Users’ overall acceptance of the system over a number of dimensions regarding the impact of the empathic communication of the character indicates that this novel

modality of delivery for behavior-change intervention could have a significant impact in terms of users' motivation to continue to use such systems. For example, users reported *30%* more *intention to use* the DCU intervention delivered by the ODVIC virtual character over the one delivered by the text-only system.

These results are very promising, particularly since it has been established that, although computer-based brief motivational behavior-change interventions can truly help people toward healthy lifestyles, too many people drop-out before benefiting. My approach may therefore lead to systems that decrease drop-out rates from behavior change interventions, which is a significant problem with, not only computer-based interventions, but also with face-to-face interventions [DCS⁺12, WP93].

Furthermore, because BMIs are adaptable and my system is modular, this approach can be adapted to target behaviors, such as overeating and lack of exercise, by adding the interventions to the database of psychometric instruments. We could therefore contribute to address several epidemic behavioral issues and promote healthy lifestyles for people in need.

The main result of this early study established that, indeed, a character's empathic and rapport-building abilities improve human-agent interaction. Results of the experiments depicted that even with a simple rule-based approach, we can affect the user acceptance and the character's perceived features positively.

CHAPTER 5

Modeling Rapport Using Machine Learning

As discussed in Chapter 4, I implemented a preliminary *Empathy Model*, which used a set of pre-defined rules in a **decision tree** to decide about the next counselor's empathic reaction to the client. As discussed before, the main goal of that preliminary implementation was to find out whether empathy and rapport, even with a simple rule-based approach, can positively affect the user acceptance and the perceived character features. The experiment results (see Section 4.4) clearly confirmed that, in an emotional context, such as behavior-change health counseling, empathic communication improves the user acceptance in terms of user's attitude, intention to use, perceived enjoyment, perceived sociability, perceived usefulness, social presence, trust, and social influence. Moreover, results showed that rapport improves the character's perceived anthropomorphism, animacy, likability, intelligence, and safety.

The major limitation of that preliminary rule-based model was that *social communication* and *psychology* expertise was needed to generate the rules. Especially, when the number of the input features (i.e., attributes) increases, the time complexity and the needed expertise are critical, because the number of the rules can increase exponentially as the number of the input features increases. For example, if we have 10 input features, each of which can take 2 values, up to 2^{10} combinations should be considered. On the other hand, if we want to generate rapport communication models for counselors that have specific features, such as specific ethnicities, cultures, and personalities, we need to have expertise in each field, too. For example, in order to generate a rule-based virtual rapport-enabled extrovert Chinese health counselor, we need an expert in Chinese culture who knows the extroversion personality as well.

Therefore, the approach that I present in this chapter generates models of non-verbal rapport communication using machine learning techniques from video and text corpora, in order to address these limitations of the rule-based approaches.

5.1 Overview

I consider modeling different non-verbal behaviors of the health counselor using machine learning, including head gestures (i.e., nod, shake, lateral sweep, and nod-shake), head movements (i.e., yaw, roll, pitch, and their combinations), eye gaze (i.e., left, right, up, and down), smile (i.e., neutral, subtle smile, and large smile), hand gestures (i.e., formless flick, pointing, contrast, iconic, closed, and opened), emotional facial expressions (i.e., neutral, happy, sad, surprised, angry/puzzled, afraid, and disgusted), eyebrow movement (i.e., up and down), and lean (i.e., forward, left, right, and back).

As shown in Figure 5.1, my approach for developing a rapport enabled virtual character using machine learning involves the following tasks: (1) providing an **annotation schema** for annotating the video and conversation transcript corpora (discussed in Section 5.3); (2) **annotating** the video and text corpora (discussed in Section 5.4); (3) **pre-processing the data** (discussed in Section 5.5); (4) **aligning** the data annotated manually and automatically (discussed in Section 5.6); (5) **selecting features** that are the most relevant ones for modeling each non-verbal behavior (discussed in Section 5.7); (6) **inducting the models**, in which a model is learned for each non-verbal behavior (discussed in Section 5.8); (7) **testing** the individual models (discussed in Section 5.10.1); (8) applying the models in **runtime** to the character as a compound non-verbal rapport model; and (9) performing **subjective tests** through user studies, in order to evaluate the perceived performance of the rapport-enabled character from users' point of view (discussed in Section 5.10.2).



Figure 5.1: The overview of the modeling phases.

This approach is close to the one taken in previous research by Lee et al. [LPNM09, LM09] in terms of the steps taken to model the non-verbal behaviors. However, it is different in the following aspects: (1) I use real-time interactive features, such as the client’s smile, emotional facial expressions, eyebrow movements, head movements, which are not used in previous research; (2) I model multiple non-verbal behaviors, whereas in previous research, only head nod and smile were modeled; (3) I model the non-verbal models of the counselor in both speaker and listener roles, while other research studies only cover either the speaker or the listener role; and (4) in addition to objective evaluation of individual models, I applied the models to a real application and evaluated their impact on the users subjectively in user studies, which is not performed in previous studies.

5.2 Data Collection

The **input** to the learning technique includes the data derived from the annotated video corpora and the data derived from the conversation transcript of Motivational Interviewing (MI) counseling sessions between real human clients and human counselors.

I video recorded four one-hour sessions of MI sessions. This comprises four hours of video from clients, and four hours of video from the counselor (the total of eight hours). These counseling sessions are delivered by Maya Boustani, a PhD student in the Clinical Science in child and adolescent psychology program at Florida International University (FIU). Maya is an expert in MI. These counseling sessions include

real face-to-face interactions of Maya with different human clients. The clients include four female FIU students.

I annotated 40 minutes of the videos, including 20 minutes from counselor and its corresponding 20 minutes of the client. For *each* of the non-verbal behaviors, the dataset includes 5,281 samples. In total, the number of times that each non-verbal behavior occurred in the dataset is reported in Table 5.1.

Table 5.1: Frequency of the speaker (i.e., counselor) and the listener (i.e., client) non-verbal behaviors in the dataset.

Gesture \ Role	Speaker	Listener	Total
Neutral head	2,378	1,181	3,559
Head nod	272	1,271	1,543
Head nod-shake	45	3	48
Head shake	106	9	115
Head lateral sweeps	12	4	16
Neutral hand	1,154	2,050	3,204
Hand formless flicks	759	51	810
Hand point	46	4	50
Hand contrast	220	20	240
Iconic hand	324	12	336
Closed hands	155	321	476
Opened hands	155	11	166
Forward head direction	5,281	0	5,281
Forward gaze	861	950	1,811
Left gaze	1,286	1,429	2,715
Right gaze	577	178	755
No smile	2,731	2,363	5,094
Subtle smile	70	91	161
Large smile	12	14	26
Neutral face	2,198	1,858	4,056
Happy face	110	186	296
Surprised face	351	149	500
Puzzled face	75	143	218
Afraid face	9	2	11
Disgusted face	70	130	200
Neutral brows	2,280	1,692	3,972
Down brows	513	768	1,281
Up brows	20	8	28
Neutral body lean	1,766	1,942	3,708
Forward lean	948	514	1462
Left lean	95	9	104
Right lean	5	2	7

For the gestures that have very few samples, such as different head movements, large smile, afraid faces, and right leans, there is not enough data to learn a model. Therefore, more data annotation is needed to collect more data for these gestures. However, for other gestures, there is enough data to learn models. In the next sections, I will explain the processes taken to generate these non-verbal behavior models.

5.3 Annotation Schema

I considered two main types of input features in modeling the non-verbal behaviors: video and text. Accordingly, I have multiple visual and textual features to be annotated. I designed an annotation schema including these two feature types. Many features are taken into consideration in the annotation schema and the annotation phase, however, in the feature selection phase (described in Section 5.7) the most relevant features to each non-verbal behavior are selected. In the next two sub-sections, I list all the features and their corresponding values.

5.3.1 Visual Features

The following visual features and values are used for annotation of the videos:

1. **Head gestures** of the counselor. Values of this feature include: *neutral*, head *nod* (AUM59), head *shake* (AUM60), head *nod-shake*, and lateral head *sweep*.
2. **Head movements** of the client and the counselor. Values of this feature include: head *yaw* (left or AU51, right or AU52), head *pitch* (up or AU53, down or AU54), head *roll* (roll-left or AU55, roll-right or AU56), and all 12 combinations of the above head AUs.

3. **Hand gestures** of the counselor. The values I used include: *neutral*, *formless flick*, *point*, *contrast*, *iconic* (represents some object or action), *opened*, and *closed* gestures.
4. **Eye gaze** of the client and the counselor. The values I used include: *forward*, *left* (AU61), *right* (AU62), *up* (AU63), and *down* (AU64).
5. **Smile** of the client and the counselor. These values are considered for this feature: *neutral*, *subtle smile* (AU12), and *open-mouth large smile* (AU12 + AU25 + AU26).
6. **Emotional facial expressions** of the client and the counselor. The values selected for this feature include the Ekman's standard emotions: *neutral*, *happy* (AU6 + AU12), *sad* (AU1 + AU4 + AU15), *angry/puzzled* (AU4 + AU5 + AU7 + AU23), *afraid* (AU1 + AU2 + AU4 + AU5 + AU20 + AU26), *surprised* (AU1 + AU2 + AU5 + AU26), and *disgusted* (AU9 + AU15 + AU16).
7. **Eyebrow movements** of the client and the counselor. The values for this feature include *neutral*, *up* (AU1 + AU2), and *down* (AU4 + AU42).
8. **Lean** of the counselor. I used five values for the lean feature: *neutral*, *lean forward*, *lean left*, *lean right*, and *lean back*.

5.3.2 Textual Features

In addition to the visual features listed above, I used different features of the surface text of the conversation. I used the following list of textual features (and values) for annotating the utterances of both client and counselor:

1. **Part of Speech (POS)** tags, which compose a linguistic category of words, generally defined by the syntactic behavior of the lexical words. I used the list

of the part of speech tags presented by the Stanford Natural Language Toolkit¹ as the possible values for this feature. A list of the POS values is presented in Table 5.2.

Table 5.2: Part of speech tags of the Stanford Natural Language Toolkit.

Tag	Description	Example	Tag	Description	Example
CC	Coordinating conjunction	and	PRP\$	Possessive pronoun	my
CD	Cardinal number	15, third	RB	Adverb	usually
DT	Determiner	the	RBR	Adverb, comparative	better
EX	Existential there is	there	RBS	Adverb, superlative	best
FW	Foreign word	d'hoevre	RP	Particle	give up
IN	Preposition or subordinating conjunction	on, in, of	SYM	Symbol	@, *
JJ	Adjective	green	TO	to	to
JJR	Adjective, comparative	greener	UH	Interjection	uhhuhh
JJS	Adjective, superlative	greenest	VB	Verb, base form	take
LS	List item marker	1)	VBD	Verb, past tense	took
MD	Modal	may, could	VBG	Verb, gerund or present participle	taking
NN	Noun, singular or mass	table	VBN	Verb, past participle	taken
NNS	Noun, plural	tables	VBP	Verb, non-3rd person singular present	take
NNP	Proper noun, singular	Alex	VBZ	Verb, 3rd person singular present	takes
NNPS	Proper noun, plural	Vikings	WDT	Wh-determiner	which
PDT	Predeterminer	both	WP	Wh-pronoun	who
POS	Possessive ending	's	WP\$	Possessive wh-pronoun	whose
PRP	Personal pronoun	I, he, it	WRB	Wh-adverb	where

2. **Dialog act**, which is a specialized utterance that has a performative function in language and communication. The list of the values I used for this feature includes: *greeting*, *question*, *interjection*, *negation*, *affirmation*, *assumption*, *obligation*, *contrast*, *inclusivity*, *intensification*, *response request*, and *word search*.
3. **Phrase boundaries** of the utterance. The values I selected for this feature include: *verb-phrase start*, *verb-phrase end*, *noun-phrase start*, *noun-phrase end*, *sentence start*, and *sentence end*.
4. **Sentence valence** of the utterance. Research shows that using the affective information of the sentences can help in modeling human non-verbal behaviors [LPNM09]. Especially, among word-level, phrase-level, and sentence-level analyses, the sentence-level affective information helps more in predicting the

¹<http://nlp.stanford.edu/software/>

non-verbal behaviors (such as head nod) [LPNM09]. Following these results, I used sentence-level valence of the text as another feature in my modeling process. I used three values for valence of the sentences: *negative*, *neutral*, and *positive*.

5. **New-word**, which is an indicator to find out if a new piece of information is being transferred between the interlocutors. Therefore, I take the “new word” (or “word newness”) as another feature in textual features, which shows whether the word being spoken is a new word or has been used before during the conversation. Accordingly, the possible values for this feature are *old* and *new*.

5.4 Data Annotation

Data annotation is a time-intensive job (order of a month for annotating 10 minutes of video) performed both manually by watching the videos and automatically using different recognizer software tools. For many of the features (explained next), I implemented automatic recognizers and annotators to make the annotation process faster, however, a few of the features still needed to be annotated manually.

In order to validate the automatic annotations, a human annotator re-annotated 25% of the automatic annotations randomly. In order to evaluate the reliability of the automatic annotations, I used the Cronbach α and measured the correlation between the automatic and manual annotations. For all features, the Cronbach α value was greater than 0.7, which indicates a high correlation between the automatic and manual annotations, and shows the reliability of the automatic annotations.

Although I was able to annotate many of the features automatically, at the time of performing this research, there was no available automatic tools to annotate the rest of the features (namely, head gesture, hand gesture, and lean). Therefore, I annotated these features manually. For that, I needed a tool that helps me align

the annotated features with the words in the surface text. I used the Anvil² [Kip01] annotation software, which allows us to (1) manually annotate videos, (2) align individual annotation features, and (3) export the annotations in well-known formats, such as Microsoft Excel, and Comma-Separated Values (CSV), and Tab-Separated Values (TSV). Below, I explain the annotation method I used for each individual feature.

5.4.1 Face, Head Movement, and Eye Gaze Recognizer

I implemented a face recognizer utilizing the InsightSDK³, which is a commercial face recognizer SDK from the SightCorp⁴ company. InsightSDK is a C++ SDK, which can recognize different facial expressions, head movements, and eye gaze directions in realtime. Figure 5.2 shows a snapshot of the implemented face recognizer.

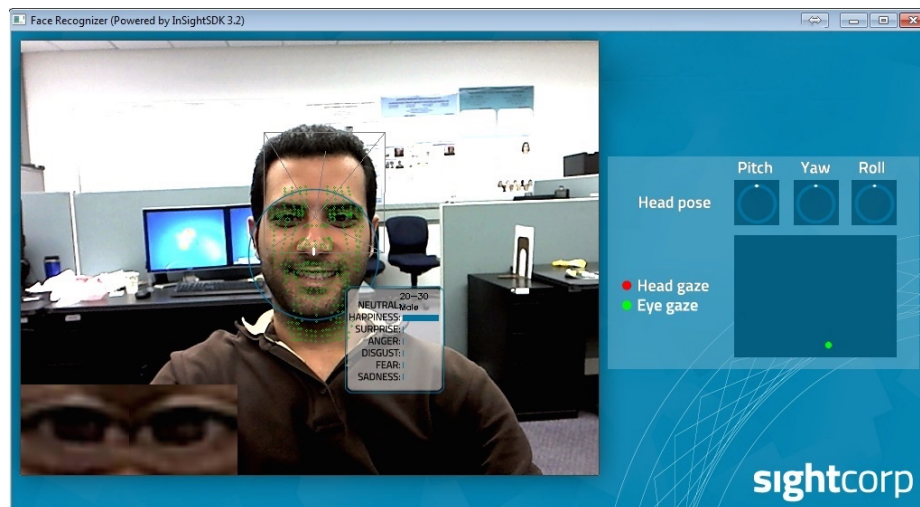


Figure 5.2: Snapshot of the implemented face recognizer.

²<http://www.anvil-software.org/>

³<http://sightcorp.com/insight/>

⁴<http://sightcorp.com/>

My face recognizer takes two types of inputs: (1) video files, and (2) video stream from camera. Also, it provides two types of outputs: (1) text file, and (2) stream of recognition results to another application through message passing. For the video annotation purpose, I passed the video files as the input to the face recognizer, and received the output as a text file including all the visual annotations. For the runtime face recognition phase (discussed in Section 5.9), I passed the camera video stream as the input and sent the recognition results to the non-verbal rapport modeling module. Figure 5.3 shows the face recognizer design.

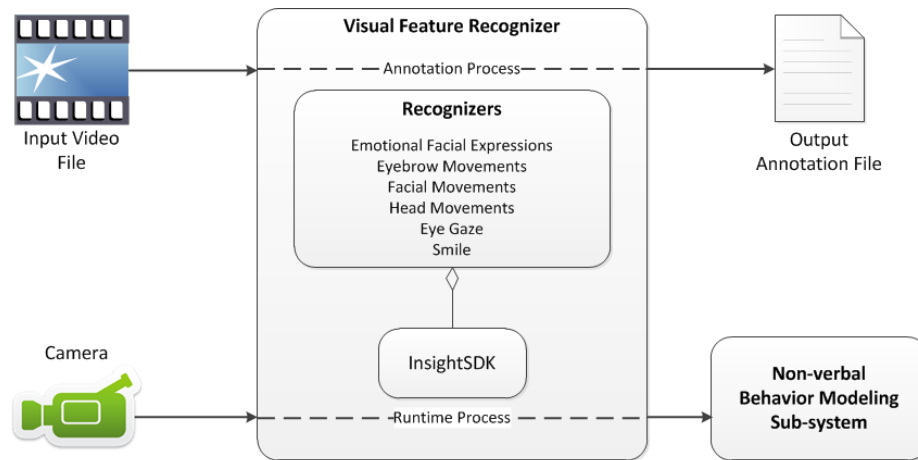


Figure 5.3: Face recognizer design.

The visual annotations that are performed using this automatic approach are head movements, emotional facial expressions, smile, eyebrow movements, and eye gaze. In addition, the face recognizer also returns other facial movements, namely: vertical upper/lower lip movements (involved in AU25 and AU26), vertical/horizontal mouth corner movements (involved in AU10, AU12, AU14, and AU15), vertical eyebrow movement (involved in AU1, AU2, AU4, and AU42), and vertical cheek movement (AU6).

SightCorp company reported the recognition accuracy of the InsightSDK as shown in Table 5.3. To get the best recognition accuracy, the best setup is reported as (1)

using 640×480 pixel resolution webcam (suggested Logitech HD Pro Webcam C910), and (2) user distance of approximately 60 cm from the camera. In both annotation and runtime phases, I zoomed in the videos to get an approximately 60 cm distance from the camera. Also, I used the same suggested webcam with the same resolution.

Table 5.3: Recognition accuracies reported by SightCorp for the InsightSDK.

Feature	Value	Accuracy
Facial expression	Neutral	88.2%
	Happy	95.2%
	Surprised	100%
	Angry (puzzled)	98.3%
	Disgusted	81.1%
	Afraid	94.2%
	Sad	95.6%
	Average	93.2%
Eye gaze	-	$\sim 2.1^\circ$
Head movement	Pitch	$5.2^\circ (\pm 4.6^\circ)$
	Yaw	$6.1^\circ (\pm 5.79^\circ)$
	Roll	$3.0^\circ (\pm 2.82^\circ)$

Head movements are reported as three features of head yaw, pitch, and roll. Yaw value is between -1 and 1, where -1 shows -40 degrees (left), and +1 shows +40 degrees (right) away from forward direction. Pitch value is between -1 and 1, where -1 indicates -30 degrees (up), and +1 indicates +30 degrees (down) away from forward. Also, roll value is between -1 and 1, where -1 shows -30 degrees (left), and +1 shows +30 degrees (right) away from forward direction. In order to take into account the recognition errors, and also consider a movement freedom range for the person to move, in all the head movement cases, I consider the values between -0.2 and $+0.2$ as the indicators of “forward” direction, which means I consider a 16 degree range for forward direction in head yaw recognition and 12 degree range in head pitch and roll recognition.

Emotional facial expressions are reported as seven features of neutral, happy, sad, surprised, angry, disgusted, and afraid. Each feature is represented by a float

number between 0 to 1. The larger values indicate higher probability of classifying the facial expression in a specific category. In order to enhance the recognition results of the emotional facial expressions, I also used a combination of the recognized facial movements (i.e., AUs). Table 5.4 shows the combinations used to recognize the emotional facial expressions.

Table 5.4: Feature comparisons used to recognize emotional facial expressions.

Features	Happy	Sad	Surprised	Angry	Afraid	Disgusted
Classified Emotion	happy	sad	surprised	Angry	Afraid	Disgusted
Vertical Upper Lip						> 0.1
Vertical Lower Lip					> 0.1	> -0.01
Ver. Left/Right Mouth Corner	> 0	< -0.1				
Hor. Left/Right Mouth Corner	> 0				> 0.1	
Vertical Cheek Movement	$0 \leq \dots \leq 1.8$					
Vertical Left/Right Eyebrow		< 0.1		< -0.1		

Eyebrow movements are reported as one feature with float values between -3 to $+3$. Negative values indicate eyebrow down movement, positive values indicate up movement, and zero indicated no eyebrow movement (i.e., neutral). I used a larger “neutral” range by considering values between -0.1 and 0.5 as neutral.

Smile is reported as horizontal/vertical different mouth movements, which can be combined to recognize open mouth smile (large smile) and subtle smile. Table 5.5 shows the combination of these features used to recognize large and subtle smile. As shown in Tables 5.4 and 5.5, the main difference between smile and happy is presence of AU6 in happiness.

Table 5.5: Feature comparisons used to recognize smile.

Features	Large Smile	Subtle Smile
Vertical Upper Lip	> 0.5	$0 \leq \dots < 0.5$
Vertical Lower Lip	> 0.1	$0 \leq \dots < 0.1$
Ver. Left/Right Mouth Corner	> 0	> 0
Hor. Left/Right Mouth Corner	> 0	> 0
Classified Emotion	happy	happy

Eye gaze is reported as a point in the 2D plane (origin is on top left corner of the screen). However, since eye gaze recognition using webcam is not as accurate as professional eye tracker devices, I was only able to recognize the eye gaze roughly. In other words, I was able to recognize if the person is gazing toward left, right, up, or down, which is good enough to recognize if the person is gazing toward or away from the other interlocutor. The eye gaze recognizer returns an (x, y) tuple, where for a screen of size 1920×1080 , values $0 \leq x \leq 1920$ indicate the left/right and $0 \leq y \leq 1080$ indicates the up/down gaze. I consider $900 < x < 1000$ and $500 < y < 600$ as gaze forward, $x > 1000$ as gaze left, $x < 900$ as gaze right, $y > 600$ as gaze down, and $y < 500$ as gaze up.

