Ruth Aylett
Brigitte Krenn
Catherine Pelachaud
Hiroshi Shimodaira (Eds.)

LNAI 8108

Intelligent Virtual Agents

13th International Conference, IVA 2013
Edinburgh, UK, August 2013
Proceedings

Springer

Lecture Notes in Artificial Intelligence 8108

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

LNAI Founding Series Editor

Ruth Aylett Brigitte Krenn
Catherine Pelachaud Hiroshi Shimodaira (Eds.)

Intelligent Virtual Agents

13th International Conference, IVA 2013
Edinburgh, UK, August 29-31, 2013
Proceedings

Springer

Volume Editors

Ruth Aylett
Heriot-Watt University
Edinburgh, UK
E-mail: r.s.aylett@hw.ac.uk

Brigitte Krenn
Austrian Research Institute for Artificial Intelligence (OFAI)
Vienna, Austria
E-mail: brigitte.krenn@ofai.at

Catherine Pelachaud
CNRS-LTCI, Télécom-ParisTech
Paris, France
E-mail: catherine.pelachaud@telecom-paristech.fr

Hiroshi Shimodaira
The University of Edinburgh
Edinburgh, UK
E-mail: h.shimodaira@ed.ac.uk

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-40414-6 e-ISBN 978-3-642-40415-3
DOI 10.1007/978-3-642-40415-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013945590

CR Subject Classification (1998): I.2.11, I.2, H.5, H.4, H.3, I.6

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Welcome to the proceedings of the 13th International Conference on Intelligent Virtual Agents. While this conference represents a field of specialization with in computer science and artificial intelligence, it celebrates an endeavor that requires the integration of knowledge, methodologies, and theories from a wide range of fields such as sociology, psychology, linguistic, cognitive science, and interactive media.

Intelligent virtual agents are animated characters that not only move, but also exhibit human-like competence when dealing with the world around them, be it virtual or real. In particular these agents communicate with humans or with each other using natural human modalities such as speech and gesture. They are capable of real-time perception, cognition, and action that allows them to participate autonomously in dynamic social environments. Intelligent virtual agents are not built overnight or by lone practitioners. These are complex systems, built layer by layer, integrating numerous components that address important functions such as visual object tracking, speech recognition, perceptual memory, language understanding, reactive behavior, reasoning, planning, action scheduling, and articulation. Advances are made by sharing knowledge, components, and techniques. Therefore the annual IVA conference is central to advancing the state of the art. It is an interdisciplinary forum for presenting research on modeling, developing, and evaluating IVAs with a focus on communicative abilities and social behavior.

IVA was started in 1998 as a workshop at the European Conference on Artificial Intelligence on Intelligent Virtual Environments in Brighton, UK, which was followed by a similar one in 1999 in Salford, Manchester. Then dedicated stand alone IVA conferences took place in Madrid, Spain, in 2001, Irsee, Germany,in 2003, and Kos, Greece, in 2005. Since 2006, IVA has become a full-fledged annual international event, which was first held in Marina del Rey, California,then Paris, France, in 2007, Tokyo, Japan, in 2008, Amsterdam, The Netherlands,in 2009, Philadelphia, Pennsylvania, in 2010, Reykjavik, Iceland, in 2011 and Santa Cruz, USA, in 2012. Since 2005, IVA has also hosted the Gathering of Animated Lifelike Agents (GALA), a festival to showcase state-of-the-art agents created by students, academic or industrial research groups. This year's conference in Edinburgh, Scotland, represented a range of expertise,from different scientific and artistic disciplines, and highlighted the value of both theoretical and practical work needed to bring intelligent virtual agents to life.

The special topic of IVA 2013 was cognitive modelling in virtual agents. This topic touches on many aspects of intelligent virtual agent theory and application such as models of personality; theory of mind; learning and adaptation; motivation and goal-management; creativity; social and culturally specific behavior. Several papers deal directly with these topics. The remaining papers cover important themes linked to the design, modelling, and evaluation of IVAs as well as system implementation and applications of IVAs.

IVA 2013 received 94 submissions. Out of the 61 long paper submissions,only 18 were accepted for the long papers track. Furthermore, therewere 18 short papers presented in the single-track paper session, and 34 poster papers were on display. Since IVA 2011, the proceedings are distributed only digitally.

This year's IVA also included four workshops that focused on "Computers as Social Actors," "Cultural Characters In Games and Learning", "Multimodal Corpora: Beyond Audio and Video," "Techniques Towards Companion Technologies." There was also a Doctoral Consortium where PhD students receive feedback from peers and established researchers.