5.4.2 Part of Speech Tagger

I implemented a Part of Speech (POS) tagger utilizing the Stanford Natural Language Processing (NLP) POS tagger API [TKMS03]. This is an API that reads text in some language (English, Arabic, Chinese, French, and German) and assigns POS tags (e.g., noun, verb, adjective) to each word. The tagger's accuracy is reported as 97.24% on the Penn Treebank WSJ [MS94], and in terms of time, it is reported to tag 15000 words per second on an Intel Server in 2008. The dialog act tagger (see Section 5.4.3), phrase boundary tagger (see Section 5.4.4), and new word tagger (see Section 5.4.6) use the Stanford NLP parser and POS tagger to perform their functions, therefore, all of them will have the same accuracy. Stanford NLP software is originally implemented in Java, but also provided in other programming/scripting languages including C#, which I used in my implementation.

My POS tagger accepts two types of inputs in English: (1) a text file including the conversation transcript, and (2) a string including a single utterance. Accordingly, the POS tagger returns two types of outputs: (1) a text file including all POS tags of

the input text file, and (2) a string array including the POS tags of the input string. During the annotation process, I passed the complete utterance of each interlocutor as a file to the POS tagger and received the annotated utterance as a text file in the output. During the runtime process (see Section 5.9), I passed each utterance of the speaker and the listener to the POS tagger and received the tags as string arrays. Figure 5.4 shows the POS tagger design.

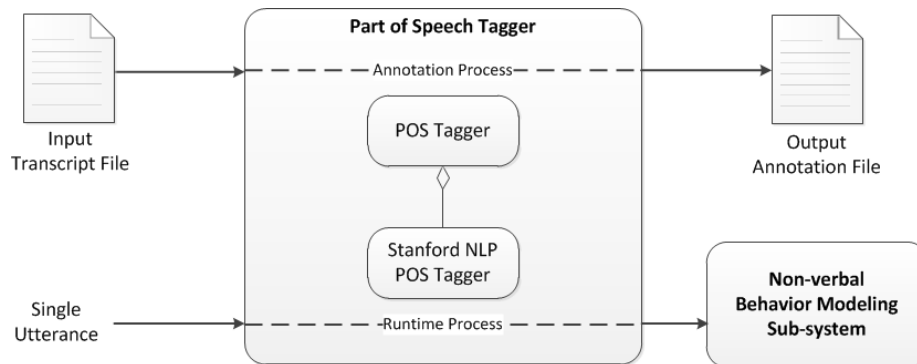


Figure 5.4: POS tagger design.

5.4.3 Dialog Act Tagger

I implemented a Dialog Act tagger using a dictionary of phrases and words that are most frequently used in different dialog acts and are mostly indicative of special dialog actions. The following list shows some examples of the words and phrase used for each of the dialog acts used in my tagger:

- Interjection: all right, of course, well, right, yes, yeah, no, and nope.
- Negation: nothing, cannot, can't, not, no, none, and nope.
- Affirmation: true, OK, yes, yeah, right, I am, he/she is, you/we are, all right, I/we have, he/she has, I/we do, and he/she does.
- Assumption: I guess, I suppose, I think, maybe, perhaps, could, probably, and assume.

- Obligation: have to, has to, need to, ought to, and should.
- Contrast: but, however, although, though, whereas, and while.
- Inclusivity: every, all, whole, several, plenty, and full.
- Intensification: really, very, quite, completely, wonderful, lot, great, absolutely, gorgeous, huge, fantastic, amazing, important, and much.
- Response request: you know.
- Word search: um, uh, well, mm, hmm, like, kind of, and I mean.
- Greeting: hi, hello, how are you, good morning, good afternoon, good evening, how do you do, what's up, how is it going, and how are you doing.
- Question: WH questions, and one of the following words in the beginning of the sentence: do, does, have, has, is, are.

First, I tokenize the input sentence and then, look for the above list of words/phrases in each sentence. If a word/phrase, which shows a special dialog act, is found in a sentence the sentence is tagged with that dialog act. A single sentence can be tagged with more than one dialog act. For example, “Hi, I am very happy today” is tagged with greeting, affirmation, and intensification.

Similar to the POS tagger, the dialog act tagger accepts two types of inputs: (1) a text file including the conversation transcript, and (2) a string including a single utterance. Accordingly, it returns two types of outputs: (1) a text file including the list of dialog acts included in each single utterance of the input text file, and (2) a list of the dialog acts included in the input string. During the annotation process, I used a text file as the input/output, and during the runtime process (see Section 5.9) I used a single utterance as input and received the list of its dialog acts. Figure 5.5 shows the dialog act tagger design.

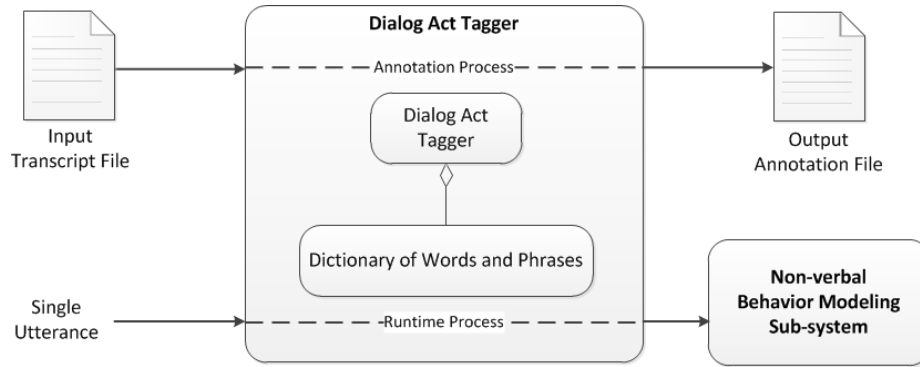


Figure 5.5: Dialog act tagger design.

In order to evaluate the *Dialog Act Tagger*, I asked 3 subjects to tag 130 sentences taken from AUDIT [BHBSM01], DrInC [MT95], SADQ [SMH83], BDP [MM84], and MAST psychometric instruments. Each sentence was tagged by all three subjects and the union of the subjects’ tags was used as the set of tags for each sentence. Then, I tagged the same sentences using the *Dialog Act Tagger*, and calculated the performance metrics for it. Results show an accuracy of 0.9581, precision of 0.7119, recall of 0.9225, and F1-measure of 0.8036.

5.4.4 Phrase Boundary Tagger

I implemented a phrase boundary tagger utilizing the API provided by the Stanford NLP. I used the natural language parser and the POS tagger included in the NLP package. Natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as “phrases”) and which words are the subject or object of a verb. Stanford NLP parser is a Java implementation of probabilistic natural language parsers (optimized PCFG, lexicalized dependency parsers, and lexicalized PCFG parser). I used a C# extension of the original software. As well as providing an English parser, the parser can be adapted

to work with other languages, such as Chinese, German, Arabic, Italian, Bulgarian, and Portuguese.

Based on previous studies [LPNM09, LM09], among other phrase boundaries in a sentence, noun phrase start/end, verb phrase start/end, and sentence start/end are the most effective ones in non-verbal behavior generation. Therefore, I used the same feature values for the *phrase boundary* feature, and accordingly, designed my phrase boundary tagger to tag them. First, I parse the input sentence and then, pass the parsed sentence to the POS tagger and create a POS tree for the sentence. Using the POS tree, I tag the phrase boundaries. If a word is not tagged with any of the above values, it is tagged with “None” tag.

My phrase boundary tagger accepts two types of inputs: (1) a text file including the conversation transcript, and (2) a string including a single utterance. Accordingly, it returns two types of outputs: (1) a text file including the list of phrase boundaries associated with each word of the input text file, and (2) a list of the phrase boundaries associated with each word of the input string. During the annotation process, I used the text file as the input/output, and during the runtime process (see Section 5.9) I used the single utterance as input and received the list of its phrase boundaries. Figure 5.6 shows the phrase boundary tagger design.

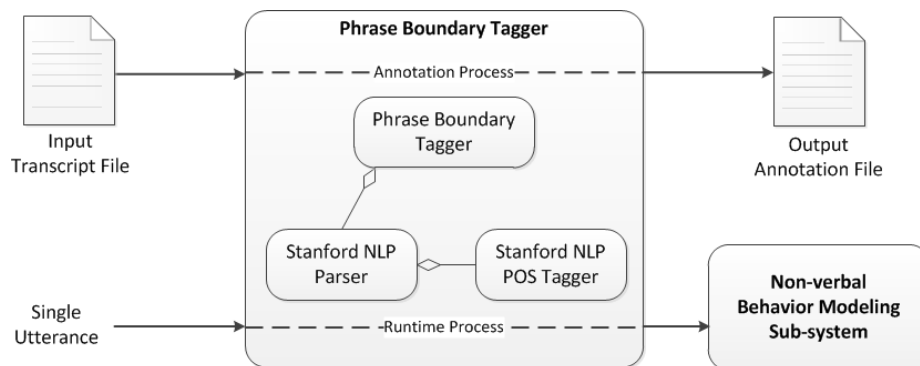


Figure 5.6: Phrase boundary tagger design.

5.4.5 Sentence Valence Tagger

I implemented a sentence valence tagger utilizing three sentiment analyses technologies: (1) SentiWordNet [BES10], (2) Stanford NLP Sentiment Analysis [SPW⁺13], and (3) Synesketchn⁵.

SentiWordNet is a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. SentiWordNet is the result of automatically annotating all WordNet [Mil95] synsets (i.e., set of synonym words) according to their degrees of positivity, negativity, and neutrality. It includes the total of 115000 words. In order to predict the valence of a sentence using the SentiWordNet, we can give positive points for positive words and negative points for negative words and then sum up these points. I consider valences greater than 0.05 as positive, less than -0.05 as negative and between -0.05 and 0.05 as neutral, which means that I consider a 5% threshold for valence tagging error.

Stanford NLP Sentiment Analysis uses a deep learning model, which builds up a representation of the whole sentence based on the sentence structure. Then, it computes the sentiment based on how words compose the meaning of longer phrases. This model is trained based on the Stanford Sentiment Treebank dataset [SPW⁺13]. Sentiment Treebank includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences. The Stanford NLP Sentiment Analysis accuracy for single sentence positive/negative classification is 85.4%. The accuracy of predicting fine-grained sentiment labels for all phrases is 80.7%. I used this software to classify each sentence into positive, negative or neutral classes.

Synesketchn is a free open-source software for textual emotion recognition and visualization. Synesketchn analyses the emotional content of sentences in terms of emotional types (namely happiness, sadness, anger, fear, disgust, and surprise), weights

⁵<http://synesketchn.krcadinac.com/>

(i.e., emotion intensity), and valence (i.e., neutral, positive, or negative). The recognition technique is grounded on a refined keyword spotting method, which employs a set of heuristic rules, a WordNet-based word lexicon, and a lexicon of emoticons (i.e., emotion icon) and common abbreviations. In my valence tagger, I just used the valence value returned by the Synesketch, which is a float number between -1 and +1. I consider valences greater than 0.05 as positive, values less than -0.05 as negative, and values between -0.05 and 0.05 as neutral.

Combining the above three valence recognition approaches, to classify the valence of a sentence, (1) it is passed to these three sentiment analyzers, (2) each of which assigns a positive, negative, or zero value to the sentence, and finally, (3) with averaging these points, the sentence valence is calculated. If the final average valence score is greater than 0.05, the sentence is labeled as positive; if it is less than -0.05, the sentence is labeled as negative; and otherwise, the sentence is labeled as neutral.

My sentence valence tagger accepts two types of inputs: (1) a text file including the conversation transcript, and (2) a string including a single utterance. Accordingly, it returns two types of outputs: (1) a text file including all sentences along with their valences, and (2) an integer label (i.e., -1, 0, or +1) indicative of the sentence valence for a single utterance input string. During the annotation process, I used the text file as the input/output, and during the runtime process (see Section 5.9) I used the single utterance as input and received its valence. Figure 5.7 shows the phrase boundary tagger design.

I evaluated the overall performance of the sentence valence detector on 130 sentences from AUDIT [BHBSM01], DrInC [MT95], SADQ [SMH83], BDP [MM84], and MAST psychometric instruments. I asked three human subjects to classify the valence of each question as negative, neutral, or positive. I used the maximum frequency of the human taggers' classifications as the label for each sentence. If all three human

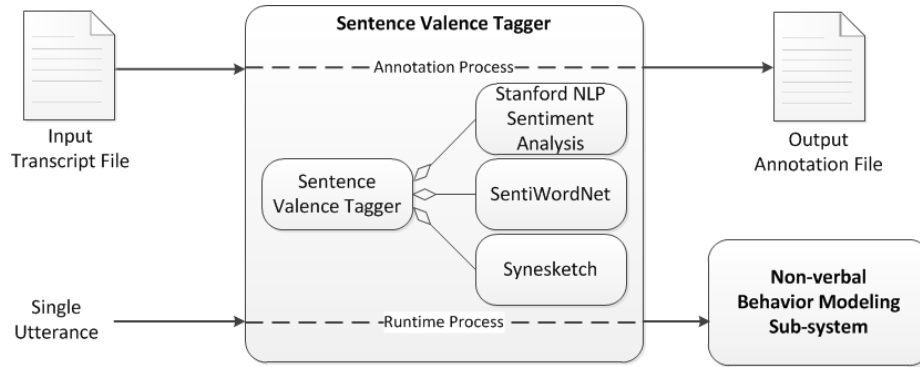


Figure 5.7: Sentence valence tagger design.

taggers classify a sentence differently, I label that sentence as neutral. Then, I tagged the same set of sentences using the *Sentence Valence Tagger*, in order to calculate the performance of the tagger. Results show an accuracy of 0.7846, precision of 0.6769, recall of 0.6769, and F1-measure of 0.6769.

5.4.6 New-Word Tagger

I implemented a new-word tagger in C# using dictionaries of words. The new-word tagger includes two sets of words, one for the client and one for the counselor. When a sentence is passed to the new-word tagger, it checks the corresponding list for every single word in the sentence, and tags them as *new* if they have not been used before by that person, otherwise words are tagged as *old*. Every time a word is checked against a dictionary, it is added to the dictionary for next look ups, if it is a new word.

Similar to other textual feature taggers, the new-word tagger also accepts two types of inputs: (1) a text file including the conversation transcript, and (2) a string including a single utterance. Also, it returns two types of outputs: (1) a text file including the list of the words and their tags, and (2) a list of the tags for the words in a string utterance. During the annotation process, I used the text file as

the input/output, and during the runtime process (see Section 5.9) I used the single utterance as input. Figure 5.8 shows the new-word tagger design.

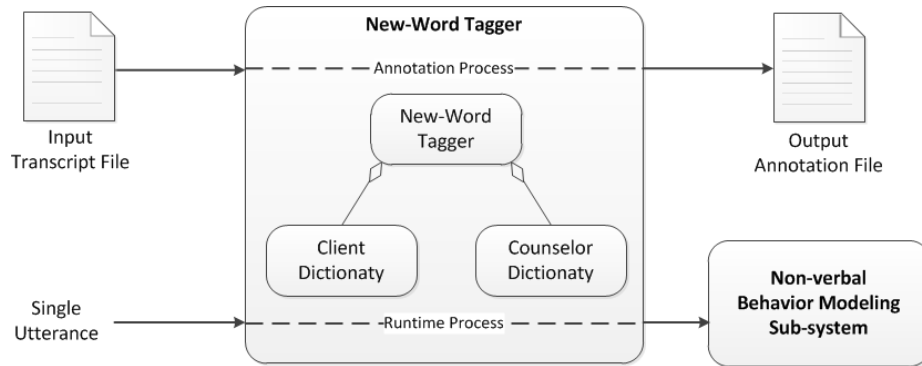


Figure 5.8: New-word tagger design.

5.5 Data Pre-Processing

The data annotations have different types including string, integer, and float. In order to use them in the HMM, I needed to change all the data types into integer. For features with string values, e.g., POS, I simply represented each string value with an integer value. For features with float values, e.g., emotional facial expressions, each float number is the probability of a category, therefore, I selected the highest probable category and represented that with an integer. For example, emotional facial expressions of neutral, happy, sad, ..., and disgust are represented by integer numbers 1, 2, 3, ..., and 7.

Afterwards, I put together every three consecutive words (and their corresponding feature vectors) and form a set of trigrams, which is used as my data set to the HMM. For each trigram, a target gesture is determined by the majority vote method [LPNM09], i.e., if 2 or 3 out of 3 words co-occur with the target gesture, the trigram is classified as an instance of the target gesture. For example, let's say we have three consecutive feature vectors of $\{\vec{a}, \vec{b}, \vec{c}\}$ with the output labels of

$\{nod, notNod, nod\}$. This trigram generates a data sample with input of $\{\vec{a}, \vec{b}, \vec{c}\}$ and output label of $\{nod\}$.

5.6 Data Alignment

I aligned each **word** in the transcript with the vector of visual and textual features that co-occur with that word. The annotated data set, which is used for the learning process, is a set of vectors, each of which co-occurs with one word in the transcript. For the manually annotated features, the alignment process is done by the Anvil annotation software. For the visual features that are annotated automatically, I matched and aligned the frame numbers reported by Anvil and the automatic annotator, in order to align the frames and consequently the words to their corresponding annotations. For the textual features that are annotated automatically, I matched and aligned the words reported by Anvil with the ones reported by the automatic annotator.

5.7 Feature Selection

For the particular kind of model I am training (i.e., Hidden Markov Models), adding another feature means I need more data samples to learn the combinations of all the features and how they affect the outcome I am trying to classify. With a limited number of data samples, I want to keep the number of features low by eliminating uncorrelated features (i.e., features that do not affect the target gestures).

I took a two-phase feature selection approach. In the first phase, for each target gesture, I reduced the number of features by counting the frequency of the gesture co-occurrence with each feature value, and selecting a subset of them that have the highest frequency (i.e., maximum frequency), as recommended in earlier related work

[LM09, LPNM09]. I called the resulted list of features as F_{mf} vector. It is important to know that some feature values may never appear in the data or may appear very few times, therefore, we can easily remove those feature values from that feature.

Tables 5.6 and 5.7 list the selected features for **speaker** and **listener** models respectively, as well as the counts of the selected features for each modeled gesture (after the maximum frequency feature selection phase).

In the second phase of the feature selection, which is called model selection (or cross validation), I took the selected list of features in the first phase (i.e., F_{mf}) as input, and used a 10-fold Cross Validation (CV) phase to select the best features out of them. For this purpose, I performed a step-wise backward elimination approach.

The data is extracted from recorded videos of human-human interactions. So, it is possible that in some intervals of the interaction, a specific gesture, say G , is expressed very rarely (or not expressed at all). Therefore, those intervals are not good sources for training and cross validation of the G model. In order to prevent this problem, I selected a different range of the data for training and cross validation. This approach is called cross validation over multiple splits of data, or multi-fold cross validation. Figure 5.9 shows the 10-fold cross validation approach.

The following algorithm shows the combination of step-wise backward elimination with the cross validation over multiple splits of data:

1. Randomly select 20% of the dataset and keep it unseen for model testing (discussed in Section 5.10.1);
2. From the remaining data, select 10 different splits of data (they have overlaps) for cross validation, each of which includes 20% of the complete dataset.
3. For each of the 10 data splits, keep the remaining 60% of the data for training (they have overlaps).

Table 5.6: Max. frequency features for speaker models (numbers show frequencies).

Feature	Model	Value	Head Nod	Head Nod-Shake	Head Shake	Subtle Smile	Gaze Left	Gaze Right	Happy	Surprised	Angry (Puzzled)	Disgust	Eyebrow Up	Eyebrow Down	Hand Formless-Flick	Hand Point	Hand Contrast	Hand Iconic	Hand Closed	Hand Opened	Lean Forward	Lean Left		
	Model																							
Total Occurrence	-	272	45	106	70	1286	577	110	351	75	70	20	513	759	46	220	324	155	155	948	95			
Counselor	New-Word	Old New	246	32	87	49	1069	463	78	307	63	59	15	419	628	41	162	256	102	115	682	84		
	POS	DT	-	-	-	-	95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
		IN	-	-	-	-	129	53	-	36	15	7	-	64	81	-	22	29	14	-	100	9	-	
		JJ	22	-	-	-	-	-	-	-	-	10	-	-	-	-	-	22	-	-	-	-	-	
		NN	-	9	-	-	137	50	-	30	-	-	-	-	-	-	-	-	16	21	81	-	-	
		NNS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19	-	-	-	-	
		PRP	-	-	11	15	123	-	20	46	12	-	-	-	57	92	12	-	33	-	-	102	9	
		PRP\$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	-	-	-	-	
		RB	19	-	16	-	111	-	10	-	-	-	-	-	43	73	-	30	27	-	19	94	13	
		UH	101	-	-	-	103	41	10	20	-	6	-	-	41	-	-	-	-	-	21	-	-	-
		VB	-	-	10	-	-	-	-	-	-	-	-	-	-	53	-	-	-	-	-	-	-	-
		VBP	-	-	-	-	-	43	-	-	-	-	-	-	-	59	-	-	25	-	15	-	-	-
	Dialog Act	Aff.	210	35	76	54	999	441	79	248	53	42	13	393	583	42	157	256	122	120	711	92		
		Assu.	-	-	-	-	113	47	-	-	-	-	-	-	-	-	-	30	51	-	-	-	-	
		Con.	34	-	-	-	128	87	13	54	-	9	-	48	102	-	23	49	-	-	-	-	43	
		Inc.	147	38	35	33	677	302	57	175	46	41	8	259	384	23	109	177	97	112	555	65		
		Inten.	-	-	-	-	166	62	-	-	-	16	-	59	128	-	-	35	-	-	-	-	-	
		Inter.	60	18	83	-	486	258	27	148	25	14	-	209	351	21	100	118	53	74	446	67		
		Neg.	37	14	83	-	457	242	26	142	25	10	-	202	332	21	97	103	47	73	426	66		
		RR	-	-	-	-	148	61	10	50	-	-	-	67	111	-	25	-	-	-	-	-	53	
		WS	45	15	-	15	563	232	26	164	16	20	5	191	379	-	67	147	42	67	368	53		
	Ques.	-	-	-	-	-	-	-	-	-	-	5	46	-	-	-	-	-	21	-	-	-		
	Phrase Bound.	SS	66	12	29	24	335	170	39	116	16	17	6	130	223	13	53	87	38	40	253	21		
		NPS	55	13	26	23	338	142	35	105	20	15	9	137	201	15	60	91	36	39	252	22		
		NPE	-	-	-	-	127	50	-	-	-	-	-	44	-	-	43	-	-	-	-	-		
		VPS	63	11	40	16	312	164	27	86	15	19	4	119	235	9	64	83	35	39	265	21		
		VPE	-	-	-	-	91	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
	Valence	Neg.	90	39	34	35	773	350	50	223	42	34	9	296	515	25	110	222	55	84	494	67		
		Neu.	140	-	10	23	215	87	34	53	12	17	-	91	75	7	42	46	55	-	-	-		
		Pos.	42	-	62	12	298	140	26	75	21	19	9	126	169	14	68	56	45	63	310	-		
	Head Gest.	Neu.	-	-	-	55	1066	487	87	289	58	56	16	415	705	33	172	297	112	134	815	90		
		Nod Shake	-	-	-	11	153	41	18	-	7	14	-	64	-	9	-	44	-	28	-	-		
Head Mov.	Fwd.	272	45	106	70	1286	577	110	351	75	70	20	513	759	46	220	324	155	155	948	95			
Hand Gest.	Neu.	183	-	26	32	533	237	47	108	40	37	16	265	-	-	-	-	-	-	-	32			
	Cont.	-	-	44	-	-	-	-	57	8	-	-	-	-	-	-	-	-	-	-	-			
	FF	37	-	11	10	368	155	12	100	14	14	5	121	-	-	-	-	-	-	-	43			
	Icon	-	-	-	-	128	83	-	52	-	-	-	-	-	-	-	-	-	-	-	-			
	Close Open	28	-	12	21	-	-	38	-	-	-	-	-	-	-	-	-	-	-	-	-			
Eye Gaze	Fwd	78	13	30	24	-	-	38	103	21	17	5	115	236	12	90	113	64	51	464	-			
	Left	153	13	47	32	-	-	50	156	33	37	10	302	368	29	88	128	68	72	-	58			
	Right	41	19	29	14	-	-	22	92	21	16	5	96	155	-	42	83	-	-	-	-			
Smile	Neu.	259	42	104	-	1250	561	28	-	75	70	20	488	748	46	216	321	129	151	904	94			
	Sub.	-	-	-	-	-	-	70	-	-	-	-	-	-	-	-	-	-	-	-	-			
	Big	-	-	-	-	-	-	12	-	-	-	-	-	-	-	-	-	-	-	-	-			
Facial Emotion	Neu.	211	24	72	-	1007	421	-	-	-	-	8	362	616	39	148	257	97	122	716	75			
	Hap.	18	-	-	70	-	-	-	-	-	-	-	-	-	-	-	-	38	-	-	-			
	Sur.	21	15	25	-	156	92	-	-	-	-	12	-	100	-	57	52	-	-	125	12			
Ang.	-	-	-	-	-	-	-	-	-	-	-	-	69	-	-	-	-	-	-	-	-			
Eyebrow Mov.	Neu.	205	35	83	-	974	476	80	303	-	56	-	-	633	36	192	276	119	142	837	70			
	Down	64	10	22	-	302	96	30	36	69	14	-	-	121	9	43	-	-	-	-	-			
Lean	Neu.	218	24	45	36	966	331	48	214	45	52	18	379	466	37	66	182	-	103	-	-			
	Fwd	49	21	61	33	259	225	61	125	23	18	-	109	250	9	150	138	122	43	-	-			
	Left	-	-	-	-	-	-	-	-	7	-	-	-	-	-	-	-	-	-	-	-			
Head Mov.	Fwd.	151	18	78	54	784	335	84	125	40	32	14	307	515	22	131	164	125	97	625	95			
	Roll-L	113	27	28	15	479	239	25	218	34	37	6	205	232	20	89	160	-	58	-	-			
	Fwd	220	30	76	46	1061	461	72	258	62	45	17	406	633	34	169	246	118	133	703	86			
Eye Gaze	Left	30	15	30	18	190	88	29	87	12	19	-	89	81	12	68	58	32	-	-	-			
	Right	22	-	-	-	-	28	-	-	-	-	-	-	-	-	-	-	-	-	-	-			
Smile	Neu.	272	45	106	70	1286	577	110	-	75	70	20	513	759	46	220	324	155	155	948	95			
Facial Emotion	Neu.	272	45	106	70	1286	577	110	351	75	70	20	513	759	46	220	324	155	155	948	95			
Eyebrow Mov.	Neu.	258	45	103	64	1226	552	104	343	67	70	19	493	727	45	212	323	151	141	923	88			
	Down	-	-	-	-	-	-	-	-	8	-	-	-	-	-	-	-	-	-	-	-			

Table 5.7: Max. frequency features for listener models (numbers show frequencies).

Feature \ Model		Value	Head Nod	Subtle Smile	Gaze Left	Gaze Right	Happy	Surprised	Angry (Puzzled)	Disgust	Eyebrow Down	Hand Closed	Lean Forward
		-	1271	91	1429	178	186	149	143	130	768	321	514
Client	Total Occurrence	-	1271	91	1429	178	186	149	143	130	768	321	514
	New-Word	Old	906	67	985	100	116	94	89	65	487	161	-
		New	236	-	-	-	35	-	-	35	163	112	-
	POS	DT	-	-	-	-	-	-	-	-	-	18	30
		IN	-	-	-	10	-	-	-	-	57	20	-
		JJ	-	-	-	-	-	-	-	-	-	18	-
		NN	-	-	-	-	-	-	-	-	60	25	35
		PRP	135	-	133	17	19	15	22	14	76	38	55
		RB	139	-	122	11	-	-	-	-	73	30	-
		UH	-	-	106	-	17	-	-	16	-	21	29
		VB	-	-	-	-	-	-	-	-	-	19	-
	VBP	-	-	-	-	-	-	-	-	48	28	39	
	Dialog Act	Aff.	873	64	843	82	125	87	88	64	437	195	271
		Assu.	-	-	170	-	-	-	-	-	86	-	-
		Inc.	872	68	1037	127	128	110	94	95	553	229	338
		Inten.	328	-	367	26	-	33	-	36	160	80	74
		Inter.	571	-	614	51	76	52	56	43	321	155	245
		Neg.	528	-	552	44	63	48	47	39	301	138	220
		WS	598	-	724	71	86	77	53	57	332	144	164
		Con.	-	-	355	22	42	33	29	-	200	-	-
Obl.	-	-	-	-	-	-	-	-	81	-	-		
Phrase Bound.	SS	279	22	278	33	38	31	35	27	142	67	102	
	NPS	300	27	318	36	45	33	41	28	167	76	112	
	NPE	-	-	-	-	-	-	-	-	60	-	-	
	VPS	304	18	315	29	36	36	30	30	166	80	122	
Valence	Neg.	663	42	607	62	79	55	82	62	332	215	252	
	Pos.	410	30	482	45	62	48	37	29	262	-	115	
Head Mov.	Fwd.	827	68	979	117	142	77	110	84	522	213	351	
	Roll-L	342	-	364	53	-	60	-	-	189	75	115	
Eye Gaze	Fwd	682	34	716	113	75	90	82	54	354	135	234	
	Left	341	37	441	40	76	-	36	57	251	130	207	
	Right	-	-	270	25	35	-	-	-	163	-	-	
Smile	Neu.	1271	91	1429	178	186	149	143	130	768	321	514	
Facial Emotion	Neu.	1271	91	1429	178	186	149	143	130	768	321	514	
Eyebrow Mov.	Neu.	1167	83	1312	160	165	140	83	122	703	297	476	
	Down	-	-	-	-	-	-	59	-	-	-	-	
Counselor	Head Mov.	Fwd.	1271	91	1429	178	186	149	143	130	768	321	514
	Hand Gest.	Neu.	1106	68	1285	153	133	120	107	88	630	-	254
		Close	159	-	-	-	49	-	36	41	124	-	234
	Head Gest.	Neu.	-	42	728	78	87	60	51	48	346	160	261
		Nod	-	49	695	97	99	87	91	82	420	159	249
	Eye Gaze	Fwd	479	33	-	-	75	46	76	67	266	222	355
		Left	695	50	-	-	96	76	60	55	468	92	135
	Smile	Neu.	1222	-	1372	169	81	149	143	130	726	293	466
		Sub.	-	-	-	-	91	-	-	-	-	-	-
	Facial Emotion	Neu.	910	91	1142	121	-	-	-	-	528	178	302
Hap.		-	-	-	-	-	-	-	-	-	-	76	
Sur.		-	-	-	27	-	-	-	-	-	-	-	
Ang.		-	-	-	-	-	-	-	-	110	36	63	
Eyebrow Mov.	Neu.	845	57	954	144	114	129	32	89	-	197	358	
	Down	420	34	468	34	72	-	110	41	-	124	156	
Lean	Neu.	1021	57	1289	151	110	120	80	86	612	87	-	
	Fwd	249	34	-	24	76	29	63	44	156	234	-	

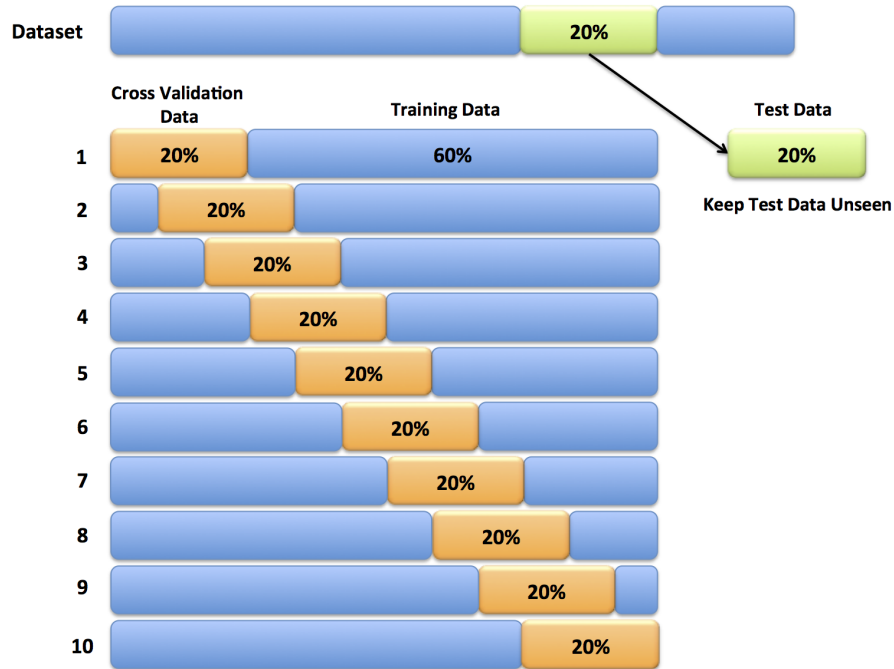


Figure 5.9: 10-fold cross validation.

4. For gesture G , start with its corresponding F_{mf} , and eliminate a single feature, f , from F_{mf} .
5. For each split of training and cross validation data:
 - (a) Learn a new model for G using the training data.
 - (b) Use the cross validation data set to evaluate the derived model.
 - (c) Save the accuracy, precision, recall, and F1-measure calculated using cross validation data.
6. Take average over all the 10 measurement sets.
7. If the average measurements are lower than the previous model in which feature f is not removed yet, it shows that the removed feature, f , is an important feature and removing that causes information loss. If removing f causes a very small amount of information loss comparing to other features, we can remove that feature.

8. Go to step 4 and repeat for other features in F_{mf} .
9. Select the model (i.e., set of features) with the lowest cross validation error (i.e., greatest average cross validation performance). Call this final feature set as F_{cv} .

Tables 5.8 and 5.9 show the final list of features selected for each of the modeled gestures for speaker and listener roles after cross validation (model selection) phase.

5.8 Model Induction

I use a “one-versus-all” approach for modeling each of the gesture classes, i.e., instead of multi-class classification, I performed a binary classification for each individual gesture class. A binary classification for gesture G classifies the input data into either class G or Not- G . This approach enables us to generate an individual model for each non-verbal behavior.

To determine whether a trigram should be classified as a target gesture G , I trained a Hidden Markov Model (HMM) [Rab89] for G classification. HMM is a statistical model that is used for learning patterns where a sequence of observations is given. In my application, the input is a sequence of feature vectors representing consecutive words. So, the sequential property of this problem led me to use HMMs to predict gestures.

The **input** to the modeling process is a vector of visual and textual features representing each spoken word (by client or counselor) during the session. The **output** of each gesture model is (1) a category, which represents presence/absence of the target gesture, and (2) the likelihood of the classification (i.e., classification correctness probability).

For each target gesture of the counselor, I trained two models, one as a **speaker** and one as a **listener**, because the non-verbal behaviors of a speaker and a listener

Table 5.8: Final set of features selected for speaker models.

Feature \ Model		Value	Counselor																			
			Head Nod	Head Nod-Shake	Head Shake	Subtle Smile	Gaze Left	Gaze Right	Happy	Surprised	Angry (Puzzled)	Disgust	Eyebrow Up	Eyebrow Down	Hand Formless-Flick	Hand Point	Hand Contrast	Hand Iconic	Hand Closed	Hand Opened	Lean Forward	Lean Left
Counselor	New-Word	Old New	-	-	✓	-	-	-	-	-	✓	✓	-	-	✓	-	-	✓	✓	✓	✓	-
	POS	IN JJ NN NNS PRP PRP\$ RB UH VB VBP	-	-	✓	-	-	-	-	-	✓	✓	-	-	✓	-	-	✓	✓	✓	✓	-
	Dialog Act	Aff. Assu. Con. Inc. Inten. Inter. Neg. RR WS Ques.	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	✓	✓	✓	✓	✓
	Phrase Bound.	SS NPS NPE VPS	✓	-	-	✓	✓	✓	-	-	✓	✓	-	✓	✓	-	-	✓	✓	✓	✓	-
	Valence	Neg. Neu. Pos.	-	✓	✓	✓	-	-	-	✓	✓	✓	-	✓	✓	-	-	✓	✓	✓	✓	✓
	Head Gest.	Neu. Nod Shake	-	-	-	✓	-	-	✓	-	✓	-	✓	-	✓	-	✓	-	✓	✓	✓	✓
	Hand Gest.	Neu. Cont. FF Icon Close	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	-	-	-	-	-	-	✓
	Eye Gaze	Fwd Left Right	✓	✓	✓	✓	-	-	✓	-	✓	-	-	-	✓	✓	-	-	-	-	✓	✓
	Smile	Neu. Sub. Big	-	✓	-	-	-	-	✓	-	✓	-	✓	-	-	-	-	-	✓	-	-	-
	Facial Emotion	Neu. Hap. Sur. Ang.	-	✓	-	✓	-	-	-	-	-	-	-	✓	✓	-	✓	-	✓	-	✓	-
	Eyebrow Mov.	Neu. Down	-	-	-	-	-	-	-	-	✓	✓	-	-	✓	-	-	-	-	✓	-	-
	Lean	Neu. Fwd Left	-	✓	-	✓	✓	✓	-	-	✓	✓	-	-	✓	✓	✓	✓	✓	✓	-	-
	Client	Head Mov.	Fwd.	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	✓	-	-	✓
		Eye Gaze	Fwd Left Right	✓	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓	✓	✓	✓	✓	✓	✓
		Eyebrow Mov.	Neu. Down	-	-	-	-	-	-	-	-	✓	✓	-	✓	✓	-	-	-	-	-	-

Table 5.9: Final set of features selected for listener models.

Feature \ Model		Value	Head Nod	Subtle Smile	Gaze Left	Gaze Right	Happy	Surprised	Angry (Puzzled)	Disgust	Eyebrow Down	Hand Closed	Lean Forward	
Client	New-Word	Old New	✓ ✓	✓ -	- -	- -	- -	- -	✓ -	- -	- -	- -	- -	
	POS	PRP	✓	-	-	-	✓	-	-	-	-	-	-	-
		RB	✓	-	-	-	-	-	-	-	-	-	-	-
		UH	-	-	-	-	✓	-	-	-	-	-	-	-
	Dialog Act	Aff.	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓
		Assu.	-	-	-	-	-	-	-	-	-	✓	-	-
		Inc.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Inten.	✓	-	✓	✓	✓	-	✓	✓	-	✓	✓	-
		Inter.	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Neg.	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		WS	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Con.	-	-	✓	✓	✓	✓	✓	✓	-	✓	-	-		
Phrase Bound.	SS	✓	-	-	-	-	-	-	-	-	✓	-	-	
	NPS	✓	-	✓	-	-	-	✓	-	-	✓	-	-	
	NPE	-	-	-	-	-	-	-	-	-	✓	-	-	
	VPS	✓	-	-	-	-	-	✓	-	-	✓	-	-	
Valence	Neg.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	Pos.	✓	✓	✓	✓	✓	✓	✓	-	-	✓	-	✓	
Head Mov.	Fwd.	✓	-	✓	✓	✓	✓	✓	✓	-	-	-	-	
	Roll-L	✓	-	-	-	-	-	-	-	-	-	-	-	
Eye Gaze	Fwd	✓	✓	✓	✓	✓	-	✓	✓	✓	-	-	-	
	Left	✓	✓	✓	✓	✓	-	✓	✓	-	-	-	-	
	Right	-	-	✓	✓	✓	-	-	-	-	-	-	-	
Smile	Neu.	-	-	✓	✓	✓	✓	✓	✓	-	✓	-		
Facial Emotion	Neu.	-	✓	✓	✓	✓	✓	✓	✓	✓	-	✓		
Eyebrow Mov.	Neu.	-	-	-	-	✓	✓	✓	-	-	-	-		
	Down	-	-	-	-	-	-	✓	-	-	-	-		
Counselor	Head Mov.	Fwd.	-	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	Hand Gest.	Neu.	-	✓	-	-	✓	-	✓	✓	✓	-	✓	
		Close	-	-	-	-	✓	-	✓	✓	✓	-	✓	
	Head Gest.	Neu.	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
		Nod	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
	Eye Gaze	Fwd	✓	✓	-	-	-	-	✓	✓	-	-	✓	
		Left	✓	✓	-	-	-	-	✓	✓	-	-	✓	
	Smile	Neu.	-	-	-	-	✓	✓	✓	✓	-	✓	-	
Sub.		-	-	-	-	✓	-	-	-	-	-	-		
Facial Emotion	Neu.	-	-	-	-	-	-	-	-	✓	✓	✓		
	Hap.	-	-	-	-	-	-	-	-	-	✓	✓		
	Ang.	-	-	-	-	-	-	-	-	✓	✓	✓		
Eyebrow Mov.	Neu.	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	-		
	Down	-	-	-	-	✓	-	✓	✓	-	✓	-		
Lean	Neu.	✓	✓	✓	✓	✓	✓	-	✓	-	✓	-		
	Fwd	✓	✓	-	-	✓	-	-	✓	-	✓	-		

are different. For example the head nod pattern (i.e., model) of a listener and a speaker are different. I used 60% of the total dataset for the training purpose.

When using HMM, we do not observe the actual sequence of states (i.e., the gesture). Rather, we can only observe some feature values happened at each state (i.e., visual and textual features). Formally, an HMM is a Markov model, for which there is a series of observations $x = \{x_1, x_2, \dots, x_T\}$ drawn from an input alphabet $V = \{v_1, v_2, \dots, v_{|V|}\}$, i.e., $x_t \in V, t = 1..T$. Also, there is a series of states $y = \{y_1, y_2, \dots, y_T\}$ drawn from a state alphabet $S = \{s_1, s_2, \dots, s_{|S|}\}$, i.e., $y_t \in S, t = 1..T$, but the values of the states are unobserved. The transition between states i and j is represented by the corresponding value in the state transition matrix A_{ij} , where $A \in \mathbb{R}^{(|S|+1) \times (|S|+1)}$. The value A_{ij} is the probability of transitioning from state i to state j at any time t .

The probability of an observation is modeled as a function of the hidden state. We make the *observation independence assumption* (i.e., current observation is statistically independent of the previous observations) and define:

$$P(x_t = v_k | x_1, \dots, x_T, y_1, \dots, y_T) = P(x_t = v_k | y_t = s_j) = B_{jk} \quad (5.1)$$

where matrix B encodes the probability of observing v_k given that the state at the corresponding time was s_j .

For example, you can think of the problem of modeling the “head nod” with a single feature (let’s say part of speech), shown in Figure 5.10. In this example,

$$S = \{nod, notNod\}, V = \{UH, NN, IN, PRP\}.$$

Assume that our training data includes a single sequence:

$$\begin{aligned} x &= \{UH, UH, PRP, VB, NN\} \\ y &= \{nod, nod, notNod, notNod, nod\}. \end{aligned}$$

We can combine the x and y vectors as a single input vector:

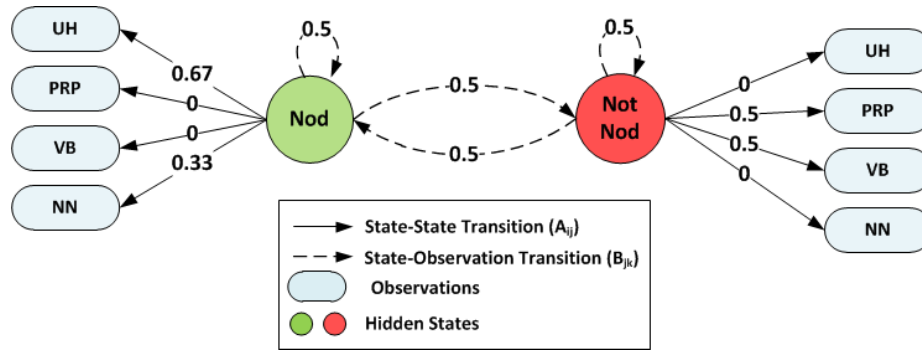


Figure 5.10: Sample Hidden Markov Model.

$$\{(UH, nod), (UH, nod), (PRP, notNod), (VB, notNod), (NN, nod)\}.$$

Given this data, the following information can be retrieved, which helps us calculating the A_{ij} and B_{jk} matrices.

- when we are in *nod* state, 50% of the times we transit to *nod* state (i.e., $A_{11} = 0.5$) and 50% of the times we transit to *not-nod* state (i.e., $A_{12} = 0.5$).
- when we are in *not-nod* state, 50% of the times we transit to *nod* state (i.e., $A_{21} = 0.5$) and 50% of the times we transit to *not-nod* state (i.e., $A_{22} = 0.5$).
- when we are in *nod* state, there is a 67% chance to observe a *UH* (i.e., $B_{11} = 0.67$), and 33% chance to observe a *NN* (i.e., $B_{14} = 0.33$).
- when we are in *not-nod* state, there is a 50% chance to observe a *PRP* (i.e., $B_{12} = 0.50$), and 50% chance to observe a *VB* (i.e., $B_{13} = 0.50$).

$$A = \begin{matrix} & \begin{matrix} 0 & nod & not - nod \end{matrix} \\ \begin{matrix} 0 \\ nod \\ not - nod \end{matrix} & \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix} \end{matrix}$$

$$B = \begin{matrix} & \begin{matrix} UH & PRP & VB & NN \end{matrix} \\ \begin{matrix} nod \\ not - nod \end{matrix} & \begin{pmatrix} 0.67 & 0 & 0 & 0.33 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix} \end{matrix}$$

There are three fundamental questions that can be asked from an HMM. *Evaluation* problem: what is the probability of an observed sequence, i.e., $P(\text{observation sequence}|\text{parameters})$; *Decoding* problem: what is the most likely series of states to generate the observations; and *Learning* problem: how the values for the HMM's parameters, A and B , can be learned given some data? In my application, first I solved the learning problem, in order to calculate the model parameters A and B , then for each sentence, I solved the decoding problem, in order to find the most probable sequence of states (i.e., gestures) for the input sequence of observations (i.e., visual and textual features). Next, I will explain these three HMM problems.