IVA 2013 was locally organized by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, with the generous support of the School of Mathematical & Computer Sciences at Heriot-Watt University, UK; the Interaction Technologies Austrian Research Institute for Artificial Intelligence (OFAI), Austria; and CNRS – LTCI at Télécom-ParisTech, France.

We would like to wholeheartedly thank the scientific committees that helped shape a quality conference program, the Senior Program Committee for taking on great responsibility and the Program Committee for their time and genuine effort. We also want to thank our keynote speakers Jacqueline Nadel, from the Centre Emotion, La Salpétrière Hospital, Paris, France, Charles Sutton, University of Edinburgh, UK, Steve Holmes, vice president of the Mobile and Communications Group, Intel, USA, and Alessandro Vinciarelli, University of Glasgow, UK, for crossing domains and sharing their insights with us. Furthermore, we would like to express great appreciation for the work put in by Matthew Aylett who oversaw the poster and demo session, by Jonas Beskow, who coordinated the workshops, by Ana Paiva, who was responsible for the publicity of the conference, by Lynne Hall, who managed the Doctoral Consortium, by Magalie Ochs, who organized GALA, and by our tireless student volunteers and their coordinator David A. Braude, who kept everything running smoothly. We are grateful to Peter Bell and Atef Ben-Youssef for website design and the timely conference system support from Hiroshi Shimodaira. Atef Ben-Youssef helped to assemble the proceedings book. Finally, we would like to express deep gratitude to Hiroshi Shimodaira, Avril Heron, Samira Reuter, and Nicola Drago-Ferrante, who managed everything from registration and financials to decoration and local travel logistics.

Of course IVA 2013 would not have been possible without the valuable contributionsof the authors, whose dedication extends beyond the creation of intelligentvirtual agents to the creation and support of a vibrant research communitythat nurtures our passion for the field.

September 2013

Ruth Aylett
Brigitte Krenn
Catherine Pelachaud
Hiroshi Shimodaira

Organization

Committees

Conference Co-chairs

Ruth Aylett	Heriot-Watt University, UK
Brigitte Krenn	OFAI, Austria
Catherine Pelachaud	CNRS-LTCI, Telecom-ParisTech, France
Hiroshi Shimodaira	University of Edinburgh, UK

Demonstrations

Matthew Aylett	University of Edinburgh, UK

Workshops Chair

Jonas Beskow	KTH, Sweden

Publicity Chair

Ana Paiva	INESC-ID, Portugal

Doctoral Consortium Chair

Lynne Hall	University of Sunderland, UK

GALA Chair

Magalie Ochs	CNRS-LTCI, Telecom-ParisTech, France

Website

Peter Bell	University of Edinburgh, UK
Atef Ben Youssef	University of Edinburgh, UK

Finance and Local Arrangements

Hiroshi Shimodaira	University of Edinburgh, UK
Nicola Drago-Ferrante	University of Edinburgh, UK
Avril Heron	University of Edinburgh, UK
Samira Reuter	University of Edinburgh, UK

Senior Program Committee

Jan Allbeck	George Mason University, USA
Elisabeth André	Augsburg University, Germany

Tim Bickmore	Northeastern University, USA
Cristina Conati	University of British Columbia, Canada
Dirk Heylen	University of Twente, The Netherlands
Michael Kipp	University of Applied Sciences Augsburg, Germany
Stefan Kopp	Bielefeld University, Germany
Brian Magerko	Georgia Institute of Technology, USA
Stacy Marsella	University of Southern California, USA
Jean-Claude Martin	LIMSI-CNRS, France
Yukiko I. Nakano	Seikei University, Japan
Michael Neff	University of California - Davis, USA
Reid Swanson	University of Southern California, USA
Hannes Vilhjalmsson	Reykjavk University, Iceland
Marilyn Walker	University of California - Santa Cruz, USA
R. Michael Young	North Carolina State University, USA