5.8.1 Probability of an Observed Sequence

In an HMM, the data is assumed to be generated by the following process: posit the existence of a series of states \vec{y} over the length of our time series. This state sequence is generated by a Markov model parametrized by a state transition matrix A . At each time step t , an observation x_t is selected as a function of the state y_t . Therefore, to get the probability of a sequence of observations, the likelihood of the data \vec{x} is added up given every possible series of states.

$$\begin{aligned}
 P(\vec{x}; A, B) &= \sum_{\vec{y}} P(\vec{x}, \vec{y}; A, B) \\
 &= \sum_{\vec{y}} P(\vec{x}|\vec{y}; A, B)P(\vec{y}; A, B)
 \end{aligned}
 \tag{5.2}$$

Although Equation 5.2 is true for any probability distribution, the HMM assumptions allow us to simplify the expression further:

$$\begin{aligned}
P(\vec{x}; A, B) &= \sum_{\vec{y}} P(\vec{x} | \vec{y}; A, B) P(\vec{y}; A, B) \\
&= \sum_{\vec{y}} \left(\prod_{t=1}^T P(x_t | y_t; B) \right) \left(\prod_{t=1}^T P(y_t | y_{t-1}; A) \right) \\
&= \sum_{\vec{y}} \left(\prod_{t=1}^T B_{y_t x_t} \right) \left(\prod_{t=1}^T A_{y_{t-1} y_t} \right)
\end{aligned} \tag{5.3}$$

The derivation in the second line of Equation 5.3 follows the HMM assumptions: the *output independence assumption* (i.e., current observation is statistically independent of the previous observations), *Markov assumption* (i.e., the next state depends only on the current state), and *stationary process assumption* (i.e., state transition probabilities are independent of the actual time, at which the transitions takes place). However, the sum is over every possible assignment to \vec{y} , because y_t can take one of $|S|$ possible values at each time step, evaluating this sum directly requires $O(|S|^T)$ operations.

A faster means of computing $P(\vec{x}; A, B)$ is via a dynamic programming algorithm called the Forward Procedure. For that, a quantity $\alpha_i(t) = P(x_1, x_2, \dots, x_t, y_t = s_i; A, B)$ is defined. Given that we are in state s_i at time t , $\alpha_i(t)$ represents the total probability of all the observations up through time t (by any state assignment). Given this quantity, the probability of the full set of observations $P(\vec{x})$ is represented as:

$$\begin{aligned}
P(\vec{x}; A, B) &= P(x_1, x_2, \dots, x_T; A, B) \\
&= \sum_{i=1}^{|S|} P(x_1, x_2, \dots, x_T, y_T = s_i; A, B) \\
&= \sum_{i=1}^{|S|} \alpha_i(T)
\end{aligned} \tag{5.4}$$

Algorithm 1 presents an efficient way to compute $\alpha_i(t)$. At each time step, only $O(|S|)$ operations are performed, resulting in a final algorithm complexity of $O(|S|.T)$ to

compute the total probability of an observed state sequence $P(\vec{x}; A, B)$. A similar algorithm known as the Backward Procedure can be used to compute an analogous probability $\beta_i(t) = P(x_T, x_{T-1}, \dots, x_{t+1}, y_t = s_i; A, B)$.

Algorithm 1 Forward Procedure for computing $\alpha_i(t)$.

Base case: $\alpha_i(0) = A_{0i}, i = 1 \dots |S|$

Recursion: $\alpha_j(t) = \sum_{i=1}^{|S|} \alpha_i(t-1) A_{ij} B_{jx_t}, j = 1 \dots |A|, t = 1 \dots T$

5.8.2 Maximum Likelihood State Assignment: Viterbi

In this problem, we ask the HMM for the most likely series of states $\vec{y} \in S^{|T|}$ given a series of observations $\vec{x} \in V^{|T|}$. Formally:

$$\arg \max_{\vec{y}} P(\vec{y} | \vec{x}; A, B) = \arg \max_{\vec{y}} \frac{P(\vec{x}, \vec{y}; A, B)}{\sum_{\vec{y}} P(\vec{x}, \vec{y}; A, B)} = \arg \max_{\vec{y}} P(\vec{x}, \vec{y}; A, B) \quad (5.5)$$

The first simplification follows from Bayes rule. The second simplification follows the observation that the denominator does not directly depend on \vec{y} . Naively, every possible assignment to \vec{y} can be tried and the one with the highest joint probability assigned by our model can be taken. However, this would require $O(|S|^T)$ operations just to enumerate the set of possible assignments. If the $\arg \max_{\vec{y}}$ is replaced with the $\sum_{\vec{y}}$, the current task is exactly analogous to the expression, which motivated the forward procedure. The *Viterbi Algorithm* is just like the forward procedure except that instead of tracking the total probability of generating the observations seen so far, only the maximum probability should be tracked and its corresponding state sequence should be recorded.

5.8.3 Parameter Learning for HMMs

The parameter learning problem is that, given a set of observations, what are the values of the state transition probabilities A and the output emission probabilities B that make the data most likely? Solving the learning problem for my dataset allows me to train the HMM before asking for the maximum likelihood state assignment of a candidate gesture. Since my training examples contain both the inputs and outputs of a process, supervised training can be performed by equating inputs to observations, and outputs to states.

For the supervised training, each training example is annotated with the correct classification. Two sets are defined: $\{y_1, \dots, y_N\}$ is the set of classes, which is equal to the HMM state set $\{s_1, \dots, s_N\}$; $\{x_1, \dots, x_M\}$ is the set of words, which is equal to the HMM observation set $\{v_1, \dots, v_M\}$. So, with this model the gesture modeling is framed as decoding the most probable hidden state sequence of classes given an observation sequence of words. To determine the model parameters, I used Maximum Likelihood Estimates (MLE) from a corpus containing sentences tagged with their correct gesture tags. For the transition matrix:

$$A_{ij} = P(y_i|y_j) = \frac{\text{Count}(y_i, y_j)}{\text{Count}(y_j)} \quad (5.6)$$

where $\text{Count}(y_i, y_j)$ is the number of times y_j followed y_i in the training data. For the observation matrix:

$$B_{jk} = P(x_k|y_j) = \frac{\text{Count}(x_k, y_j)}{\text{Count}(y_j)} \quad (5.7)$$

where $\text{Count}(x_k, y_j)$ is the number of times that when we observe the x_k , the gesture was classified as y_j in the training data. And lastly the initial probability distribution π_i is the probability of starting a sequence at state i , where s_1 is the starting state:

$$\pi_i = P(s_1 = y_i) = \frac{\text{Count}(s_1 = y_i)}{\text{Count}(s_1)} \quad (5.8)$$

I used the Accord.NET framework for implementing the HMMs. Accord.NET is a framework for scientific computing including statistical data processing, machine learning, pattern recognition, computer vision, computer audition, etc.

5.9 Runtime Operation

Since I would like to model the character's non-verbal behavior as both speaker and listener, I provided the ability for the client (i.e., user) to speak out his/her answers to the virtual counselor, which is a more natural way than typing. I used the Microsoft Speech Recognizer to recognize the clients' verbal answers. To increase the accuracy of the speech recognizer, I limited the recognition vocabulary to specific options for each question and asked the users to read their choice.

At the runtime, the textual feature recognizers recognize the part of speech, dialog acts, phrase boundaries, word newness, and valence of the dialog content of the client and counselor (i.e., virtual character). The textual feature recognizers are explained earlier in Sections 5.4.2 to 5.4.6. Also, the visual feature recognizer uses the camera (i.e., webcam) as its input and returns its classifications (i.e., emotional facial expressions, eyebrow movements, head movements, smile, and gaze) to the rapport model using message passing. The visual feature recognizer is explained before in Section 5.4.1.

For each sentence uttered by the speaker (virtual character or user) or listened by the character, all the above features are recognized and returned to the non-verbal behavior models (i.e., composite non-verbal rapport model). This set of feature values are used as the observations of the non-verbal behavior HMMs. The HMMs return the sequence of non-verbal behaviors to be expressed when this sentence is being uttered or listened by the character. The outputs of the non-verbal models are passed to the HapGest module, which resolves the possible conflicts between the gestures

(using priorities) and synchronizes them with the character’s verbal utterance. In order to generate the facial expressions, HapGest uses the hypertexts provided by the HapFACS (described in Section 3.1). Figures 5.11 and 5.12 depict the runtime process for speaker and listener roles.

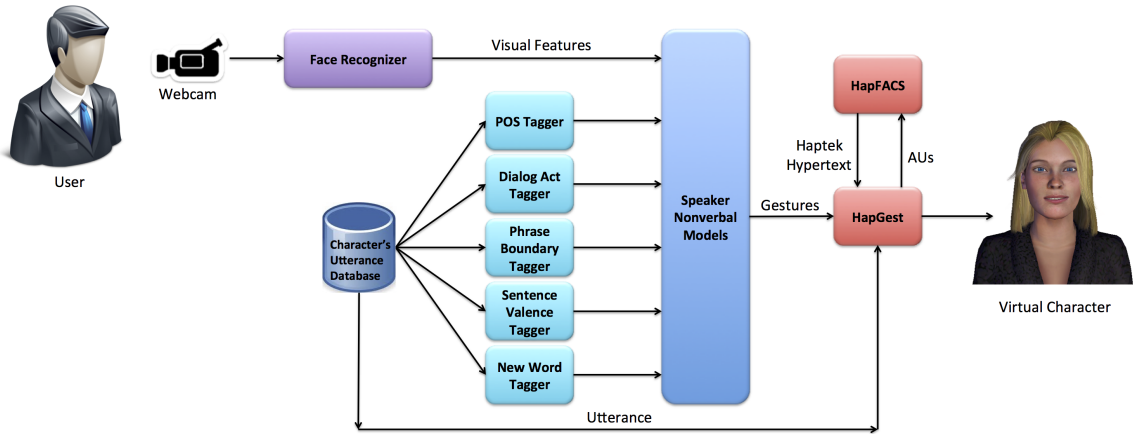


Figure 5.11: Runtime process for the virtual character’s speaker role.

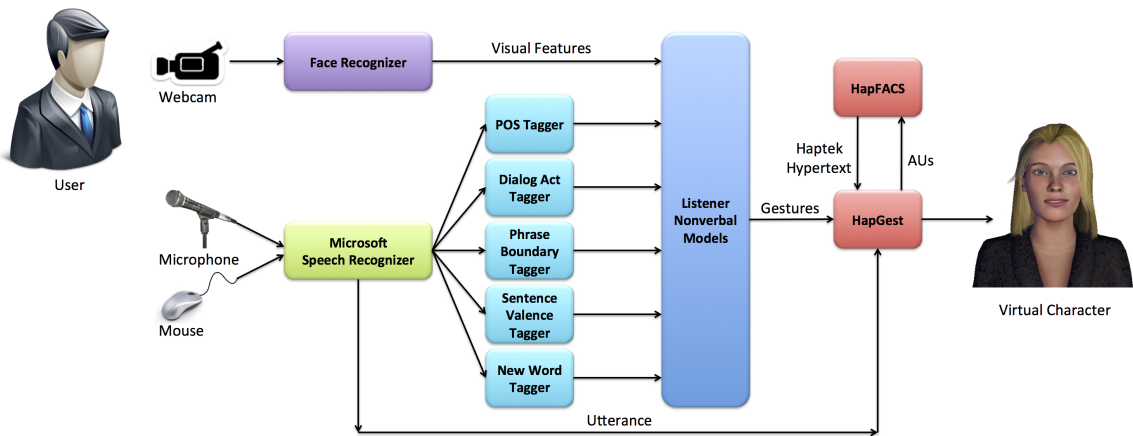


Figure 5.12: Runtime process for the virtual character’s listener role.

In the speaker role, utterances come from the database. Utterances are passed to the textual feature recognizers, and the latest visual features of the user are also perceived using the camera and the visual feature recognizer. All these feature values are passed to the speaker inducted models and their outputs are passed to the HapGest module.

In the listener role, responses are perceived from the user using microphone (or mouse, if the speech recognizer cannot recognize the user's voice). Then similar to the speaker role, these inputs are passed to the feature recognizers. Feature recognizers send their outputs to the listener inducted models, which decide about the best gestures and sent their decision to the HapGest to animate the character.

In an interactive application, time is a major concern in the runtime operation phase. The processing time of the system should be fast enough so the users do not feel large delays in the interaction. In my application, modules that are involved in the runtime process are (1) the textual feature recognizers, (2) the facial/head expression recognizer, (3) the classification, (4) synchronizing the verbal and non-verbal modalities by HapGest, and (5) the animation of the character. The fore-mentioned natural language processing toolkits are able to recognize the features from the text in the order of hundreds of milliseconds. The Stanford NLP tags more than 15000 words per second. For example, the Stanford NLP returns all the syntactical tags of the sentence "How often do you have a drink containing alcohol?" in 0.087 seconds. Also, the InsightSDK facial recognizer is able to recognize the facial expressions in realtime. The modeling process, which takes a few seconds, is performed offline before being used for classification. Therefore, the classification process is performed in realtime too. Synchronizing the verbal and non-verbal modalities can sometime take a few hundreds of milliseconds if the sentence is too long, because for each word in the sentence, HapGest may need to handle some events. However, this delay is not large enough to be recognized by the users. Finally, the Haptik virtual character system is able to animate the characters in realtime, i.e., the character can animate the gestures in realtime by receiving the appropriate hypertexts from the HapGest. Therefore, the overall runtime process, from recognition to animation, takes a time in

order of hundreds of milliseconds, which does not disturb the natural flow of speech and can generate natural animations.

5.9.1 Modeling Non-verbal Rapport Communication

Since the non-verbal behaviors are modeled based on the video corpus of the interactions of rapport-building counselors with clients, I used the **combination** of the learned non-verbal models to model the counselor’s **Rapport** communication as speaker and listener.

As mentioned in Section 2.1, based on Tickle-Degnen and Rosenthal [TDR90], the three essential components of rapport are *mutual attentiveness*, *positivity*, and *coordination*. In my rapport model, I (1) modeled the mutual *attentiveness* and *coordination* using the hand gestures, body lean, head gestures, and eye gaze models; and (2) modeled the *positivity* using the head nod, smile, emotional facial expressions, and eyebrow movement models.

My Non-verbal Rapport Communication Model is similar to the Rapport Agent 2.0 [HMG11] in the sense that (1) both of them are modeling the rapport as a three-component paradigm based on the Tickle-Degnen and Rosenthal [TDR90] theory; (2) both of them are modeling the attentiveness and coordination using the backchannel models; and (3) both of them are modeling the positivity using smile and head nods. However, there are multiple differences between my rapport model and the Rapport Agent 2.0, including (1) in the Rapport Agent 2.0 only smile and head nod are modeled, whereas in addition to these gestures, I also modeled other head gestures, hand gestures, eyebrow movements, eye gaze, emotional facial expressions, and body lean. Modeling these non-verbal behaviors improves the rapport communication and naturalness of the character; (2) in the Rapport Agent 2.0, the dialog content is not used for modeling the rapport (silence, head nod, eye gaze and smile are used), whereas in

my method I used the dialog content. This allows us to improve the attentiveness and coordination components; and (3) in the Rapport Agent 2.0, the dataset used for the modeling process is derived from videos of actors who role-play in a story telling scenario with a virtual character, which is not necessarily an emotional context, whereas I used videos of real human-human Motivational Interviewing counseling sessions, in which more emotional dialogs are exchanged.

5.10 Evaluation and Validation - Hypotheses Testing

I evaluated the individual non-verbal models and the overall rapport model in two phases: (1) objective evaluation of the non-verbal models, and (2) subjective evaluation of the character naturalness and perceived rapport.

5.10.1 Objective Evaluation of the Non-verbal Models

In order to evaluate the machine learning based approach, I measured the performance of each individual learned model using 20% of my annotated dataset, called the test dataset, which was kept unseen during the feature selection and learning phases. I applied the test data to the learned models and calculated the **accuracy** (i.e., ratio of the gestures correctly expressed), **precision** (i.e., ratio between the number of the gestures expressed correctly and the total number of the expressed gestures), **recall** (i.e., ratio between the number of the gestures expressed correctly and the number of the gestures in the actual data), and **F1-measure** (i.e., weighted harmonic mean of precision and recall) of the learned model. Equations 3.1 to 3.4, in Section 3.1, provide the mathematical formulas for calculating the above measurements.

Objective Evaluation Results and Discussion

Tables 5.10 and 5.11 respectively show the evaluation results of the objective measures for speaker and listener models.

Table 5.10: Objective evaluation results of the speaker models.

Measure	Accuracy	Precision	Recall	F1-Measure
Model				
Head Nod	0.703	0.750	0.871	0.803
Head Shake	0.982	0.997	0.984	0.991
Head Nod-Shake	0.890	0.992	0.895	0.939
Subtle Smile	0.768	0.991	0.765	0.851
Gaze Left	0.611	0.603	0.374	0.437
Gaze Right	0.773	0.860	0.882	0.860
Happy	0.876	0.981	0.881	0.925
Surprised	0.759	0.914	0.811	0.855
Angry (Puzzled)	0.836	0.975	0.849	0.904
Disgust	0.885	0.959	0.915	0.934
Eyebrow Up	0.893	0.995	0.897	0.942
Eyebrow Down	0.750	0.787	0.921	0.847
Hand Formless-Flick	0.743	0.905	0.771	0.830
Hand Point	0.904	0.993	0.909	0.948
Hand Contrast	0.827	0.966	0.842	0.895
Hand Iconic	0.771	0.951	0.794	0.864
Hand Closed	0.814	0.934	0.830	0.879
Hand Opened	0.806	0.975	0.822	0.888
Lean Forward	0.619	0.733	0.739	0.692
Lean Left	0.902	1.000	0.902	0.948
Average	0.8056	0.9131	0.8327	0.8616

As stated in Section 2.4.2, there are very few research studies in which machine learning is used to model the non-verbal behaviors of a human. Also, they modeled very few non-verbal behaviors. For example, Lee et al. [LM09] modeled head nods of a human speaker with accuracy of 0.8528, precision of 0.8249, recall of 0.8957, and F1-measure of 0.8588. Lee et al. [LPNM09] expanded their head nod model later by using affective information during the learning process, to improve the prediction metrics compared to accuracy of 0.8957, precision of 0.8909, recall of 0.9018, and

Table 5.11: Objective evaluation results of the listener models.

Measure Model	Accuracy	Precision	Recall	F1-Measure
Head Nod	0.759	0.885	0.747	0.808
Subtle Smile	0.846	0.973	0.862	0.912
Gaze Left	0.559	0.533	0.454	0.473
Gaze Right	0.746	0.854	0.848	0.845
Happy	0.827	0.985	0.826	0.891
Surprised	0.813	0.912	0.880	0.894
Angry (Puzzled)	0.738	0.976	0.743	0.841
Disgust	0.763	0.967	0.776	0.852
Eyebrow Down	0.719	0.795	0.844	0.816
Hand Closed	0.845	0.928	0.879	0.902
Lean Forward	0.672	0.767	0.717	0.713
Average	0.7515	0.8694	0.7825	0.8131

F1-measure of 0.8963. My speaker head nod model metrics are comparable to the model presented by Lee et al. [LM09, LPNM09].

In an other study [Kip06], which uses both hand crafted rules and machine learning to generate the gestures (facial expression, gaze, and head movement), they reported the maximum models' cumulative evaluation metrics as 0.338 precision and 0.321 recall. which are much lower than the evaluation metrics reported in my study.

5.10.2 Subjective Evaluation of the Character

In order to evaluate the perceived rapport of the character by the users and the perceived **naturalness** of the character, I applied the models to a virtual health counselor similar to the one implemented in Chapter 4. I replaced the rule-based empathy model, shown in Figure 4.2, with the new data-driven rapport model. The new system architecture and interface are depicted in Figures 5.13 and 5.14. Also, the class diagram of the implemented system is provided in Figure 5.15. I used user studies to compare the user acceptance (e.g., perceived rapport, believability,

likability, enjoyability, usefulness, engagement) and character features (e.g., perceived intelligence) of (1) the rapport-building counselor, and (2) a neutral counselor (i.e., a virtual counselor with neutral facial expressions and no non-verbal gestures). I hypothesize that the rapport-building virtual counselor is better accepted by the users and its character features are perceived better than a neutral one.

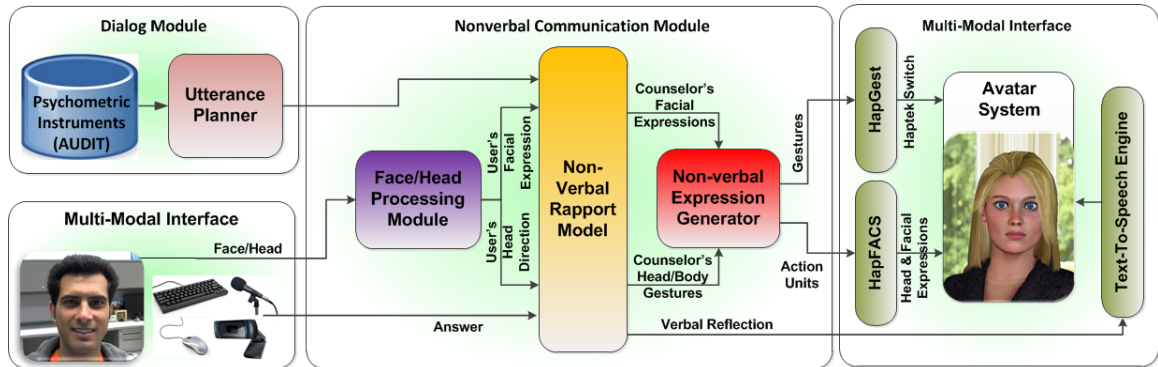


Figure 5.13: New architecture of the ODVIC.

According to Tickle-Degnen and Rosenthal [TDR90], non-verbal behaviors are measured in two ways: **molecular** and **molar**. Molecular measures are calculated internally during the interaction of the client with the system and are appropriate for measuring the user engagement, attention, and positivity components of rapport. The molecular measures consist of *counts/durations* of specific behaviors, such as head nodding and eye contact. Molar measures are defined in terms of the psychological impression, gestalt image, or perceived function they create, such as negative/positive facial expressions, social presence, helpfulness, distraction, and naturalness [BH02]. The molar measures are appropriate for measuring the coordination component of rapport. The molar measures are measured using both internal calculations (e.g., clients' facial expressions) and after-experiment questionnaires (e.g., asking the clients to provide feedback about the naturalness of the character's non-verbal behaviors).

I video recorded all of the interactions between the clients and the virtual counselor. After the experiment, two human coders (one FACS-certified and one non-



Figure 5.14: Snapshot of the virtual counselor.