Program Committee

Samer Al Moubayed	Royal Institute of Technology, Sweden
Matthew Aylett	University of Edinburgh, UK
Sabarish Babu	Clemson University, USA
Atef Ben-Youssef	University of Edinburgh, UK
Jonas Beskow	Royal Institute of Technology, Sweden
Elisabetta Bevacqua	Lab-STICC, ENIB, France
Bridget Cooper	University of Sunderland, UK
Ginevra Castellano	University of Birmingham, UK
Mohamed Chetouani	University of Pierre and Marie Curie, France
Matthieu Courgeon	Lab-STICC, France
Carlos Delgado-Mata	Universidad Panamericana, Mexico
Joao Dias	INESC-ID and IST, Portugal
Jens Edlund	Royal Institute of Technology, Sweden
Victoria Eyharabide	Paris-Sorbonne University, France
Björn Granström	Royal Institute of Technology, Sweden
Agneta Gulz	University of Linköping, Sweden
Joakim Gustafson	Royal Institute of Technology, Sweden
Lynne Hall	University of Sunderland, UK
Helen Hastie	Heriot-Watt University, UK
Anna Hjalmarsson	KTH, Sweden
Colette Hume	University of Sunderland, UK
Lynne Humphries	University of Sunderland, UK
Jesús Ibánez	UPF Barcelona, Spain
Amy Isard	University of Edinburgh, UK
Susan Jones	University of Sunderland, UK
Sophie Joerg	Clemson University, USA
H. Chad Lane	University of Southern California, USA

Jina Lee	Sandia National Labs, USA
Margaux Lhommet	University of Southern California, USA
Sandy Louchart	Heriot-Watt University, UK
Maurizio Mancini	University of Genoa, Italy
Joshua McCoy	University of California Santa Cruz, USA
Kasia Muldner	Arizona State University, USA
Asad Nazir	Heriot-Watt University, UK
Radoslaw Niewiadomski	Telecom ParisTech, France
Mohammad Obaid	University of Canterbury, New Zealand
Magalie Ochs	CNRS-LTCI, Telecom-ParisTech, France
Sabine Payr	OFAI, Austria
Christopher Peters	Royal Institute of Technology, Sweden
Paolo Petta	OFAI, Austria
Ronald Poppe	University of Twente, The Netherlands
Rui Prada	INESC-ID and Instituto Superior Tecnico-Technical University of Lisbon, Portugal
Ken Prepin	CNRS-LTCI, Telecom-ParisTech, France
Stefan Rank	OFAI, Austria
Laurel Riek	University of Notre Dame, USA
Astrid Marieke Rosenthal-von der Pütten	University of Duisburg-Essen, Germany
Isaac Rudomin	Barcelona Supercomputing Center (BSC), Spain
Nicolas Sabouret	LIMSI-CNRS, France
Stefan Scherer	University of Southern California, USA
Daniel Schulman	Veterans Health Administration, USA
Mei Si	Rensselaer Polytechnic Institute, USA
Gabriel Skantze	Royal Institute of Technology, Sweden
Marcin Skowron	OFAI, Austria
Sarah Tazzyman	University of Sunderland, UK
Laura Vardoulakis	Northeastern University, USA
Hannes Vilhjalmsson	Reykjavik University, Iceland
Preben Wik	Royal Institute of Technology, Sweden
Arlette van Wissen	VU University Amsterdam, The Netherlands

Reviewers

Angelo Cafaro	Reykjavik University, Iceland
Hazaël Jones	University Pierre and Marie Curie, France

Sponsoring Institutions

Scottish Informatics and Computer Science Alliance

Table of Contents

Cognitive Models

Using a Parameterized Memory Model to Modulate NPC AI 1
Weizi (Philip) Li, Tim Balint, and Jan M. Allbeck

DyBaNeM: Bayesian Episodic Memory Framework for Intelligent Virtual Agents 15
Rudolf Kadlec and Cyril Brom

Self-organizing Cognitive Models for Virtual Agents 29
Yilin Kang and Ah-Hwee Tan

Are You Thinking What I'm Thinking? An Evaluation of a Simplified Theory of Mind 44
David V. Pynadath, Ning Wang, and Stacy C. Marsella

Censys: A Model for Distributed Embodied Cognition 58
Tiago Ribeiro, Marco Vala, and Ana Paiva

Applications

Automated Promotion of Technology Acceptance by Clinicians Using Relational Agents 68
Timothy Bickmore, Laura Vardoulakis, Brian Jack, and Michael Paasche-Orlow

Virtual Agents as Daily Assistants for Elderly or Cognitively Impaired People: Studies on Acceptance and Interaction Feasibility 79
Ramin Yaghoubzadeh, Marcel Kramer, Karola Pitsch, and Stefan Kopp

Modeling Brief Alcohol Intervention Dialogue with MDPs for Delivery by ECAs 92
Ugan Yasavur, Christine Lisetti, and Napthali Rishe