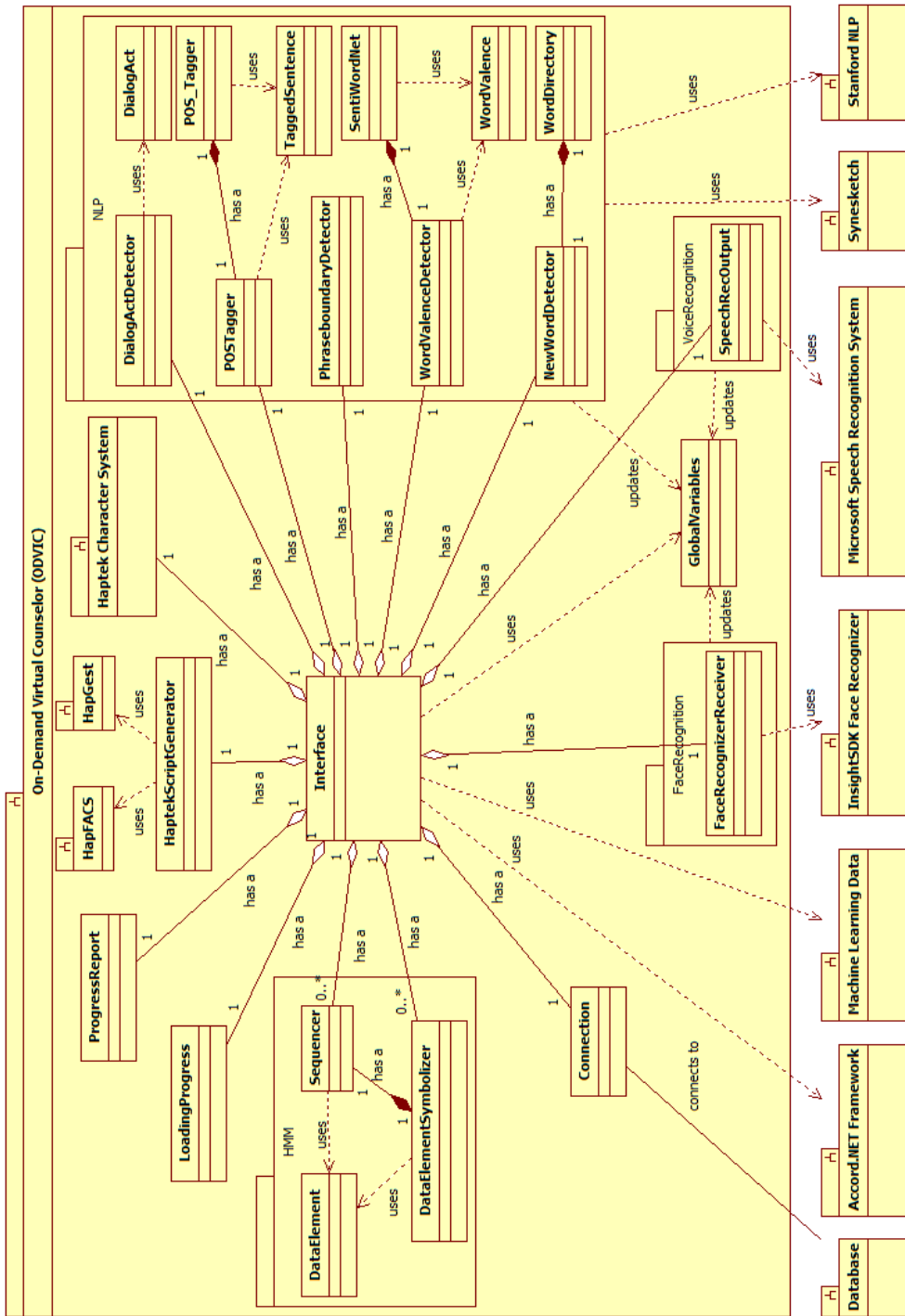


Figure 5.15: Class diagram of the ODVIC.

FACS-certified) manually watched the videos and reported the duration of positive facial expressions (happy and surprised) and negative facial expressions (angry, afraid, and disgusted). I compared them for the two experiment conditions as a measure of the perceived rapport by the users.

In addition, I used a collection of validated questionnaires as follow, in order to debrief the clients about their interaction experience with the virtual counselor:

1. Heerink's [HKEW09] questionnaire for evaluating the user acceptance. Items in this questionnaire are provided in Section 4.5.2.
2. Bartneck's [BKC08] Godspeed questionnaire for evaluating character features. Items for this questionnaire are provided in Section 4.5.2.
3. Virtual Experience Test [CGL10], which measures the virtual environment experiences based upon different dimensions of experiential design: sensory, cognitive, and affective:
 - (a) Sensory (perception of sensory input (visual, aural, haptic, etc): (1) "I found the virtual character to be of high quality," (2) "I found the character's voice to be of high quality."
 - (b) Cognitive (mental engagement with an experience): (1) "I found that the content in the interaction was helpful in informing me of my current drinking behavior."
 - (c) Affective (refers to the user's emotional state, and the degree to which a person's emotions are similar to those in real-world situations): (1) "I had emotional reactions while interacting with the counselor," (2) "I think if I was talking to a real counselor, I would experience the same emotional reactions," (3) "I felt that the character conveyed emotions."

4. Question provided by Kipp et al. [KKNG06] for evaluating user acceptance:
“The character managed to get my attention”.
5. Social presence assessment [BBBL01], which evaluates rapport and naturalness:
 - (a) “I perceived that I am in the presence of another person in the room with me.”
 - (b) I felt that the character was watching me and was aware of my presence.”
 - (c) “I perceived the character as being only a computerized image, not as a real person.”
6. Rapport scale presented in [HMG10a, KWG09, GWGF07], which evaluates different dimensions of rapport:
 - (a) Positivity and Close Connection: “I felt a close connection between me and the agent?”
 - (b) Mutual attentiveness: “The agent appeared to be interested in listening to me?”
 - (c) Perceived Rapport: “I felt rapport between the agent and myself?”
 - (d) Coordination: “I think that the character and I understood each other.”
 - (e) Perceived (1) Precision: “How often do you think the agent used inappropriate gestures (e.g., head, hand, smile, facial expressions, brows, gaze, body lean)?” and (2) Recall: “How often do you think the agent missed gesture opportunities?”
 - (f) Naturalness: “I think the virtual agent’s behavior was natural?”
7. Scales presented by Ruttkay and Pelachaud for evaluating perceived character features:

- (a) Fidelity/realism: perception of how lifelike or real the ECA and its capabilities appear.
 - (b) Expressiveness: the diversity and intensity of expressions, such as facial expressions, gesture, emotions.
 - (c) Personality: is the ECA dominant or humble?
 - (d) Coordination of multiple modalities.
8. Engagement evaluation tool suggested by Webster and Ho [WH97]:
- (a) Challenge: “The virtual counselor encouraged me to think about my drinking.”
 - (b) Attention focus: “The character kept me absorbed in the interaction.”
 - (c) Intrinsic interest: “The character presentation was interesting.”
 - (d) Overall: “The character presentation was engaging.”
9. Information disclosure (ID): user’s intention to disclose personal information:
- (a) “I would feel better interacting with the virtual counselor than a human counselor in terms of revealing personal information.”
 - (b) “I was comfortable to disclose information about my drinking.”
 - (c) “I would disclose more information about my drinking to the virtual counselor than a human.”
10. A specific question suggested in [HMG11] to evaluate the overall naturalness: “I think the virtual agent’s overall behavior was natural.”

In order to be able to measure and compare the users’ information disclosure, I used an additional option to the possible answers of each single question during the

interaction as “I prefer not to answer”. I counted the number of times that users select this response and compare them in the two evaluation conditions, as a measure of information disclosure. Therefore, lower number of selecting this option shows more intention to disclose information to the character.

Experiment Setup

The total of 56 subjects were recruited from the college students through fliers. Subject included 39 males with an average age of 25.5 years old and 17 females with an average age of 26.5 years old. Subjects included 21% White, 11% Black, 45% Hispanic, 14% Asian, 5% Caucasian, and 4% Indian ethnicities. The goal of this study was more focused on evaluating the non-verbal models in terms of perceived rapport, naturalness, and engagement. Since my goal is not to evaluate the effectiveness of the system on real problem-drinkers at this stage of the research, I did not perform clinical studies with real problem-drinkers.

Subjects were randomly assigned to one of the neutral or the rapport-enabled characters (27 subjects to neutral character and 29 subjects to rapport-enabled one). I guided the subjects to interact with the virtual character, which was applied to a virtual counseling framework. In this experiment, the virtual character acted as a virtual health counselor and steps through a series of assessment questions one by one. The human subjects were interviewed by the virtual counselor. For the interaction, I used a combination of the AUDIT assessment instrument provided in Section 4.3.4 and 5 other questions from the Drinker’s Inventory of Consequences (DrInC) [MT95] instrument (which assesses the negative consequences of drinking). The order of virtual character’s questions were predefined based on the assessment instrument.

Before the experiment starts, I provided the system introductions for the subjects and told them that the virtual counselor would ask them several questions about their drinking behaviors, and their task is to answer as best as they can. I asked the subjects to answer each question verbally by choosing the best choice from the provided answers (on the screen). I told the subjects that the virtual counselor listens to their answers, however, if the subjects ask any questions from the virtual counselor, it does not provide any answers to their questions. Clients sit in front of the monitor, from which the virtual counselor interacts with them. A camera is mounted on top of the monitor for both recognizing the visual features and video recording the interaction. After the interaction, subjects were redirected to a website, where the after-experiment questionnaire was implemented. Subjects were asked to assess the virtual character's performance and their experience with the character by answering the questions.

I recorded the users' facial expressions during the interaction using the same webcam used for facial expression recognition. Two subjects, one FACS-certified coder and one non-FACS coder, were asked to report the amount of time that users expressed positive and negative facial expressions. This time is considered as a measure of users' perceived rapport.

Subjective Evaluation Results and Discussion

The total of 40 questions (provided above in Section 5.10.2) were asked from the users. Subjects answered each question in a 5-level Likert scale (-2 to +2). So, for each question, a 2×5 table is created which compares the two experiment conditions (i.e., rapport-enabled vs. neutral). The table rows are the experiment conditions, and the columns are the Likert scales (i.e., -2, -1, 0, +1, and +2). Users' responses were analyzed using the Mantel-Haenszel-Chi-Square statistical method (degree of freedom

$df = 1$) [MDK03]. I followed the null-hypothesis of “characters with different levels of rapport abilities (rapport-enabled vs. neutral) have the same effects on the users”. Therefore, under the assumption of the null-hypothesis, a Chi-Square p value of less than 0.05 ($df = 1$) rejects the null-hypothesis.

As presented in Table 5.12, Chi-Square analyses show that, rapport-building ability of the character made a significant difference between the neutral and rapport-enabled characters in all measured aspects, except in *perceived ease of use*, *information disclosure*, and *perceived precision of expressions*. In other words, there is not enough statistical evidence to show that rapport-building ability affects these aspects significantly.

Also, I compared the mean values of the same statements in the two experimental conditions to calculate the possible improvement/deterioration of them upon each other. The improvement/deterioration is calculated using Equation 4.1 provided in Chapter 4. As indicated in Table 5.12, the *rapport-enabled* character was perceived positively in all measured aspects. Although the *neutral* character was also perceived positively in many of the measured aspects, it was perceived negatively in terms of **intention to use**, **perceived sociability**, *social presence*, *affective experience*, **naturalness**, **rapport**, *recall of expressions*, *anthropomorphism*, *animacy*, and *expressiveness*. This means that, subjects perceived these aspects of the neutral character negatively. As shown in Table 5.12, in all of the measured aspects, mean value comparison indicates an improvement of the rapport-enabled character over the neutral one.

In the neutral and rapport-enabled conditions, users selected the “I prefer not to answer” option 9 and 8 times respectively, which does not show a significant difference between the information disclosure of users in these two conditions. This result

Table 5.12: Subjective evaluation mean value comparison and Chi-Square test results.

Evaluated Aspect	Agent	Mean	Std. Dev.	χ^2	p	Improvement	Hypothesis																																																																																																																																																																																																																																													
Attitude	Neutral	0.22	1.12	27.92	0.0000	26.53%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.28	0.71					Intention to Use	Neutral	-0.15	1.24	16.53	0.0000	34.54%	<i>Rejected</i>	Rapport	1.23	0.88	Perceived Enjoyment	Neutral	0.00	1.19	30.82	0.0000	30.42%	<i>Rejected</i>	Rapport	1.22	0.78	Perceived Ease of Use	Neutral	1.26	0.58	3.78	0.0518	7.69%	<i>Not rejected</i>	Rapport	1.57	0.56	Perceived Sociability	Neutral	-0.29	1.12	71.66	0.0000	32.18%	<i>Rejected</i>	Rapport	1.00	0.75	Perceived Usefulness	Neutral	0.22	1.10	10.33	0.001	22.78%	<i>Rejected</i>	Rapport	1.13	0.80	Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>	Rapport	0.82	0.86	Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312
Intention to Use	Neutral	-0.15	1.24	16.53	0.0000	34.54%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.23	0.88					Perceived Enjoyment	Neutral	0.00	1.19	30.82	0.0000	30.42%	<i>Rejected</i>	Rapport	1.22	0.78	Perceived Ease of Use	Neutral	1.26	0.58	3.78	0.0518	7.69%	<i>Not rejected</i>	Rapport	1.57	0.56	Perceived Sociability	Neutral	-0.29	1.12	71.66	0.0000	32.18%	<i>Rejected</i>	Rapport	1.00	0.75	Perceived Usefulness	Neutral	0.22	1.10	10.33	0.001	22.78%	<i>Rejected</i>	Rapport	1.13	0.80	Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>	Rapport	0.82	0.86	Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94						
Perceived Enjoyment	Neutral	0.00	1.19	30.82	0.0000	30.42%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.22	0.78					Perceived Ease of Use	Neutral	1.26	0.58	3.78	0.0518	7.69%	<i>Not rejected</i>	Rapport	1.57	0.56	Perceived Sociability	Neutral	-0.29	1.12	71.66	0.0000	32.18%	<i>Rejected</i>	Rapport	1.00	0.75	Perceived Usefulness	Neutral	0.22	1.10	10.33	0.001	22.78%	<i>Rejected</i>	Rapport	1.13	0.80	Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>	Rapport	0.82	0.86	Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																	
Perceived Ease of Use	Neutral	1.26	0.58	3.78	0.0518	7.69%	<i>Not rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.57	0.56					Perceived Sociability	Neutral	-0.29	1.12	71.66	0.0000	32.18%	<i>Rejected</i>	Rapport	1.00	0.75	Perceived Usefulness	Neutral	0.22	1.10	10.33	0.001	22.78%	<i>Rejected</i>	Rapport	1.13	0.80	Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>	Rapport	0.82	0.86	Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																												
Perceived Sociability	Neutral	-0.29	1.12	71.66	0.0000	32.18%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.00	0.75					Perceived Usefulness	Neutral	0.22	1.10	10.33	0.001	22.78%	<i>Rejected</i>	Rapport	1.13	0.80	Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>	Rapport	0.82	0.86	Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																							
Perceived Usefulness	Neutral	0.22	1.10	10.33	0.001	22.78%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.13	0.80					Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>	Rapport	0.82	0.86	Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																		
Social Presence	Neutral	-0.38	1.12	45.71	0.0000	30.12%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.82	0.86					Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>	Rapport	1.07	0.75	Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																													
Trust	Neutral	0.31	1.01	17.22	0.0000	18.80%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.07	0.75					Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>	Rapport	0.69	1.05	Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																								
Information Disclosure	Neutral	0.39	1.43	2.33	0.1267	7.35%	<i>Not rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.69	1.05					Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>	Rapport	1.07	0.73	Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																			
Sensory Experience	Neutral	0.28	1.16	16.35	0.0000	19.72%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.07	0.73					Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>	Rapport	0.97	0.79	Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																														
Cognitive Experience	Neutral	0.37	1.19	4.54	0.0331	14.91%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.97	0.79					Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>	Rapport	0.40	0.96	Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																									
Affective Experience	Neutral	-0.47	0.92	29.75	0.0000	21.73%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.40	0.96					Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>	Rapport	1.1	0.71	Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																				
Engagement	Neutral	0.35	1.09	32.64	0.0000	18.70%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.1	0.71					Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>	Rapport	0.53	1.12	Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																															
Naturalness	Neutral	-0.56	1.16	20.92	0.0000	27.22%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.53	1.12					Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>	Rapport	0.74	0.82	Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																										
Rapport	Neutral	-0.38	1.08	58.38	0.0000	28.03%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.74	0.82					Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>	Rapport	0.8	1.11	Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																					
Precision of Expressions	Neutral	0.78	1.23	0.005	0.9432	0.56%	<i>Not rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.8	1.11					Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>	Rapport	0.60	1.33	Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																
Recall of Expressions	Neutral	-1.11	0.83	20.53	0.0000	42.78%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.60	1.33					Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>	Rapport	0.71	0.87	Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																											
Anthropomorphism	Neutral	-0.69	1.09	95.99	0.0000	35.06%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.71	0.87					Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>	Rapport	1.42	0.69	Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																																						
Likability	Neutral	0.53	0.96	62.95	0.0000	22.17%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.42	0.69					Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>	Rapport	1.02	0.81	Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																																																	
Animacy	Neutral	-0.11	1.21	40.42	0.0000	28.29%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.02	0.81					Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>	Rapport	1.18	0.85	Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																																																												
Perceived Intelligence	Neutral	0.15	1.19	22.90	0.0000	25.88%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.18	0.85					Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>	Rapport	1.10	0.75	Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																																																																							
Expressiveness	Neutral	-0.70	1.33	23.42	0.0000	45.09%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	1.10	0.75					Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>	Rapport	0.80	0.94																																																																																																																																																																																																																																		
Perceived Personality	Neutral	0.26	0.84	4.64	0.0312	13.52%	<i>Rejected</i>																																																																																																																																																																																																																																													
	Rapport	0.80	0.94																																																																																																																																																																																																																																																	

confirms the results of the Chi-Square test and mean value comparison performed for *Information Disclosure*.

Two human subjects reviewed the recorded interaction videos, and reported that users in the neutral condition had expressed 33 instances of positive facial expressions (83 seconds in total) and 15 instances of negative facial expressions (47 seconds in total) all together. Also, in the rapport-enabled condition, they reported 98 instances of positive facial expressions (total of 458 seconds) and 6 instances of negative facial expressions (11 seconds in total), which shows that the rapport model was able to convey the positivity feature of the rapport to the users, therefore, users had more positive facial expressions during the interaction with rapport-enabled character. Also, research [GM04, GWO07] shows that, positive facial expressions are indicative of rapport, while negative facial expressions are indicative of lack of rapport.

As stated above, one of the questions for evaluating the the affective experience of the users with the characters was “I had emotional reactions while interacting with the counselor”. I asked the users to list their emotional reactions, if there was any. The following are the emotional reactions that users mentioned after using the rapport-enabled character: *happiness/joy/pleased* (7 times), *content* (1 time), *pleasure* for being understood (2 times), *surprised* (1 time), *sympathy* (1 time), *curiosity* (1 time), *excitement* (1 time), *calm* (2 times), *confidence* (1 time), and *shame* (1 time). For the neutral character condition, users mentioned *happiness* (1 time), *neutral* (1 time), *board* (1 time), *annoyed* (2 times), *confused* (1 time), *regret* (1 time), *guilt* (2 times), *sadness* (2 times), *bad memories* (2 times), and *defensive* (1 time), as their emotional reactions during the interaction.

I also asked the users to mention some of the inappropriate gestures that they felt in the characters’ behavior. They reported the constant forward eye gaze of the

neutral character as inappropriate, which was addressed by the gaze away models (i.e., gaze left and right) in the rapport-enabled character.

In Chapter 4, I discussed my approach to enable the character to empathize with the users using a decision tree (i.e., rule-based). In this chapter, I used data-driven modeling using machine learning for modeling the rapport. I compared the results of the user studies performed to evaluate these two studies. The evaluation schema of the two studies have common measurements including the Heerink’s questionnaire [HKEW09] and Bartneck’s questionnaire [BKC08], which enabled me to compare the users’ perceptions of the machine learning approach and the decision tree approach. In other words, the mean value (and standard deviation) is compared only for those statements that were common in evaluation of the decision tree and machine learning approaches. Table 5.13 shows the results of comparing the user response mean values in each of the categories.

Table 5.13: Comparing evaluations of decision tree and machine learning (ML) approaches.

Evaluated Aspect	Agent	Mean	Std. Dev.	Improvement
Attitude	Dec. tree	0.78	0.9	12.58%
	ML	1.28	0.71	
Intention to Use	Dec. tree	0.8	0.89	10.83%
	ML	1.23	0.88	
Perceived Enjoyment	Dec. tree	1.08	0.52	3.42%
	ML	1.22	0.78	
Perceived Ease of Use	Dec. tree	1.6	0.49	-0.83%
	ML	1.57	0.56	
Perceived Sociability	Dec. tree	0.8	0.87	5.00%
	ML	1.00	0.75	
Perceived Usefulness	Dec. tree	0.64	0.89	12.33%
	ML	1.13	0.80	
Social Presence	Dec. tree	0.35	1.04	11.81%
	ML	0.82	0.86	
Trust	Dec. tree	0.78	0.83	7.17%
	ML	1.07	0.75	
Anthropomorphism	Dec. tree	0.28	1.05	10.83%
	ML	0.71	0.87	
Likability	Dec. tree	1.29	0.64	3.25%
	ML	1.42	0.69	
Animacy	Dec. tree	0.52	0.98	12.51%
	ML	1.02	0.81	
Perceived Intelligence	Dec. tree	0.97	0.77	5.58%
	ML	1.18	0.85	

Although both objective and subjective results are acceptable, and show improvements over the other approaches of modeling non-verbal behaviors and rapport, there are still some limitations to be addressed in future studies (discussed in more details in Section 6.2). Some of the limitation are as follow:

- There are some non-verbal behaviors for which I did not have enough data in the videos for model generation (i.e., speaker head lateral sweep, speaker large smile, speaker afraid face, puzzled face, speaker head movements, listener head lateral sweep, listener head nod-shake, listener head shake, listener large smile, listener afraid face, listener raised brows, listener opened hands, listener point hands, listener contrast hands, listener hand formless flicks, listener lean left, and listener right lean). This limitation can be addressed by annotating more videos and generating more data to enable us model the missing non-verbal behaviors.
- Some of the models of non-verbal behaviors have lower objective performance in comparison to others (e.g., speaker gaze left, speaker/listener lean forward), which can be improved with more data collection.
- In the data annotation phase, a few of the features were annotated manually (e.g., body lean, head gesture, hand gesture), which was a time intensive process. Designing and implementing automatic visual feature recognizers can automate this phase and reduce the time required to generate the non-verbal models.
- In this research, the focus was on modeling the non-verbal behaviors, therefore, a simple dialog planner was used, which was selecting the utterances in a pre-defined order without taking into account the user's responses as a trigger for selecting the next utterances. More dynamic dialog planners and spoken dialog management systems can address this limitation and enable the character to have a more believable verbal interaction with the users too.

- The current character system (i.e., Haptik) is a platform dependent (i.e., Windows) application. Therefore, we are limited in using this character on other platforms. This limitation can be addressed by using character systems that are less platform dependent or rendering the character using a platform independent approaches such as WebGL (see detailed discussion in Section 6.2).