Gendering the Machine: Preferred Virtual Assistant Gender and Realism in Self-Service 106
Jeunese Payne, Andrea Szymkowiak, Paul Robertson, and Graham Johnson

Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training 116
Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer

A Question-Answering Agent Using Speech Driven Non-linear Machinima 129
Anuj Tewari, Timothy Brown, and John Canny

A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance 139
Kirsten Bergmann and Manuela Macedonia

Using Linguistic Alignment to Enhance Learning Experience with Pedagogical Agents: The Special Case of Dialect 149
Vivien Kühne, Astrid Marieke Rosenthal-von der Pütten, and Nicole C. Krämer

A Virtual Patient to Assess Pediatric Intensive Care Unit (PICU) Nurses' Pain Assessment and Intervention Practices 159
Sangyoon Lee, Cynthia M. LaFond, Andrew E. Johnson, Catherine Vincent, Jason Leigh, and Luc Renambot

Dialogue, Language, Speech

Controlling the Listener Response Rate of Virtual Agents 168
Iwan de Kok and Dirk Heylen

The Influence of Prosody on the Requirements for Gesture-Text Alignment 180
Yingying Wang and Michael Neff

Prediction of Visual Backchannels in the Absence of Visual Context Using Mutual Influence 189
Derya Ozkan and Louis-Philippe Morency

Modeling the Semantic Coordination of Speech and Gesture under Cognitive and Linguistic Constraints 203
Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp

Modeling Multimodal Behaviors from Speech Prosody 217
Yu Ding, Catherine Pelachaud, and Thierry Artières

An Examination of Whether People Prefer Agents Whose Gestures Mimic Their Own 229
Pengcheng Luo, Victor Ng-Thow-Hing, and Michael Neff

Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor 239
Klaus Förger, Timo Honkela, and Tapio Takala

Non-verbal Behaviour

Conversational Gaze Aversion for Virtual Agents 249
Sean Andrist, Bilge Mutlu, and Michael Gleicher

From a User-Created Corpus of Virtual Agent's Non-verbal Behavior to a Computational Model of Interpersonal Attitudes 263
Brian Ravenet, Magalie Ochs, and Catherine Pelachaud

Explaining the Variability of Human Nonverbal Behaviors in Face-to-Face Interaction 275
Lixing Huang and Jonathan Gratch

Head Motion Analysis and Synthesis over Different Tasks 285
Atef Ben Youssef, Hiroshi Shimodaira, and David A. Braude

Using Virtual Agents to Guide Attention in Multi-task Scenarios 295
Philipp Kulms and Stefan Kopp

Gesture with Meaning 303
Margaux Lhommet and Stacy C. Marsella

Social, Cultural Models and Agents

Conceptualizing Social Power for Agents 313
Gonçalo Pereira, Rui Prada, and Pedro A. Santos

Social Importance Dynamics: A Model for Culturally-Adaptive Agents 325
Samuel Mascarenhas, Rui Prada, Ana Paiva, and Gert Jan Hofstede

Looking Real and Making Mistakes 339
Yuqiong Wang, Peter Khooshabeh, and Jonathan Gratch

What's Going on? Multi-sense Attention for Virtual Agents 349
Tim Balint and Jan M. Allbeck

A Qualitative Evaluation of Social Support by an Empathic Agent 358
Janneke M. van der Zwaan, Virginia Dignum, and Catholijn M. Jonker

Tools and Techniques

All Together Now: Introducing the Virtual Human Toolkit 368
Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch

Real-Time Behavioral Animation of Humanoid Non-Player Characters with a Computational Ecosystem ... 382
Rui Filipe Antunes and Frederic Fol Leymarie

Judging IVA Personality Using an Open-Ended Question ... 396
Kris Liu, Jackson Tolins, Jean E. Fox Tree, Marilyn Walker, and Michael Neff

An Affective Virtual Agent Providing Embodied Feedback in the Paired Associate Task: System Design and Evaluation ... 406
Christian Becker-Asano, Philip Stahl, Marco Ragni, Matthieu Courgeon, Jean-Claude Martin, and Bernhard Nebel

Posters

Expressive Animation: Towards a Procedural Solution ... 417
Ewan McGregor and Robin J.S. Sloan

Interactive Machine Learning for Virtual Agents ... 419
Marco Gillies and Andrea Kleinsmith

An Eye Movement Model for the Purpose of Emotional Expression ... 421
Chen Jiang, Jun Yu, and Zengfu Wang