5.11 Summary

In this chapter, I studied modeling human non-verbal behaviors from video and conversation text corpora, using machine learning. I modeled different non-verbal behaviors for both speaker and listener roles of a virtual character, including head gesture (i.e., nod, shake, and nod-shake), eye gaze (i.e., left, and right), subtle smile, hand gestures (i.e., formless flick, pointing, contrast, iconic, closed, and opened), emotional facial expressions (i.e., neutral, happy, sad, surprised, angry/puzzled, and disgusted), eyebrow movement (i.e., up and down), and lean (i.e., forward and left).

I evaluated each individual non-verbal behavior using objective tests, and also evaluated their combination, as a rapport model, using subjective tests (i.e., user studies). Evaluation results show high accuracy of the individual models and also improvements of a rapport-enabled character over a neutral one. Evaluations compare the neutral and rapport-enabled characters in terms of engagement, naturalness, rapport, attitude, intention to use, perceived enjoyment, perceived ease of use, perceived sociability, perceived usefulness, social presence, trust, information disclosure, sensory experience, cognitive experience, affective experience, perceived precision of expressions, perceived recall of expressions, anthropomorphism, likability, animacy, perceived intelligence, expressiveness, and perceived personality. Evaluation show high improvements of the rapport-enabled character over the neutral character. Also, comparing with the decision tree approach presented in Chapter 4, subjective eval-

uations show high improvements of the machine learning approach presented in this chapter.

CHAPTER 6

Conclusions

This chapter describes the summary of my contributions and possible future research directions.

6.1 Summary

In this dissertation, I mainly focused on the design and development of a data-driven computational model of non-verbal rapport for virtual characters, using the data derived from video corpora of human-human interactions. In this approach, I used machine learning to learn human non-verbal behaviors, and modeled non-verbal rapport using the combination of those models.

The following list summarizes the contributions of this research:

1. Extracting information from the lexical and syntactical structure of the surface text to support the automatic generation of believable non-verbal behaviors using machine learning techniques.
2. Extracting information from human-human counseling video corpora to support the automatic generation of believable non-verbal behaviors using machine learning techniques.
3. Using interactive realtime features, such as facial expressions, head movements, and gaze directions, for modeling the non-verbal behaviors.
4. Modeling a set of non-verbal behaviors for the virtual character in both speaker and listener roles. For the speaker role, head gesture (nod, shake, non-shake), subtle smile, gaze (left and right), facial expression (happy, surprised, angry/puzzled, and disgusted), eyebrow movement (up and down), hand gesture

(formless-flick, point, contrast, iconic, closed, and open), and body lean (forward and left) are modeled. For listener role, head nod, subtle smile, gaze (left and right), facial expression (happy, surprised, angry/puzzled, disgusted), eyebrow down, hand closed gesture, and body lean forward are modeled.

5. Combining a set of non-verbal behavior models (stated above), generated using machine learning techniques, to model non-verbal rapport communication for a virtual character.
6. Mapping all the possible facial muscle movements, head movements, and head gestures of a virtual character to the Action Units (AUs) of the Facial Action Coding System (FACS).
7. Applying the non-verbal rapport-enabled communication model to a virtual health counselor to improve the user acceptance of the character and perceived character features.

My contribution has impacts on two areas of human-computer interaction, and computer based health intervention systems. In addition, I developed computational resources to map FACS action units on virtual characters' faces, in order to generate standard facial expressions, which impacts on the areas of psychology and emotion theory research as well.

6.2 Future Directions

As mentioned in Section 3.1, some of the HapFACS action unit expressions still need improvements (i.e., AUs 11, 13, 14, 16, 20, 41, 42, 44). Also, for further studies, I suggest to provide non-linear changing of the intensity for video generation, because AU activation can be non-linear from a geometric point of view. Therefore, future

versions of HapFACS will improve the expressiveness of the imperfect AUs and will enable non-linear AU activations.

Moreover, the automatic non-verbal behavior generation can be extended in several directions. In order to model new non-verbal behaviors (that are missing in this dissertation, i.e., speaker head lateral sweep, speaker large smile, speaker afraid face, puzzled face, speaker head movements, listener head lateral sweep, listener head nod-shake, listener head shake, listener large smile, listener afraid face, listener raised brows, listener opened hands, listener point hands, listener contrast hands, listener hand formless flicks, listener lean left, and listener right lean) and improve the currently modeled behaviors, more videos can be annotated and more data can be collected. Also, automatic visual feature recognizers can be implemented for annotating those visual features that were annotated manually in this research. Therefore, the video annotation phase will be automated completely and the data annotation time will be reduced significantly.

Recent growth of using the smart phones and other portable smart devices is a motivation to port the current system to other platforms which increases the availability and accessibility significantly. New improvements in HTML5 and WebGL technologies enable us to render 3D virtual characters in web browsers that support HTML5 (including Safari, Chrome, Firefox, and Internet Explorer). Therefore, generating new characters in WebGL and integrating them with my system (as web services) enable us to deliver rapport-enabled virtual characters on different platforms. As explained in Section 3.2, HapGest software is responsible for generating the non-verbal behaviors and synchronizing them with the verbal behaviors (i.e., words in the sentence). HapGest uses events to perform this function. For animating the WebGL characters on the browser, a similar event handling process needs to be implemented for synchronizing the new characters' verbal and non-verbal behaviors. Therefore, the

WebGL character receives a tagged text (from rapport model web-service), in which words are tagged with appropriate non-verbal behaviors using bookmark events. The WebGL character uses the Text-To-Speech (TTS) engine to read the sentence and throw events for each reached bookmark (i.e, non-verbal behavior). Finally, the event handler animates the character with corresponding non-verbal behaviors.

In this research, the focus was on modeling the non-verbal behaviors, therefore, I used a simple dialog planner, which was selecting the utterances in a pre-defined order without taking into account the user's responses as a trigger for selecting the next utterance. The future version of the dialog planner can be a spoken dialog management system, which enables the character to select the best utterances based on the user's responses.

Last but not least, the completed web-based virtual health counselor can be used to deliver health interventions to real problem drinkers (or people who have other unhealthy life styles), in order to (1) study the system's effects on their behavior change, and (2) compare the system's performance with other web-based interventions such as Drinker's Check-Up, based on which the On-Demand Virtual Counselor (ODVIC) was implemented.

BIBLIOGRAPHY

- [ADJE06] Ali Arya, Steve DiPaola, Lisa Jefferies, and James T. Enns. Socially Communicative Characters for Interactive Applications. In *Proceedings of the 14th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG'06)*, Plzen, Czech Republic, 2006. UNION Agency Science Press.
- [AFB⁺13] Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, Mike Eheler, Zybnek Kysela, and Javier Pahlen. Digital Ira: Creating a Real-Time Photoreal Digital Actor. In *Proceedings of the ACM SIGGRAPH'13*, page 4. ACM, 2013.
- [AL13] Reza Amini and Christine Lisetti. HapFACS: an Open Source API/Software to Generate FACS-Based Expressions for ECAs Animation and for Corpus Generation. In *Proceedings of the Fifth Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII'13)*, pp 270–275, Geneva, Switzerland, 2013. IEEE Computer Society.
- [All02] Jens Allwood. Bodily Communication Dimensions of Expression and Content. *Multimodality in Language and Speech Systems*, 7(26):pp 1–15, 2002.
- [ALY14] Reza Amini, Christine Lisetti, and Ugan Yasavur. Emotionally Responsive Virtual Counselor for Behavior-Change Health Interventions. In *Proceedings of the 9th International Conference on Design Science Research in Information Systems and Technology (DESRIST), LNCS8463*, pp 433–437, Miami, FL, 2014. Springer International Publishing Switzerland.
- [ALYR13] Reza Amini, Christine Lisetti, Ugan Yasavur, and Naphtali Rishe. On-Demand Virtual Health Counselor for Delivering Behavior-Change Health Interventions. In *Proceedings of the IEEE International Conference on Healthcare Informatics 2013 (ICHI'13)*, pp 46–55, Philadelphia, PA, USA, 2013. IEEE.
- [AN14] Márcio Alencar and José Francisco Netto. TUTOR Collaborator Using Multi-Agent System. *Collaboration Technologies and Social Computing*, 460 (Communications in Computer and Information Science):153–159, 2014.

- [AR92] N. Ambady and Robert Rosenthal. Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: a Meta-Analysis. *Psychological Bulletin*, 111(2):256–274, 1992.
- [ARL⁺09] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. Creating a Photoreal Digital Actor: The Digital Emily Project. In *Proceedings of the Conference for Visual Media Production (CVMP'09)*, pp 176–187. IEEE Computer Society, November 2009.
- [AWL11] Carl Arrington, Dale-Marie Wilson, and Lorrie Lehmann. Improving Performance and Retention in Computer Science Courses Using a Virtual Game Show. In *Proceedings of the 49th Annual Southeast Regional Conference on ACM-SE'11*, page 320, New York, New York, USA, 2011. ACM Press.
- [AYL12] Reza Amini, Ugan Yasavur, and Christine L. Lisetti. HapFACS 1.0: Software/API for Generating FACS-Based Facial Expressions. In *Proceedings of the ACM 3rd International Symposium on Facial Analysis and Animation (FAA'12)*, Article 17, Vienna, AUSTRIA, 2012. ACM Press.
- [BAD02] B. L. Burke, H. Arkowitz, and C. Dunn. The Efficacy of Motivational Interviewing and Its Adaptation. In *Motivational Interviewing: Preparing People for Change*, pages 217–250. Guilford Press, New-York,NY, 2nd edition, 2002.
- [BAW09] Christian Becker-Asano and Ipke Wachsmuth. Affective Computing with Primary and Secondary Emotions in a Virtual Human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49, May 2009.
- [BBA07] Hana Boukricha and Christian Becker-Asano. Simulating Empathy for the Virtual Human Max. In Dirk Reichardt and Paul Levi, editors, *Proceedings of the 2nd Workshop at KI2007 on Emotion and Computing Current Research and Future Impact*, pp 23–28, Osnabrück, Germany, 2007.
- [BBBL01] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. Equilibrium Theory Revisited : Mutual Gaze and Personal Space. *Presence*, 10(6):583–598, 2001.

- [BBLM86] Janet Beavin Bavelas, Alex Black, Charles R. Lemery, and Jennifer Mullett. “I Show How You Feel” - Motor Mimicry as a Communicative Act. *Journal of Personality and Social Psychology*, 50(2):322–329, 1986.
- [BBLM87] Janet Beavin Bavelas, Alex Black, Charles R. Lemery, and Jennifer Mullett. Empathy and its Development. In Nancy Eisenberg and Janet Strayer, editors, *Motor Mimicry as Primitive Empathy*, pages 317–338. Cambridge University Press, Cambridge, UK, 1987.
- [BCL00] S. Battista, F. Casalino, and C. Lande. MPEG-4: a Multimedia Standard for the Third Millennium, Part 1. In *IEEE Multimedia*, volume 7, pages 74–83. IEEE, 2000.
- [BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Daniel Tapias Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, editor, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, volume 0, pages 2200–2204, Valletta, Malta, 2010. European Language Resources Association (ELRA).
- [Beu11] Leon Beutl. *A Simulation for the Creation of Soft-Looking, Realistic Facial Expressions*. Master thesis, University of Wien, 2011.
- [BFA09] Nikolaus Bee, Bernhard Falk, and Elisabeth Andre. Simplified Facial Animation Control Utilizing Novel Input Devices: A Comparative Study. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI’09)*, pp 197–206, Sanibel Island, Florida, USA, 2009. ACM.
- [BFS87] C. D. Batson, J. Fultz, and P. A. Schoenrade. Adults’ Emotional Reactions to the Distress of Others. In N. Eisenberg and J. Strayer, editors, *Empathy and its Development*, pages 163–185. Cambridge University Press, Cambridge, 1987.
- [BG92] Thomas F. Babor and Marcus Grant. Programme on Substance Abuse: Project on Identification and Management of Alcohol-Related Problems. Report on Phase II, a Randomized Clinical Trial of Brief Interventions in Primary Health Care. Technical Report, World Health Organization (WHO), 1992.

- [BH02] Frank Biocca and Chad Harms. Defining and Measuring Social Presence: Contribution to the Networked Minds Theory and Measure. *Proceedings of PRESENCE*, 2002(517):1–36, 2002.
- [BH05] M. G. Beaupre and U. Hess. Cross-cultural Emotion Recognition Among Canadian Ethnic Groups. *Journal of Cross-Cultural Psychology*, 36:355–370, 2005.
- [BHBSM01] Thomas F. Babor, John C. Higgins-Biddle, John B. Saunders, and Maristela G. Monteiro. *AUDIT: The Alcohol Use Disorders Identification Test. Guidelines for Use in Primary Health Care*. World Health Organization (WHO), Department of Mental Health and Substance Dependence, 2 edition, 2001.
- [BIC⁺07] Farah Benamara, Sabatier Irit, Carmine Cesarano, Napoli Federico, and Diego Reforgiato. Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. In *Proceedings of International Conference on Weblogs and Social Media*, pp 1–4, 2007.
- [Bie93] JS. Bien TH., and Miller WR., and Tonigan. Brief Interventions for Alcohol Problems: A Review. *Addiction*, 88:315–336, 1993.
- [BKC08] Christoph Bartneck, Dana Kulic, and Elizabeth Croft. Measuring the Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety of Robots. In *Proceedings of the Metrics for Human-Robot Interaction Workshop in Affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI'08)*, *Technical Report 471*, volume 471, pp 37–44, Amsterdam, 2008. University of Hertfordshire.
- [BL07] Margaret M. Bradley and P. J. Lang. The International Affective Digitized Sounds Affective Ratings of Sounds and Instruction Manual. *Emotion*, pp 29–46, 2007.
- [Bla05] R. J. R. Blair. Responding to the Emotions of Others: Dissociating Forms of Empathy Through the Study of Typical and Psychiatric Populations. *Consciousness and Cognition*, 14(4):698–718, December 2005.
- [BLL⁺04] M.S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine Learning Methods for Fully Automatic Recognition of Facial Expressions and Facial Actions. In *Proceedings of the 2004 IEEE In-*

ternational Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), volume 1, pp 592–597. IEEE, 2004.

- [BM91] P. S. Bellet and M. J. Maloney. The Importance of Empathy as an Interviewing Skill in Medicine. *JAMA: the Journal of the American Medical Association*, 266(13):1831–2, October 1991.
- [Bou13] Hana Boukricha. *Simulating Empathy in Virtual Humans*. Ph.D. Thesis, Bielefeld: Bielefeld University, 2013.
- [BP05] Timothy Wallace Bickmore and Rosalind W. Picard. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):617–638, 2005.
- [BPI07] Werner Breitfuss, Helmut Prendinger, and Mitsuru Ishizuka. Automated Generation of Non-Verbal Behavior for Virtual Embodied Characters. In *Proceedings of the Ninth International Conference on Multimodal Interfaces (ICMI'07)*, pp 319–322, New York, New York, USA, 2007. ACM Press.
- [BPJ09] Timothy Wallace Bickmore, Laura M. Pfeifer, and Brian W. Jack. Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents. In *Proceedings of the 27th International ACM Conference on Human Factors in Computing Systems (CHI'09)*, pp 1265–1274, New York, 2009. ACM.
- [BPN+10] Elisabetta Bevacqua, Ken Prepin, Radoslaw Niewiadomski, Etienne de Sevin, and Catherine Pelachaud. GRETA: Towards an Interactive Conversational Virtual Companion. In Yorick Wilks, editor, *Close Engagement with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, pages 143–156. John Benjamins Publishing Co., 2010.
- [BR11] Karla Bransky and Debbie Richards. Users' Expectations of IVA Recall and Forgetting. In *Proceedings of the Intelligent Virtual Agents 10th International Conference (IVA'11)*, pp 433–434. Springer-Verlag Berlin Heidelberg, 2011.
- [BS07] Timothy Wallace Bickmore and Daniel Schulman. Practical Approaches to Comforting Users with Relational Agents. In *Proceeding of*

ACM CHI 2007 Conference on Human Factors in Computing Systems, pp 2291–2296, San Jose, California, USA, 2007. ACM.

- [BST10] Tibor Bosse, Ghazanfar F. Siddiqui, and Jan Treur. An Intelligent Virtual Agent to Increase Involvement in Financial Services. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA '10)*, pp 378–384. Springer-Verlag Berlin, Heidelberg, 2010.
- [BSV⁺96] J. Brug, I. Steenhuis, P. Van Assema, and H. De Vries. The Impact of a Computer-Tailored Nutrition Intervention. *Preventive Medicine*, 25(3):236–242, 1996.
- [BT11] Tara S. Behrend and Lori Foster Thompson. Similarity Effects in Online Training: Effects with Computerized Trainer Agents. *Computers in Human Behavior*, 27(3):1201–1206, May 2011.
- [BTB⁺08] Bridgette M. Bewick, Karen Trusler, Michael Barkham, Andrew J. Hill, Jane Cahill, and Brendan Mulhern. The Effectiveness of Web-Based Interventions Designed to Decrease Alcohol Consumption—a Systematic Review. *Preventive Medicine*, 47(1):17–26, July 2008.
- [BTTS12] Tara S. Behrend, Steven Toaddy, Lori Foster Thompson, and David J. Sharek. The Effects of Avatar Appearance on Interviewer Ratings in Virtual Employment Interviews. *Computers in Human Behavior*, 28(6):2128–2133, July 2012.
- [BW11] Hana Boukricha and Ipke Wachsmuth. Empathy-Based Emotional Alignment for a Virtual Human: A Three-Step Approach. *KI-Künstliche Intelligenz*, 25(3):195–204, May 2011.
- [Cas01] Justine Cassell. More than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. *Knowledge-Based Systems*, 14(1-2):55–64, 2001.
- [CB99] T. L. Chartrand and J. A. Bargh. The Chameleon Effect: the Perception-Behavior Link and Social Interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, June 1999.
- [CGL10] Dustin B. Chertoff, Brian Goldiez, and Joseph J. LaViola. Virtual Experience Test: A Virtual Environment Evaluation Questionnaire. In *IEEE Proceedings of the Virtual Reality Conference (VR)*, pp 103–110, Waltham, Massachusetts, USA, March 2010. IEEE.

- [Cha00] Eugene Charniak. A Maximum-Entropy-Inspired Parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL' 2000)*, number c, pp 132–139. Association for Computational Linguistics, Stroudsburg, PA, USA, 2000.
- [Cho91] Nicole Chovil. Discourse-Oriented Facial Displays in Conversation. *Research on Language and Social Interaction*, 25:163–194, 1991.
- [CID⁺03] Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C Mazziotta, and Gian Luigi Lenzi. Neural Mechanisms of Empathy in Humans: a Relay from Neural Systems for Imitation to Limbic Areas. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5497–502, May 2003.
- [Cli02] Christina Cliffordson. The Hierarchical Structure of Empathy: Dimensional Organization and Relations to Social Functioning. *Scandinavian Journal of Psychology*, 43(1):49–59, February 2002.
- [CMK⁺06] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, and Kostas Karpouzis. Modeling Naturalistic Affective States via Facial and Vocal Expressions Recognition. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, pp 146–154. ACM, 2006.
- [CML05] T. L. Chartrand, W. W. Maddux, and J. L. Lakin. Beyond the Perception-Behavior Link: The Ubiquitous Utility and Motivational Moderators of Nonconscious Mimicry. In *The New Unconscious*, pages 334–361. Oxford University Press New York, 2005.
- [CNB⁺01] Justine Cassell, Yukiko I. Nakano, Timothy Wallace Bickmore, Candace L. Sidner, and Charles Rich. Non-Verbal Cues for Discourse Structure. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL'01)*, pp 114–123, 2001.
- [CSPC00] Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth F. Churchill. Embodied Conversational Agents. *Social Psychology*, 40(1):26–36, 2000.
- [CSX04] Richard Catrambone, John Stasko, and Jun Xiao. ECA as User Interface Paradigm: Experimental Findings within a Framework for Research. In Zsófia Ruttkay and Catherine Pelachaud, editors, *From*

Brows to Trust: Evaluating Embodied Conversational Agents, chapter 9, pages 239–267. Kluwer Academic Publishers, 2004.

- [CTP99] Justine Cassell, Obed E. Torres, and Scott Prevost. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. In Y. Wilks, editor, *Machine Conversations*, pages 143–154. Kluwer, 1999.
- [Cun99] A. Cunningham, J.A., Humphreys, K., & Koski-Jannes. Providing Personalized Assessment Feedback for Problem Drinking on the Internet. In *the 33rd Annual Convention of the Association for the Advancement of Behavior Therapy, Toronto.*, 1999.
- [CVB01] Justine Cassell, H.H. Vilhjálmsón, and Timothy Wallace Bickmore. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*, pp 477–486. ACM, 2001.
- [CVG⁺10] Marc Cavazza, C. Emilio Vargas, José Relación Gil, I D Telefónica, Nigel Crook, Debora Field, and S. Sheffield. 'How Was Your Day ?' An Affective Companion ECA Prototype. In *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, volume 1, pp 277–280, The University of Tokyo, 2010. Association for Computational Linguistics.
- [CW98] H. H. Clark and T. Wasow. Repeating Words in Spontaneous Speech. *Cognitive Psychology*, 37(3):201–42, December 1998.
- [Dam94] Antonio R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*, volume 33. Putnam, 1994.
- [Dav83] Mark H. Davis. Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology*, 44(1):113–126, 1983.
- [Dav94] Mark H. Davis. *Empathy: A Social Psychological Approach*. Westview Press, 1994.
- [DCP02] Berardina De Carolis, Valeria Carofiglio, and Catherine Pelachaud. From Discourse Plans to Believable Behavior Generation. In *Proceedings of the 2nd International Conference on Natural Language Generation (INLG'02)*, New York, USA, 2002.