Effects of Users' Social Skill on Evaluations of a Virtual Agent That Exhibits Self-adaptors ... 423
Tomoko Koda and Hiroshi Higashino

Activity Planning for Long-Term Relationships (Extended Abstract) ... 425
William Coon, Charles Rich, and Candace L. Sidner

Realistic Facial Animation Driven by a Single-Camera Video ... 428
Changwei Luo, Jun Yu, and Zengfu Wang

The INGREDIBLE Database: A First Step Toward Dynamic Coupling in Human-Virtual Agent Body Interaction ... 430
Igor Stanković, Pierre De Loor, Virginie Demulier, Alexis Nédélec, and Elisabetta Bevacqua

OpenBMLParser: An Open Source BML Parser/Analyzer ... 432
Herwin van Welbergen

A Companion Robot That Can Tell Stories ... 434
Carole Adam and Lawrence Cavedon

Towards Mapping and Segmentation of Very Large Scale Spaces with Intelligent Virtual Agents ... 436
Katrina Samperi, Nick Hawes, and Russell Beale

A Character Agent System for Promoting Service-Minded Communication 438
Kaoru Sumi and Ryuji Ebata

A Speech-Enabled Intelligent Agent for Pedestrian Navigation and Tourist Information 440
Srinivasan Janarthanam, Oliver Lemon, Phil Bartie, Tiphaine Dalmas, Xingkun Liu, William Mackaness, and Bonnie Webber

A Question-Answering Game for Preschool Children 442
Anuj Tewari, Peiyi Ko, and John Canny

A Software Framework for Social Cue-Based Interaction with a Virtual Recruiter 444
Ionut Damian, Tobias Baur, Patrick Gebhard, Kaśka Porayska-Pomsta, and Elisabeth André

Size Certainly Matters – At Least If You Are a Gesticulating Digital Character: The Impact of Gesture Amplitude on Addressees Information Uptake 446
Mette Clausen-Bruun, Tobias Ek, and Magnus Haake

A Procedural Approach to Simulate Virtual Agents Behaviors in Indoor Environments 448
Laura Mattos Flach, Vinícius Jurinic Cassol, Fernando Pinho Marson, and Soraia Raupp Musse

Animating Mixed Emotions: Can Basic Emotion Blends Be Recognized from a Virtual Face? 450
Meeri Mäkäräinen, Jari Kätsyri, and Tapio Takala

A Thousand Words Paint a Picture - Examining Interviewer Effects with Virtual Agents 452
Jennifer Klatt, Astrid Marieke Rosenthal-von der Pütten, Laura Hoffmann, and Nicole C. Krämer

Deep Conversations on a Flat Screen? A Comparison of Virtual Agents in Videos and Immersive Virtual Environments 455
Jennifer Klatt and Nicole C. Krämer

yaPOSH Action Selection 458
Jakub Gemrot, Jan Havlíček, Michal Bída, Rudolf Kadlec, and Cyril Brom

Social Identity Bias in Agents' Rational Decision 460
Joana Dimas, Phil Lopes, Guida Preto, Pedro A. Santos, and Rui Prada

Tell Me a Story: A Comparative Study of Child Interaction with Virtual Characters and Adults 462
Ovidiu Şerban, Alexandre Pauchet, Anne Bersoult, and Émilie Chanoni

A Dynamic Persuasive Dialogue Model for Encouraging Social Interaction for Older Adults 464
John Paul Vargheese, Somayajulu Sripada, Judith Masthoff, Nir Oren, and Matt Dennis

The University of Edinburgh Head-Motion and Audio Storytelling (UoE - HAS) Dataset 466
David A. Braude, Hiroshi Shimodaira, and Atef Ben Youssef

Incremental, Adaptive and Interruptive Speech Realization for Fluent Conversation with ECAs 468
Herwin van Welbergen, Timo Baumann, Stefan Kopp, and David Schlangen

The Effect of Multiple Modalities on the Perception of a Listening Agent 470
Ronald Poppe, Mark ter Maat, and Dirk Heylen

Who Is the Werewolf? 472
Mei Yii Lim, Ruth Aylett, João Dias, Birgit Endrass, Eva Krumhuber, Gert Jan Hofstede, and Lynne Hall

Evaluating the Impact of ECAs on User Navigation Performance and Perception 474
Ioannis Doumanis and Adalet Serengul Smith

Modelling Social Power Intelligent Agents 476
Gonçalo Pereira, Rui Prada, and Pedro A. Santos

Evaluating the Accessibility of ECA-Based Information Presentation Systems 478
Ioannis Doumanis and Adalet Serengul Smith

Finding the Timings for a Guide Agent to Intervene User-User Conversation to Provide Information Actively 482
Shochi Otogi, Hung-Hsuan Huang, Ryo Hotta, and Kyoji Kawagoe

Gestural Adaptation in Extravert-Introvert Pairs and Implications for IVAs 484
Jackson Tolins, Kris Liu, Yingying Wang, Jean E. Fox Tree, Marilyn Walker, and Michael Neff

Web-Enabled 3D Talking Avatars Based on WebGL and HTML5 486
Jonas Beskow and Kalin Stefanov

Deus-Ex - Interactive Machinima Movies for Cognitive Behavioral Therapy 488
Pablo Paredes, Anuj Tewari, and John Canny

Author Index 491

Using a Parameterized Memory Model to Modulate NPC AI

Weizi (Philip) Li, Tim Balint, and Jan M. Allbeck

Laboratory for Games and Intelligent Animation
George Mason University
4400 University Drive, MSN 4A5
Fairfax, VA 22030
{wlia,jbalint2,jallbeck}@gmu.edu

Abstract. While there continues to be exciting developments in research related to virtual characters, improvements are still needed to create plausibly human-like behaviors. In this paper, we present a synthetic parameterized memory model which includes sensory, working, and long-term memories and mechanisms for acquiring and retrieving memories. With the aid of this model, autonomous virtual humans are able to perform more reasonable interactions with objects and agents in their environment. The memory model also facilitates emergent behaviors, enhances behavioral animation, and assists in creating heterogeneous populations. To demonstrate the effectiveness of the memory model, we also provide an example in a 3D game environment and have conducted a user study in which we found general guidance in determining parameter values for the memory model, resulting in NPCs with more human-like game playing performances.

Keywords: Virtual Humans, Memory Model, Behavioral Animation.

1 Introduction

Non-Player Characters (NPCs) have become vital assets in games and simulations. They allow game authors to add depth to the world by providing valuable enemies, allies, and neutral characters. Over the past three decades, there has been a great deal of work on improving NPCs. However, while a majority of it has gone into creating more visually appealing animated characters, development of the underlying intelligence for these characters has remained fairly stagnant, creating strange and undesirable phenomenon such as repetitive behaviors and a lack of learning and knowledge understanding. A character may appear as a photo-realistic knight in shining armor, but can only greet the player or fight with the player monotonously. This lack of depth diminishes NPC believability and creates a less enjoyable gaming experience for the player.

There are many different forms and functions that NPCs need to fulfill, and these commonly correspond to their roles, relevance and importance to a player. In many of these cases, when the purpose of agents is to stay transiently and

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 1–14, 2013.

Fig. 1. Our agent (Carl) is trying to remeber objects in the environment

blend into the environment, it's less likely a player would spend a great deal of time examining each one of them. For example, if there is a squad of enemy soldiers the player must fight, it is doubtful that the player will spend considerable time monitoring each individual soldier's behaviors. In these contexts, there is not a great need for strong AI techniques, and it is generally impractical to implement and execute such techniques for a large group of agents. However, for those agents meant to serve as companions or enemies that are central to the game story and thus exist for longer periods of time, the player will most likely spend a lot of time interacting with them and also observing their behaviors. For these characters, techniques that create more believable agents are needed, and these types of NPCs are the focus of this work.

While there are many ways of improving the believability of virtual agents, in particular we see a human-like memory model helping to achieve this goal in two ways. First, given that information storage is inevitably needed for agents to reason and interact with their environments, a human-like memory model that includes false memory and memory distortion can result in more human-like performances. These human-level abilities (i.e. not sub-human nor super-human) will provide more plausible interactions with the player, including more reasonable competition and more engaging communication [26]. Secondly, an event-independent memory model can make longer simulation more plausible and allow agents to carry their knowledge to other scenarios without massive editing. To realize such a system, certain aspects of human memory are necessary such as learning, forgetting and false memories. While there are many techniques, such as scripting and behavior trees, that can be used to simulate an agent's memory model and can create the illusion of false memories and forgetful agents, these techniques usually require a great deal of crafting by a game play author in order to create some form of believability. A memory model that causes an agent to forget or create hazy memories provides the virtual character some variability in its understanding of the world and what has taken place. This variability is inherent within a reasonable human-like memory model, and these differences do not have to be enumerated and explicitly written by the game author, as they would with prior techniques.