- [DCS⁺12] Tamara L. Dunn, Leanne M. Casey, Jeanie Sheffield, Peter Newcombe, and Anne B. Chang. Dropout from Computer-Based Interventions for Children and Adolescents with Chronic Health Conditions. *Journal of Health Psychology*, 17(3):429–42, April 2012.
- [DDR01] C. Dunn, L. Deroo, and F.P. Rivara. The Use of Brief Interventions Adapted from Motivational Interviewing Across Behavioral Domains: a Systematic Review. *Addiction*, 96(12):1725–42, 2001.
- [Den87] D. C. Dennett. *The Intentional Stance*. MIT Press, 1987.
- [DMNP10] Berardina De Carolis, Irene Mazzotta, Nicole Novielli, and Sebastiano Pizzutilo. Social Robots and ECAs for Accessing Smart Environments Services. In *Proceedings of the International Conference on Advanced Visual Interfaces (AVI'10)*, pp 275–278, New York, New York, USA, 2010. ACM Press.
- [Doh00] J. Doherty, Y., Hall, D., James, P.T., Roberts, S.H., & Simpson. Change Counseling in Diabetes: The Development of a Training Programme for the Diabetes Team. *Patient Education & Counseling*, 40:263–278, 2000.
- [DRSV02] D. DeCarlo, C. Revilla, Matthew Stone, and J.J. Venditti. Making Discourse Visible: Coding and Animating Conversational Facial Displays. In *Proceedings of Computer Animation 2002 (CA'02)*, volume 2002, pages 11–16. IEEE Computer Society, 2002.
- [dVS06] Frederique de Vignemont and Tania Singer. The Empathic Brain: How, When and Why? *Trends in Cognitive Sciences*, 10(10):435–41, October 2006.
- [EBLvH07] H. A. Elfenbein, M. Beaupre, M. Le vesque, and U. Hess. Toward a Dialect Theory: Cultural Differences in the Expression and Recognition of Posed Facial Expressions. *Emotion*, 7(1):131–146, 2007.
- [EF74] Paul Ekman and Wallace V. Freisen. Detecting Deception from the Body or Face. *Journal of Personality and Social Psychology*, 29(3):288–298, 1974.
- [EF78] Paul Ekman and Wallace V. Freisen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

- [EF86] Paul Ekman and W. V. Friesen. A New Pancultural Facial Expression of Emotion. *Motivation and Emotion*, 10(2):159–168, 1986.
- [EFA80] Paul Ekman, Wallace V. Freisen, and Sonia Ancoli. Facial Signs of Emotional Experience., 1980.
- [EFH02] Paul Ekman, Wallace V. Freisen, and Joseph C. Hager. *Facial Action Coding System*, volume 160. Research Nexus eBook, Salt Lake City, UT, 2nd edition, 2002.
- [ELF83] Paul Ekman, Robert W Levenson, and Wallace V. Freisen. Autonomic Nervous System Activity Distinguishes among Emotions. *Science*, 221(4616):1208–1210, 1983.
- [EM01] Andrew J. Elliot and holly A. McGregor. A 2 x 2 Achievement Goal Framework. *Journal of Personality and Social Psychology*, 80(3):501–519, 2001.
- [ER01] K. M. Emmons and S. Rollnick. Motivational Interviewing in Health Care Settings. Opportunities and limitations. *American Journal of Preventive Medicine*, 20(1):68–74, January 2001.
- [FE83] Wallace V. Friesen and Paul Ekman. EMFACS-7: Emotional Facial Action Coding System. *Unpublished Manuscript, University of California at San Francisco*, 1983.
- [Fes87] Norma Deitch Feshbach. Parental Empathy and Child Adjustment/Maladjustment. In N. Eisenberg and J. Strayer, editors, *Empathy and its Development*, Cambridge Studies in Social and Emotional Development., pages 271–291. New York, NY, US: Cambridge University Press, 1987.
- [FMS⁺12] Angela N. Fellner, Gerald Matthews, Kevin D. Shockley, Joel S. Warm, Moshe Zeidner, Lisa Karlov, and Richard D. Roberts. Using Emotional Cues in a Discrimination Learning Task: Effects of Trait Emotional Intelligence and Affective State. *Journal of Research in Personality*, 46(3):239–247, June 2012.
- [FO08] Mary Ellen Foster and Jon Oberlander. Corpus-Based Generation of Head and Eyebrow Motion for an Embodied Conversational Agent. *Language Resources and Evaluation*, 41(3-4):305–323, February 2008.

- [Fri94] A.J. Fridlund. *Human Facial Expression: An Evolutionary View*. Academic Press, San Diego, 1994.
- [Fri03] U. Frith. *Autism: Explaining the Enigma*, volume 21 of *Cognitive Development*. Blackwells, 2003.
- [Fus02] Susan R. Fussell. *The Verbal Communication of Emotions: Interdisciplinary Perspectives*. Lawrence Erlbaum Associates, 2002.
- [Gal03] Vittorio Gallese. The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity. *Psychopathology*, 36(4):171–180, 2003.
- [GdRLV08] E. Goeleven, R. de Raedt, L. Leyman, and B. Verschuere. The Karolinska Directed Emotional Faces: A Validation Study. *Cognition and Emotion*, 22:1094–1118, 2008.
- [GKW10] Jonathan Gratch, Sin-hwa Kang, and Ning Wang. Using Social Agents Explore Theories of Rapport and Emotional Resonance. Technical Report Chap X, University of Southern California, 2010.
- [GM85] Arnold P. Goldstein and Gerald Y. Michaels. *Empathy: Development, Training, and Consequences*. Hillsdale, N.J. and L. Erlbaum Associates, 1 edition, 1985.
- [GM86] Rand J. Gruen and Gerald Mendelsohn. Emotional Responses to Affective Displays in Others: The Distinction Between Empathy and Sympathy. *Journal of Personality and Social Psychology*, 51(3):609–614, 1986.
- [GM04] Jonathan Gratch and Stacy C. Marsella. A Domain-Independent Framework for Modeling Emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.
- [GOL06] Jonathan Gratch, Anna Okhmatovskaia, and Francois Lamothe. Virtual Rapport. In *Proceedings of the Intelligent Virtual Agents Conference (IVA)*, 2006.
- [Gor85] Ronald D. Gordon. Empathy: The State of the Art and Science. In *Proceedings of the International Conference of the World Communication Association*, pp 1–16, Baguio, Philippines, 1985.

- [Gra99] J.E. Grahe. The Importance of Nonverbal Cues in Judging Rapport. *Journal of Nonverbal Behavior*, 23(4):253–269, 1999.
- [GSM⁺11] Barbara Gonsior, Stefan Sosnowski, Christoph Mayer, Jiirgen Blume, B. Radig, D. Wollherr, and K. Kuhlentz. Improving Aspects of Empathy and Subjective Performance for HRI through Mirroring Facial Expressions. In *Proceedings of the RO-MAN, 20th IEEE International Symposium on Robot and Human Interactive Communication*, pp 350–356, Atlanta, GA, USA, 2011. IEEE.
- [GWGF07] Jonathan Gratch, Ning Wang, Jillian Gerten, and Edward Fast. Creating Rapport with Virtual Agents. In *Proceedings of the Intelligent Virtual Agents Conference (IVA)*, 2007.
- [GWO07] Jonathan Gratch, Ning Wang, and Anna Okhmatovskaia. Can Virtual Humans be More Engaging than Real Ones? In *Proceedings of the 12th International Conference on Human-Computer Interaction: Intelligent Multimodal Interaction Environments, (HCI'07)*, Chamonix, France, 2007. Springer-Verlag Berlin, Heidelberg.
- [Hal77] Michael Alexander Kirkwood Halliday. *Explorations in the Functions of Language*. Elsevier North-Holland, 1977.
- [HCR94] Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. Emotional Contagion. *Current Directions in Psychological Science*, 2(3):96–99, 1994.
- [HDC15] M. Sazzad Hussain, Sidney K. D’Mello, and Rafael A. Calvo. Research and Development Tools in Affective Computing. In *The Oxford Handbook of Affective Computing*, chapter 25, pages 349–357. Oxford University Press, 1 edition, 2015.
- [Hel04] Mitsuru Ishizuka Helmut Prendinger. *Life-Like Characters: Tools, Affective Functions, and Applications*. Springer, 2004.
- [Hes97] H.D. Hester, R.K., & Delaney. Behavioral Self-Control Program for Windows: Results of a Controlled Clinical Trial. *Journal of Consulting and Clinical Psychology*, 65:685–693, 1997.
- [HFG03] Robert C. Hubal, Geoffrey A. Frank, and Curry I. Guinn. Lessons Learned in Modeling Schizophrenic and Depressed Responsive Virtual

- Humans for Training. In *Proceedings of the 2003 International Conference on Intelligent User Interfaces (IUI'03)*, pp 85–92, Miami, FL, US, 2003. ACM.
- [HJR10] Adam T. Hirsh, Mark P. Jensen, and Michael E. Robinson. Evaluation of Nurses' Self-Insight into Their Pain Assessment and Treatment Decisions. *The Journal of Pain: Official Journal of the American Pain Society*, 11(5):454–61, May 2010.
- [HKEW09] Marcel Heerink, B. Krose, Vanessa Evers, and Bob Wielinga. Measuring Acceptance of an Assistive Social Robot: A Suggested Toolkit. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'09)*, pages 528–533. IEEE, 2009.
- [HMG10a] Lixing Huang, Louis-philippe Morency, and Jonathan Gratch. Learning Backchannel Prediction Model from Parasocial Consensus Sampling: A Subjective Evaluation. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA'10)*, pp 159–172. Springer-Verlag Berlin Heidelberg, 2010.
- [HMG10b] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In Van Der Hoek, Kaminka, Lesperance, Luck, and Sen, editors, *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10)*, pp 10–14, Toronto, Canada, 2010. International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).
- [HMG11] Lixing Huang, Louis-philippe Morency, and Jonathan Gratch. Virtual Rapport 2.0. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents (IVA'11)*, pp 68–79, Reykjaavik, Iceland, 2011. Springer-Verlag Berlin, Heidelberg.
- [Hof00] Martin L. Hoffman. *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press, 2000.
- [Hoj07] Mohammadreza Hojat. *Empathy in Patient Care: Antecedents, Development, Measurement, and Outcomes*. New York, NY: Springer, 2007.
- [HSD05] Reid K. Hester, Daniel D. Squires, and Harold D. Delaney. The Drinker's Check-up: 12-Month Outcomes of a Controlled Clinical Trial

of a Stand-Alone Software Program for Problem Drinkers. *Journal of Substance Abuse Treatment*, 28(2):159–169, 2005.

- [HSW⁺06] Frank Hegel, Torsten Spexard, Britta Wrede, G. Horstmann, and T. Vogt. Playing a Different Imitation Game: Interaction with an Empathic Android Robot. In *Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots*, pp 56–61. IEEE, 2006.
- [HTW11] Rüdiger Heimgärtner, L.W. Tiede, and Helmut Windl. Empathy as Key Factor for Successful Intercultural HCI Design. *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, pages 557–566, 2011.
- [Hun67] JB Hunsdahl. Concerning Einfühlung (Empathy): A Concept Analysis of its Origin and Early Development. *Journal of the History of the Behavioral*, 2(4):298–312, 1967.
- [IY10] Ryo Ishii and I. Nakano Yukiko. An Empirical Study of Eye-Gaze Behaviors: Towards the Estimation of Conversational Engagement in Human-Agent Communication. In *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction (EGIHMI'10)*, pp 33–40, Hong Kong, 2010. ACM.
- [Iza77] Carroll Ellis Izard. *Human Emotions*. Plenum Press, New York, 1977.
- [Jac92] S. W. Jackson. The Listening Healer in the History of Psychological Healing. *The American Journal of Psychiatry*, 149(12):1623–1632, 1992.
- [JL04] W.L. Johnson and C. LaBore. A Pedagogical Agent for Psychosocial Intervention on a Handheld Computer. In *Proceedings of the AAAI Fall Symposium on Dialogue Systems for Health Communication*, pp 22–24, 2004.
- [Jos66] Joseph Weizenbaum. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [KCC96] Robert M. Krauss, Yihsiu Chen, and Purnima Chawla. Nonverbal Behavior and Nonverbal Communication: What do Conversational Hand Gestures Tell Us? In Mark P. Zanna, editor, *Advances in Experimental*

Social Psychology, volume 28, pp 389–450, San Diego, CA, US, 1996. Academic Press.

- [KGWW08a] Sin-hwa Kang, Jonathan Gratch, Ning Wang, and J. Watt. Agreeable People Like Agreeable Virtual Humans. In *Proceedings of the Intelligent Virtual Agents (IVA '08)*, pp 253–261. Springer, 2008.
- [KGWW08b] Sin-hwa Kang, Jonathan Gratch, Ning Wang, and J.H. Watt. Does the Contingency of Agents' Nonverbal Feedback Affect Users' Social Anxiety? In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 1, pp 120–127. International Foundation for Autonomous Agents and Multiagent Systems, 2008.
- [Kip01] Michael Kipp. Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp 1367–1370, 2001.
- [Kip05] Michael Kipp. *Gesture Generation By Imitation: From Human Behavior To Computer Character Animation*. Universal Publishers, 2005.
- [Kip06] Michael Kipp. Creativity Meets Automation: Combining Nonverbal Action Authoring with Rules and Machine Learning. In Jonathan Gratch, editor, *Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA '06)*, pp 230–242, Marina Del Rey, CA, USA, 2006. Springer Berlin Heidelberg.
- [KKNNG06] Michael Kipp, Kerstin H Kipp, Alassane Ndiaye, and Patrick Gebhard. Evaluating the Tangible Interface and Virtual Characters in the Interactive COHIBIT Exhibit. In J. Gratch, editor, *Proceedings of the Intelligent Virtual Agents (IVA '06)*, pp 434–444. Springer-Verlag Berlin Heidelberg, 2006.
- [KL85] Neil H. Katz and John W. Lawyer. *Communication and Conflict Resolution Skills*. Kendall Hunt, Dubuque, Iowa, 1985.
- [KMR02] J. Klein, Y. Moon, and Rosalind W. Picard. This Computer Responds to User Frustration: Theory, Design, and Results. *Interacting with Computers*, 14(2):119–140, 2002.
- [KPG⁺07] Patrick Kenny, T. Parsons, Jonathan Gratch, Anton Leuski, and A. Rizzo. Virtual Patients for Clinical Therapist Skills Training. In

C. Pelachaud et Al., editor, *Proceedings of the Intelligent Virtual Agents (IVA '07)*, pp 197–210. Springer-Verlag Berlin Heidelberg, 2007.

- [KTB⁺08] François L. A. Knoppel, Almer S. Tigelaar, Danny Oude Bos, Thijs Alofs, and Zsófia Ruttkay. Trackside DEIRA: A Dynamic Engaging Intelligent Reporter Agent. In Padgham, Parkes, Müller, and Parsons, editors, *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp 112–119, Estoril, Portugal, 2008. International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).
- [KTRS12] Eva G. Krumhuber, Lucas Tamarit, Etienne B. Roesch, and Klaus R. Scherer. FACSGen 2.0 Animation Software: Generating Three-Dimensional FACS-Valid Facial Expressions for Emotion Research. *Emotion*, 12(2):351–363, January 2012.
- [KWG09] Sin-hwa Kang, James H. Watt, and Jonathan Gratch. Associations Between Interactants Personality Traits and Their Feelings of Rapport in Interactions with Virtual Humans. In *Proceedings of the Annual Meeting of the International Communication Association*, pp 1–25, Marriott, Chicago, IL, 2009.
- [Laf79] Marianne Lafrance. Nonverbal Synchrony Panel Technique: Analysis by the Cross-Lag and Rapport. *Social Psychology*, 42(1):66–70, 1979.
- [Laf82] Marianne Lafrance. Posture Mirroring and Rapport. *Counselling Psychology Quarterly*, 14(4):267–280, 1982.
- [Lan95] P. J. Lang. The Emotion Probe. Studies of Motivation and Attention. *The American Psychologist*, 50(5):372–85, May 1995.
- [LAYR13] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rische. I Can Help You Change! An Empathic Virtual Agent Delivers Behavior-Change Health Interventions. *ACM Transactions on Management Information Systems*, 4(4):1–28, 2013.
- [LB76] Marianne Lafrance and M. Broadbent. Group Rapport: Posture Sharing as a Nonverbal Indicator. *Group & Organization Management*, 1(3):328–333, September 1976.

- [LBAM04] Christine L. Lisetti, S. M. Brown S. M. Brown, K. Alvarez K. Alvarez, and A. H. Marpaung A. H. Marpaung. A Social Informatics Approach to Human-Robot Interaction with a Service Social Robot, 2004.
- [LBC97] P. J. Lang, Margaret M. Bradley, and B. N. Cuthbert. *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*, volume 77. The Center for Research in Psychophysiology, University of Florida, 1997.
- [LCK⁺10] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, Iain Matthews, and Forbes Ave. The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, number July, pages 94–101, San Francisco, CA, 2010. IEEE.
- [LDB⁺10] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and Validation of the Radboud Faces Database. *Cognition and Emotion*, 24:1377–1388, 2010.
- [Les87] Alan M. Leslie. Pretense and Representation: The Origins of "Theory of Mind.". *Psychological Review*, 94(4):412–426, 1987.
- [LFO98] D. Lundqvist, A. Flykt, and A. Öhman. The Karolinska Directed Emotional Faces - KDEF, 1998.
- [Lis08] Christine L. Lisetti. Embodied Conversational Agents for Psychotherapy. In *Proceedings of the CHI 2008 Conference Workshop on Technology in Mental Health*, pages 1–12. ACM, 2008.
- [LJC03] J. L. Lakin, VE Jefferis, and C.M. Cheng. The Chameleon Effect as Social Glue: Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior*, 27(3):145–162, 2003.
- [LM09] Jina Lee and Stacy C. Marsella. Learning a Model of Speaker Head Nods Using Gesture Corpora. In Decker, Sichman, Sierra, and Castelfranchi, editors, *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'09)*, Budapest, Hungary, 2009. International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

- [LM12] Zheng Li and Xia Mao. Emotional Eye Movement Generation Based on Geneva Emotion Wheel for Virtual Agents. *Journal of Visual Languages & Computing*, 23(5):299–310, October 2012.
- [LMMR97] W. Litvack-Miller, D. McDougall, and D. M. Romney. The Structure of Empathy During Middle Childhood and its Relationship to Prosocial Behavior. *Genetic, Social, and General Psychology Monographs*, 123(3):303–24, August 1997.
- [LMR06] Jina Lee, Stacy C. Marsella, and Marina Del Rey. Nonverbal Behavior Generator for Embodied Conversational Agents. In Jonathan Gratch, editor, *Proceedings of the 6th International Conference on Intelligent Virtual Agents (IVA'06)*, pp 243–255, Marina del Rey, 2006. Springer Berlin/Heidelberg.
- [LPNM09] Jina Lee, Helmut Prendinger, Alena Neviarouskaya, and Stacy Marsella. Learning Models of Speaker Head Nods with Affective Information. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction (ACII'09)*, pp 1–6. IEEE, 2009.
- [LSN10] Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-Time EEG-Based Human Emotion Recognition and Visualization. In *Proceedings of the International Conference on Cyberworlds (CW)*, pp 262–269, Singapore, October 2010. IEEE Computer Society.
- [LW08a] Christine L. Lisetti and Eric Wagner. Mental Health Promotion with Animated Characters: Exploring Issues and Potential. In *Proceedings of the AAAI Spring Symposium on Emotion, Personality and Social Behavior*, 2008.
- [LW08b] Christine L. Lisetti and Eric Wagner. Mental Health Promotion with Animated Characters: Exploring Issues and Potential. In *AAAI Spring Symposium*, 2008.
- [LYL⁺12] Christine L. Lisetti, Ugan Yasavur, Claudia De Leon, Reza Amini, and Naphtali Rische. Building an On-Demand Avatar-Based Health Intervention for Behavior Change. In *Proceeding of FLAIRS'12 Association for the Advancement of Artificial Intelligence (www.aaai.org)*, number Mi, Miami, FL, US, 2012.
- [MA99] M.J. Mendelson and F.E. Aboud. Measuring Friendship Quality in Late Adolescents and Young Adults: McGill Friendship Questionnaires.

Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement, 31(2):130, 1999.

- [MBF⁺90] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography*, 3(4):235–244, 1990.
- [McN92] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. Psychology/Cognitive Science. University of Chicago Press, 1992.
- [MDI⁺11] Brian Magerko, James Dean, Avinash Idnani, Michael Pantalon, and Gail D. Onofrio. Dr. Vicky: A Virtual Coach for Learning Brief Negotiated Interview Techniques for Treating Emergency Room Patients. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium*, pp 25–32. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.
- [MDK03] Maura E. Stokes, Charles S. Davis, and Gary G. Koch. *Categorical Data Analysis Using the SAS System*. SAS Institute and Wiley, 2nd edition, 2003.
- [MEM12] Christos N. Moridis and Anastasios A. Economides. Affective Learning: Empathetic Agents with Emotional Facial and Tone of Voice Expressions. *IEEE Transactions on Affective Computing*, 3(3):260–272, 2012.
- [MGR03] Stacy C. Marsella, Jonathan Gratch, and Jeff Rickel. Expressive Behaviors for Virtual Worlds. In M. Prendinger, H., Ishizuka, editor, *Life-Like Characters, Tools, Affective Functions, and Applications*. Springer, Heidelberg, 2003.
- [Mil88] B. Miller, W., Sovereign, R. and Krege. Motivational Interviewing with Problem Drinkers: The Drinker’s Check-Up as a Preventive Intervention. *Behavioural Psychotherapy*, 16:251–268, 1988.
- [Mil94] B. Miller, W., and Sanchez. Motivating Young Adults for Treatment and Lifestyle Change. In *Proceedings of the Issues in Alcohol Use and Misuse by Young Adults*, pp 55–81. 1994.
- [Mil95] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

- [ML09] Xia Mao and Zheng Li. Implementing Emotion-Based User-Aware E-Learning. *Science*, pp 3787–3792, 2009.
- [MM84] William R. Miller and G. Alan Marlatt. Brief Drinker Profile. *Psychological Assessment Resources*, Odessa, FL, 1984.
- [MNP13] Ruud Mattheij, Marie Nilsenova, and Eric Postma. Vocal and Facial Imitation of Humans Interacting with Virtual Agents. In *2013 Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp 815–820. IEEE, September 2013.
- [MPNP13] R. Mattheij, M. Postma-Nilsenová, and E. Postma. Mirror, Mirror in the Wall: Is there Mimicry in You All? *Journal of Ambient Intelligence and Smart Environments*, 1:1–5, 2013.
- [MR02] William R. Miller and Stephen Rollnick. *Motivational Interviewing: Preparing People for Change*, volume 2nd. Guilford Press, New York, 2nd edition, 2002.
- [MR09] William R. Miller and Gary S. Rose. Toward a Theory of Motivational Interviewing. *American Psychologist*, 64(6):527–537, 2009.
- [MR10] William R. Miller and Gary S. Rose. Toward a Theory of Motivational Interviewing. *American Psychologist*, 64(6):527–537, 2010.
- [MRP08] Scott W. McQuiggan, Jennifer Robison, and Robert Phillips. Modeling Parallel and Reactive Empathy in Virtual Agents: An Inductive Approach. In Müller Padgham, Parkes and Parsons, editors, *Proceedings of 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS’08)*, pp 167–174, Estoril, Portugal, 2008. International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).
- [MS94] B. H. Marcus and L. R. Simkin. The Transtheoretical Model: Applications to Exercise Behavior. *Medicine & Science in Sports & Exercise*, 26(11):1400–1404, 1994.
- [MSB93] D. A. Matthews, A. L. Suchman, and W. T. Branch. Making “Connections”: Enhancing the Therapeutic Potential of Patient-Clinician Relationships. *Annals of Internal Medicine*, 118(12):973–977, 1993.