Toward this end, we have developed a memory model that includes components for *Sensory Memory*, *Working Memory* and *Long-term Memory* and have designed it to contain features such as forgetting and false memories. We also provided multiple parameters related to various memory capabilities. These parameters can be set by level designers or game systems to vary the difficulty level of games. To determine a reasonable range of values for these parameters, we conducted a user study in a 3D game environment and carried out a performance analysis. To summarize, our contributions include:

- A memory model that supports a variety of psychological activities such as forgetting and false memories and aims to bring the memory model for NPCs one step closer to a human-like memory system.
- A memory model that offers flexibility to users by providing several tunable parameters. A user study has been conducted to further provide general guidance and possible ranges for the model parameters.
- A memory model grounded in a 3D game environment, where agents are capable of demonstrating complex, emergent, and social behaviors, making longer duration simulations and games more believable.

2 Related Work

Memory systems, appearing to be an inevitable component in intelligent agent architectures, have been studied and developed from a variety of points-of-view. In animated, autonomous agents research, several groups have combined vision and memory systems to allow virtual characters to perform navigation and path planning tasks [21,30]. Others have used memory to facilitate and enhance agent-object interactions [13,24], crowd simulations [23], believability and intelligence in synthetic characters [4,16], and virtual actors in dramas [18]. In these efforts, memory was not the main focus. It was a part of larger agent architectures that also included other components such as perceptual units, dialogue units, action selection modules, and goal and plan generation systems. Given the overall complexity, the memory model is often treated simply as permanent or temporary information storage with relatively simple structures, and is not intended to be scrutinizingly designed as a component to achieve human-like performances in games and other applications.

More specifically, a great deal of work has explored using certain types of memory found in psychology literature. Episodic and autobiographical memory [31,32] collect individual experiences that occurred at particular times and locations. Examples include, a pedagogical agent named Steve [25] who is equipped with episodic memory and can explain its decision-making process for a short time; Brom et al [3] exploited it to achieve longer duration storytelling activity in a gaming environment; Gomes et al [11] implemented an episodic memory model for emotion appraisal. Similar to episodic memory but with broader scope, autobiographical memory has also been incorporated into many applications. To name a few, Dias et al [7] has utilized it to enable synthetic agents reporting their past experiences. Similarly to the previously mentioned figure Steve, actors in [12] can tell stories for limited time period based on their autobiographical memory. While significant results were achieved, these efforts focused primarily on retrieving episodic knowledge and had memory modules crafted for specific tasks such as storytelling, communicating, and social companionship [17].

Another set of related work is in cognitive computing research. To list a few task-dependent applications, a memory model has been implemented for cognitive robots [8], and autonomous and virtual agents [6,10,5]. In addition, a large group of work has been associated with cognitive architectures such as SOAR [14], ACT-R [1], and CLARION [29]. These architectures generally have relatively comprehensive memory models in their large software infrastructures and are capable of simulating many psychological activities. Additionally, a shared goal of these architectures is to achieve task-independence including a standalone memory fragment. For example, Nuxoll [22] has attempted to build an event-independent episodic memory system integrated with SOAR. However, there are some issues with these cognitive architectures. First of all, generality is pursued, at the loss of some specific elements. Take, for example, SOAR, one of the virtual and gaming environment friendly cognitive architectures [15,14]. While its usage has been demonstrated in several gaming applications, most of them are taken place in simple environments where agents are capable of conducting a limited number of behaviors and interactions with environmental objects. Secondly, using these architectures requires considerable effort and seasoned programming skills. In contrast, a parameterized model with tunable values could be easily adopted by users to endow virtual characters with heterogeneous features.

To summarize, while strides have been made, improvements are still needed. In particular, animated, autonomous agents have limited memory model functionality, storytelling agents rely mainly on episodic or autobiographical memory, and general cognitive architectures are not yet mature enough for use in rich 3D environments. The work presented in this paper attempts to fill in gaps in these research efforts. Toward that end, we create a synthetic parameterized memory model based on several theories of memory and lessons learned in the cognitive computing and agents development research communities. The model is grounded in a game world with autonomous virtual humans conducting life-like interactions with objects and each other.

3 Memory System

In this section, we will first explain the memory representation and then detail components of our memory model which includes *Sensory Memory*, *Working Memory* and *Long-term Memory*, and most importantly why our design choices could benefit NPCs in enhancing their believability and achieving human-like performances.