- [MT83] R.E. Maurer and J.H. Tindall. Effect of Postural Congruence on Client's Perception of Counselor Empathy. *Journal of Counseling Psychology*, 30(2):158, 1983.
- [MT95] W.R. Miller and J.S. Tonigan. *The Drinker Inventory of Consequences (DrInC)*. 1995.
- [MW02] William R Miller and Paula L Wilbourne. Mesa Grande: a Methodological Analysis of Clinical Trials of Treatments for Alcohol Use Disorders. *Addiction*, 97(3):265–277, 2002.
- [Nat11] National Center for Chronic Prevention and Health Promotion. Excessive Alcohol Use At a Glance: Addressing a Leading Risk for Death, Chronic Disease, and Injury. Technical Report, National Center for Chronic Prevention and Health Promotion, 2011.
- [NBH07] Seth M. Noar, Christina N. Benac, and Melissa S. Harris. Does Tailoring Matter? Meta-Analytic Review of Tailored Print Health Behavior Change Interventions. *Psychological Bulletin*, 133(4):673–693, 2007.
- [NdRM10] Nicole Novielli, Fiorella de Rosis, and Irene Mazzotta. User Attitude Towards an Embodied Conversational Agent: Effects of the Interaction Mode. *Journal of Pragmatics*, 42(9):2385–2397, September 2010.
- [NI10] I. Nakano and Ryo Ishii. Estimating User's Engagement from Eye-Gaze Behaviors in Human-Agent Conversations. In ACM, editor, *Proceedings of the 15th International Conference on Intelligent User Interfaces*, pp 139–148, Hong Kong, China, 2010.
- [NK11] L. Neuhauser and G.L. Kreps. Participatory Design and Artificial Intelligence: Strategies to Improve Health Communication for Diverse Audiences. Cambridge, MA: AAAI Press, 2011.
- [NM09] H. Nguyen and Judith Masthoff. Designing Empathic Computers: the Effect of Multimodal Empathic Feedback Using Animated Agent. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive09)*, Claremont, California, USA, 2009. ACM.
- [Nol85] Patricia Noller. Video primacy? A Further Look. *Journal of Nonverbal Behavior*, 9(1):28–47, 1985.

- [NPAI06] Michael Nischt, Helmut Prendinger, Elisabeth André, and Mitsuru Ishizuka. MPML3D: A Reactive Framework for the Multimodal Presentation Markup Language. *Lecture Notes in Computer Science*, 62(1):218–229, 2006.
- [NPI07a] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Analysis of Affect Expressed Through the Evolving Language of Online Communication. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI'07)*, pp 278–281, Honolulu, Hawaii, USA, 2007. ACM Press.
- [NPI07b] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Textual Affect Sensing for Sociable and Expressive Online Communication. In Ana Paiva, R. Prada, and Rosalind W. Picard, editors, *Proceedings of the 2nd International Conference in Affective Computing and Intelligent Interaction (ACII)*, pp 220–231, Lisbon, Portugal, 2007. Springer-Verlag Berlin Heidelberg.
- [OCC88] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions*, volume 18. Cambridge University Press, Cambridge, UK, 1988.
- [OSP10] Magalie Ochs, David Sadek, and Catherine Pelachaud. A Formal Model of Emotions for an Empathic Rational Dialog Agent. *Autonomous Agents and Multi-Agent Systems*, November 2010.
- [Ost98] Jörn Ostermann. Animation of Synthetic Faces in MPEG-4, 1998.
- [Ost02] Jörn Ostermann. Face Animation in MPEG-4. In Igor Pandzic and Robert Forchheimer, editors, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, chapter 2, pages 17–55. Wiley, 2002.
- [PBA06] Helmut Prendinger and Christian Becker-Asano. A Study in User's Physiological Response to an Empathic Interface Agent. *International Journal of Humanoid*, 3(3):371–391, 2006.
- [PBCS10] S. G. Pulman, J. Boye, M. Cavazza, and Cameron Smith. 'How Was Your Day?'. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems (ACL'10)*, number July, pp 37–42, Uppsala, Sweden, 2010. Association for Computational Linguistics.

- [PBS96] Catherine Pelachaud, Norman I. Badler, and Mark Steedman. Generating Facial Expressions for Speech. *Cognitive Science*, 20(1):1–46, 1996.
- [PDP11] Jernej Polajnar, B. Dalvandi, and D. Polajnar. Does Empathy Between Artificial Agents Improve Agent Teamwork? In *Proceedings of the 10th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI'11)*, pp 96–102. IEEE, 2011.
- [PI05] Helmut Prendinger and M. Ishizuka. The Empathic Companion - A Character-Based Interface that Addresses Users' Affective States. *Applied Artificial Intelligence*, 19(3-4):267–286, 2005.
- [PKG09] Astrid M. Von Der Pütten, Nicole C. Krämer, and Jonathan Gratch. Who's There? Can a Virtual Agent Really Elicit Social Presence?, 2009.
- [PLM⁺11] A. Pereira, Iolanda Leite, Samuel Mascarenhas, Carlos Martinho, and Ana Paiva. Using Empathy to Improve Human-Robot Relationships. *Human-Robot Personal Relationships*, LNICST 59:130–138, 2011.
- [PP01] Stefano Pasquariello and Catherine Pelachaud. Greta: A Simple Facial Animation Engine. In R. Roy, editor, *Proceedings 6th Online World Conference on Soft Computing in Industrial Applications Session on Soft Computing for Intelligent 3D Agents*, pp 511–525. Springer-Verlag London, 2001.
- [Pre07] SD Preston. A perception-Action Model for Empathy. In T. Farrow and P. Woodruff, editors, *Empathy in Mental Illness*, chapter 23, pages 428–446. Cambridge University Press, 2007.
- [PS94] Scott Prevost and Mark Steedman. Specifying Intonation from Context for Speech Synthesis. *Speech Communication*, 15(1-2):18, 1994.
- [PSSJC08] David B. Portnoy, Lori A. J. Scott-Sheldon, Blair T. Johnson, and Michael P. Carey. Computer-Delivered Interventions for Health Promotion and Behavioral Risk Reduction: A Meta-Analysis of 75 Randomized Controlled Trials, 1988-2007. *Preventive medicine*, 47(1):3–16, July 2008.

- [PV97] J. O. Prochaska and W. F. Velicer. The Transtheoretical Model of Health Behavior Change. *American Journal of Health Promotion*, 12(1):38–48, 1997.
- [PW78] D. Premack and G. Woodruff. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*, 1(04):515–526, 1978.
- [Rab89] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [RE95] E. Rosenberg and Paul Ekman. Conceptual and Methodological Issues in the Judgment of Facial Expressions of Emotion. *Motivation and Emotion*, 19(2):111–138, 1995.
- [Res00] J. Resnicow, K., Soler, R., Braithwaite, R. L., Ahluwalia, J. S., & Butler. Cultural Sensitivity in Substance Use Prevention. *Journal of Community Psychology*, 28:271–290, 2000.
- [RM09] S.H. Rodrigues and S.F. Mascarenhas. “I Can Feel it too!: Emergent Empathic Reactions Between Synthetic Characters. In *Affective Computing and Intelligent Interaction and Workshops (ACII’09)*, pp 1–7, Porto, Portugal, 2009. IEEE.
- [Rog59] C. R. Rogers. A Theory of Therapy, Personality and Interpersonal Relationships as Developed in the Client-Centered Framework. In S. Koch, editor, *Psychology: the Study of a Science*, volume 3, chapter 3, pages 184–256. McGraw-Hill, New York, 1959.
- [RR08] Laurel D. Riek and Peter Robinson. Real-Time Empathy: Facial Mimicry on a Robot. In *Proceedings of the ACM Workshop on Affective Interaction in Natural Environments AFFINE at the International ACM Conference on Multimodal Interfaces (ICMI’08)*, pp 1–5, 2008.
- [SAD⁺06] Abdolhossein Sarrafzadeh, Samuel Alexander, Farhad Dadgostar, Chao Fan, and Abbas Bigdeli. See Me, Teach Me: Facial Expression and Gesture Recognition for Intelligent Tutoring Systems. *Proceedings of the 2006 Innovations in Information Technology*, pages 1–5, November 2006.

- [SbJS03] Marianne Sonnby-borgström, Peter Jonsson, and Owe Svensson. Emotional Empathy as Related to Mimicry Reactions at Different Levels of Information Processing. *Journal of Nonverbal*, 27(1):3–23, 2003.
- [SBS11] Daniel Schulman, Timothy Wallace Bickmore, and Candace L. Sidner. An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change Using Motivational Interviewing. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*, pp 61–64. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2011.
- [SCB⁺10] Cameron Smith, Nigel Crook, Johan Boye, Daniel Charlton, Simon Dobnik, David Pizzi, Marc Cavazza, and Stephen Pulman. Interaction Strategies for an Affective Conversational Agent. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA'10)*, pages 301–314. Springer-Verlag Berlin, Heidelberg, 2010.
- [Sch01] Klaus R. Scherer. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York, NY, US, 2001.
- [Sch05] Klaus R. Scherer. What Are Emotions? And How Can They Be Measured? *Social Science Information*, 44(4):695–729, 2005.
- [Sch11] B. Schneier. Empathy and Security. *Security & Privacy, IEEE*, 9(5):88–88, 2011.
- [SCW01] M.A. Sayette, Jeffrey F. Cohn, and J.M. Wertz. A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression. *Journal of Nonverbal Behavior*, 25(3):167–185, 2001.
- [SD12] Julian Szymaski and Wlodzislaw Duch. Information Retrieval with Semantic Memory Model. *Cognitive Systems Research*, 14(1):84–100, April 2012.
- [SE07] Klaus R. Scherer and Heiner Ellgring. Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion (Washington, D.C.)*, 7(1):158–71, February 2007.
- [SH04] Daniel D. Squires and Reid K. Hester. Using Technical Innovations in Clinical Practice: The Drinker’s Check-Up Software Program. *Journal of Clinical Psychology*, 60(2):159–69, February 2004.

- [Sha11] Ari Shapiro. Building a Character Animation System. *Motion in Games(MIG'11)*, LNCS(7060):98–109, 2011.
- [Ski94] H. A. Skinner. Computerized Lifestyle Assessment, Toronto: Multi-Health Systems, 1994.
- [SLN11] Olga Sourina, Y. Liu, and Minh Khoa Nguyen. Emotion-Enabled EEG-Based Interaction. In *SIGGRAPH Asia 2011 Posters on SA'11*, page 1, New York, New York, USA, 2011. ACM Press.
- [SMH83] T. Stockwell, D. Murphy, and R. Hodgson. The Severity of Alcohol Dependence Questionnaire: Its Use, Reliability and Validity. *British Journal of Addiction*, 78(2):145–155, 1983.
- [SPW⁺13] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp 1631–1642, 2013.
- [SS86] D. Servan-Schreiber. Artificial Intelligence and Psychiatry. *Journal of Nervous and Mental Disease*, 174:191–202, 1986.
- [SST02] Florian Schiel, Silke Steininger, and Ulrich Türk. The SmartKom Multimodal Corpus at BAS. In *Proceedings of the 3rd Conference on Language Resources and Evaluation LREC'02*, number 34, pp 200–206, 2002.
- [ST11] Simone G. Shamay-Tsoory. The Neural Bases for Empathy. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, 17(1):18–24, February 2011.
- [Stu06] Karsten Stueber. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. MIT Press, 1 edition, 2006.
- [Stu08] Karsten Stueber. Empathy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition, 2008.
- [SV04] Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: An Affective Extension of WordNet. In *Proceedings of LREC*, volume 4, pages 1083–1086. ELRA, Citeseer, 2004.

- [SW97] Christopher D. Shaw and Orion Wilson. *Methods And Apparatuses For Controlling Transformation Of Two And Three-Dimensional Images*, 1997.
- [SW00] Christopher D. Shaw and Orion Wilson. *Methods and Apparatuses for Controlling Transformation of Two and Three-Dimensional Images*, 2000.
- [Swa99] K. R. Swanson, A. J., Pantalon, M. V., & Cohen. Motivational Interviewing and Treatment Adherence Among Psychiatric and Dually Diagnosed Patients. *Journal of Nervous & Mental Disease*, 187:630–635, 1999.
- [TDR90] L. Tickle-Degnen and Robert Rosenthal. The Nature of Rapport and its Nonverbal Correlates. *Psychological Inquiry*, 1(4):285–293, 1990.
- [THSh+11] Markku Turunen, Jaakko Hakulinen, Olov Ståhl, Björn Gambäck, Preben Hansen, Mari C. Rodríguez Gancedo, Raúl Santos de la Cámara, Cameron Smith, Daniel Charlton, and Marc Cavazza. Multimodal and Mobile Conversational Health and Fitness Companions. *Computer Speech & Language*, 25(2):192–209, April 2011.
- [TKMS03] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-Rich Part-Of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the HLT-NAACL*, pp 252–259, 2003.
- [TRM+08] Marcus Thiebaux, Marina Rey, Andrew N. Marshall, Stacy Marsella, and Marcelo Kallmann. SmartBody: Behavior Realization for Embodied Conversational Agents. In Müller Padgham, Parkes and Parsons, editors, *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS’08)*, pp 12–16, Estoril, Portugal, 2008. International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).
- [TRS09b] Jessica L. Tracy, Richard W. Robins, and Roberta A. Schriber. Development of a FACS-Verified Set of Basic and Self-Conscious Emotion Expressions. *Emotion Washington Dc*, 9(4):554–559, 2009.
- [Tur95] S. Turkle. *Life on the Screen*. Simon & Schuster, New York, 1995.

- [TW11] Yaniv Taigman and Lior Wolf. Leveraging Billions of Faces to Overcome Performance Barriers in Unconstrained Face Recognition. *Arxiv Preprint ArXiv:1108.1122*, 1(view 2):1–7, 2011.
- [UPI08] Sebastian Ullrich, Helmut Prendinger, and Mitsuru Ishizuka. MPML3D: Agent Authoring Language for Virtual Worlds. In *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology (ACE '08)*, pp 134–137, Yokohama, Japan, 2008. ACM.
- [Van03] Lyn M. Van Swol. The Effects of Nonverbal Mirroring on Perceived Persuasiveness, Agreement with an Imitator, and Reciprocity in a Group Discussion. *Communication Research*, 30(4):461–480, August 2003.
- [VbHKK04] Rick B. Van baaren, Rob W. Holland, Kerry Kawakami, and Ad Van Knippenberg. Mimicry and Prosocial Behavior. *Psychological Science*, 15(1):71–74, January 2004.
- [vBY05] A. van Breemen and X. Yan. iCat: An Animated User-Interface Robot with Personality. *Proceedings of the fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'05)*, pp 143–144, 2005.
- [vdSHFD11] J. van der Schalk, S. T. Hawk, A. H. Fischer, and B. J. Doosje. Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, 11:907–920, 2011.
- [Ver10] M.L. Vernon. A Review of Computer-Based Alcohol Problem Services Designed for the General Public. *Journal of Substance Abuse Treatment*, 38(3):203–211, 2010.
- [VF10] C. Emilio Vargas and Debora Field. 'How Was Your Day? An Architecture for Multimodal ECA Systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIG-DIAL'10)*, pp 47–50, The University of Tokyo, 2010. Association for Computational Linguistics.
- [VHFD11] Job Van Der Schalk, Skyler T. Hawk, Agneta H. Fischer, and Bertjan Doosje. Moving Faces, Looking Places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion Washington Dc*, 11(4):907–920, 2011.

- [VP10] M. Valstar and Maja Pantic. Induced Disgust, Happiness and Surprise: An Addition to the MMI Facial Expression Database. In *Proceedings of International Conference on Language Resources and Evaluation, Workshop on EMOTION*, pp 65–70, Malta, 2010.
- [VS09] S. Villagraza and A. Susín Sánchez. Face! 3d Facial Animation System Based on FACS. In O. Rodríguez, F. Serón, R. Joan-Arinyo, J. Madeiras, J. Rodríguez, and E. Coto, editors, In *Proceedings of the American Symposium in Computer Graphics (SIACG'09)*, pp 203–209, Isla Margarita, 2009.
- [Wal95] H. G. Wallbott. Mutualities in Dialogue. In I. Markova, C. F. Graumann, and K. Foppa, editors, *In 1st Eds*, volume Cambridge, Chapter Congruence, pp 82–98. Cambridge University Press, 1995.
- [WG09] Ning Wang and Jonathan Gratch. Rapport and Facial Expression. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009*, pp 1–6, Amsterdam, September 2009. IEEE.
- [WG10] Ning Wang and Jonathan Gratch. Don't Just Stare at Me! In *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems (CHI'10)*, pages 1241–1249, Atlanta, GA, USA, 2010. ACM.
- [WH97] Jane Webster and Hayes Ho. Audience Engagement in Multimedia Presentations. *ACM SIGMIS Database*, 28(2):63–77, April 1997.
- [WHO11] WHO World Health Organization. Obesity and Overweight, 2011.
- [WHT10] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective Unconstrained Face Recognition by Combining Multiple Descriptors and Learned Background Statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1–13, December 2010.
- [Wis87] L. Wispé. History of the Concept of Empathy. In Nancy Einsenber and Janet Strayer, editors, *Empathy and its Development*, chapter 2, pages 17–37. Cambridge University Press, 1987.
- [WKP⁺03] Bruno Wicker, Christian Keysers, Jane Plailly, Jean Pierre Royet, Vittorio Gallese, and Giacomo Rizzolatti. Both of Us Disgusted in My Insula: the Common Neural Basis of Seeing and Feeling Disgust. *Neuron*, 40(3):655–64, 2003.

- [WKS⁺10] Angela White, David Kavanagh, Helen Stallman, Britt Klein, Frances Kay-Lambkin, Judy Proudfoot, Judy Drennan, Jason Connor, Amanda Baker, Emily Hines, and Ross Young. Online Alcohol Interventions: A Systematic Review. *Journal of Medical Internet Research*, 12(5):e62, January 2010.
- [WMS⁺87] Rebecca M. Warner, Daniel Malloy, Kathy Schneider, Russell Knoth, and Bruce Wilder. Rhythmic Organization of Social Interaction and Observer Ratings of Positive Affect and Involvement. *Journal of Non-verbal Behavior*, 11(2):57–74, 1987.
- [WP93] Michael Wierzbicki and Gene Pekarik. A Meta-Analysis of Psychotherapy Dropout. *Professional Psychology Research and Practice*, 24(2):190–195, 1993.
- [WR05] A. Wojde and L.J.M. Rothkrantz. Parametric Generation of Facial Expressions Based on FACS. *Computer Graphics Forum*, 24(4):743–757, 2005.
- [WS96] Jack W. Wiley and Christopher D. Shaw. Tactile Interface Apparatus For Providing Physical Feedback To A User Based On An Interaction With A Virtual Environment, 1996.
- [WT00] Nigel Ward and Wataru Tsukahara. Prosodic Features Which Cue Back-Channel Responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.
- [YAL12] Ugan Yasavur, Reza Amini, and Christine L. Lisetti. User Modeling for Pervasive Alcohol Intervention Systems. In *Proceedings of the First International Workshop on Recommendation Technologies for Lifestyle Change 2012 (LIFESTYLE'12)*, pp 29–34, Dublin, Ireland, 2012.
- [YALR13] Ugan Yasavur, Reza Amini, Christine Lisetti, and Naphtali Rishe. Ontology-Based Named Entity Recognizer for Behavioral Health. In *Proceedings of the 26th International FLAIRS Conference*, St Petersburg, FL, USA, 2013. AAAI Press.
- [ZMI89] R. B. Zajonc, S. T. Murphy, and M. Inglehart. Feeling and Facial Expression: Implications of the Vascular Theory of Emotion. *Psychological review*, 96(3):395–416, July 1989.

- [ZWRE92] Carolyn Zahn-Waxler, JoAnn L. Robinson, and Robert N. Emde. The Development of Empathy in Twins. *Developmental Psychology*, 28(6):1038–1047, 1992.

VITA

REZA AMINI

February 23, 1983	Born, Isfahan, Iran
2005	B.A., Biomedical Engineering University of Isfahan Isfahan, Iran
2009	M.S., Electrical Engineering (Control Eng.) Isfahan University of Technology Isfahan, Iran
2012	M.S., Computer Science Florida International University Miami, Florida
2015	Ph.D., Computer Science Florida International University Miami, Florida

PUBLICATIONS AND PRESENTATIONS

R. Amini, C. Lisetti, G. Ruiz, (2015). HapFACS 3.0: Open-Source FACS-Based Facial Expression Generator for 3D Speaking Virtual Characters. *IEEE Transactions on Affective Computing*. (accepted with minor revisions)

C. Lisetti, R. Amini, U. Yasavur, (2015). Now All Together: Overview of Virtual Health Assistants Emulating Face-to-Face Health Interview Experience. *KI-Künstliche Intelligenz Journal*, Springer Berlin Heidelberg, pp 1–12.

R. Amini, C. Lisetti, (2015). Survey: Empathy in Computer Systems. *IEEE Transactions on Affective Computing*. (under review)

R. Amini, C. Lisetti, and U. Yasavur (2014). Emotionally Responsive Virtual Counselor for Behavior-Change Health Interventions. In *Proceedings of the Design Science Research in Information Systems and Technologies (DESRIST 2014)*, Advancing the Impact of Design Science: Moving from Theory to Practice, Lecture Notes in Computer Science Volume 8463, pp 433–437, (Miami, USA, May 2014).

C. L. Lisetti, R. Amini, U. Yasavur (2013). I Can Help You Change! An Empathic Virtual Agent Delivers Behavior Change Health Interventions. *ACM Transactions*

on Management Information Systems, Vol. 4, No. 4, Article 19, 2013.

U. Yasavur, R. Amini, C. L. Lisetti (2013). Ontology-Based Named Entity Recognizer for Behavioral Health. In *Proceedings of the 26th International FLAIRS Conference*, pp 249–254, (St. Pete Beach, USA, May 2013).

R. Amini, C.L. Lisetti, U. Yasavur, N. Rishe (2013). On-Demand Virtual Health Counselor for Delivering Behavior-Change Health Interventions. In *Proceedings of the IEEE International Conference on Healthcare Informatics*, pp 46–55, (Philadelphia, USA, September 2013).

R. Amini, C. Lisetti, (2013). HapFACS: an Open Source API/Software to Generate FACS-Based Expressions for ECAs Animation and for Corpus Generation. In *Proceedings of the 5th Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII'13)*, IEEE Computer Society, pp 270–275, (Geneva, SWITZERLAND, September, 2013).

R. Amini, U. Yasavur, C. L. Lisetti (2012). HapFACS 1.0: Software/API for Generating FACS-Based Facial Expressions. In *Proceedings of the ACM 3rd International Symposium on Facial Analysis and Animation (FAA'12)*, Article 17, (Vienna, AUSTRIA, September, 2012).

U. Yasavur, R. Amini, C. L. Lisetti (2012). User Modeling for Pervasive Alcohol Intervention Systems. In *Proceedings of the Workshop on Recommendation Technologies for Lifestyle Change, In conjunction with the 6th ACM Conference on Recommender Systems (RecSys'12)* pp 29–34, , (Dublin, IRELAND, September 2012).

C. Lisetti, U. Yasavur, C. De Leon, R. Amini, U. Visser, N. Rishe. Building On-demand Avatar-based Health Intervention for Behavior Change. In *Proceedings of the 25th International FLAIRS Conference*, pp 175–180, (Marco Island, USA, May 2012).