3.1 Memory Representation

For representing memory, we have chosen a directed graph with nodes representing concepts and edges enacting links between concepts. We chose a directed edge following observation that humans generally archive memories in a certain order, and this order does not necessarily work in reverse sequence. This representation can benefit NPCs, such as in a scenario where a NPC has been given the formula of a medicine which can be made only by adding its elements in a certain sequence. While some elements of this formula may be forgotten, the agent should still be able to create and preserve a sense of sequence in order to facilitate resolving the missing elements. An example of this would be an agent thinking, "I remember that in order to make this medicine I need one more item after I use the purple potion, but I forget which item that is. I should search in this ancient book to determine what the item after the purple potion is". Without directions between concepts, this activity is harder to capture.

In addition, we have implemented a strength factor for both nodes and edges, hereby denoted $Node_{strength}$ and $Edge_{strength}$ accordingly. $Node_{strength}$ indicates how strongly a concept is encoded within the memory model while $Edge_{strength}$ denotes the degree of ease of going from one node to another. These strength factors add believability to characters. To be specific, while many objects afford many different actions, an agent should select not a random object, but the most familiar object for a particular task. For example, if a NPC is trying to build a birdhouse, it is more believable that the NPC uses its frequently used hammer over a nail gun. While both items can perform the same tasks, the NPC is more accustom to using his hammer, and so should be more likely to do so. Currently, both $Node_{strength}$ and $Edge_{strength}$ share the same integer range, 1 to 10, and this range is further divided into two stages, a strong and a weak stage. By creating this division, our model is able to support memory distortion. The division between the two stages is chosen by the user, who does so by choosing the threshold values $Node_{threshold}$ and $Edge_{threshold}$. Therefore, if someone has chosen $Node_{threshold}$ to be 5, then a value of 1 to 4 would create a weak stage node and values between 5 to 10 would be the node's strong stage. While the threshold may be different, the same logic applies to edge values. When nodes and edges are in weak stage, the phenomenon known as false memory can occurs with probability $P_{false} = 1 - \frac{Node/Edge_{strength}}{Node/Edge_{threshold}}$. So, if $Node_{threshold}$ has been set as 5, then nodes in the weak stage with values from 1 to 4 have corresponding probabilities from 80% to 20% to be forgotten. This simple design allows us to simulate partial human fuzziness. Currently instead of having a sophisticated mechanism for selecting an incorrect node and edge to replace the correct ones during the course of false memory generation, the false concept will be picked among neighboring concepts based on an object ontology of the environment.

3.2 Sensory Memory

One module of our memory model is called *Sensory Memory*. This component maintains transient information captured by the sensory system. In our current work we only address vision. There are several reasons for us to design and develop this component including psychology research studies that have shown that information coming from environment does not contact memory directly. Instead, these studies show that information will spend a short time in a system that serves as an interface between the perception and memory [27,2]. This interface is always present in real humans, and can be seen in simple phenomenon such as the light curve observed by swinging an illuminating object swiftly in the dark. We believe this module is also necessary due to the fact that, unlike electronic devices, humans never record complete events, as noted in [20]. Instead, physical humans record key elements and later on with the aid of environmental cues, use these elements to reconstruct past scenarios and events. In this work, we use $SM_{capacity}$ to indicate how many cues will be potentially maintained in the sensory memory module. Inspired by findings in [19], which states a human can process 5 to 9 items at once, and considering that design choice that all elements from sensory memory will transit to working memory with various strength factors, we have chosen the range of $SM_{capacity}$ being 1 to 10. This implementation helps preserve *Sensory Honesty* which it has been argued plays a vital role in synthetic characters [4]. Essentially, we believe a player would not want to compete with a NPC for a task in a place where a lot of objects exists and the NPC can process everything while the player is bound by the human limitation of only processing 5 to 9 items at a time.

3.3 Working Memory

We have also included a *Working Memory* module which is seen in many virtual agent architectures. This is a module that stores information currently being used by the agent for a variety of cognitive activities such as reasoning, thinking and comprehending. Studies have shown that in order to use declarative long-term memory elements, one has to extract the material into working memory first. The working memory has been documented to have much smaller size and information maintaining time compared to long-term memory [2]. From this, we decide that only one graph structure can exist in the working memory at a time. This feature has also been adopted by several other architectures, including SOAR. In addition, while one graph could contain multiple nodes and edges, the actual reinforcement rate on each node and edge depends on the total number of nodes/edges and the information linger time. The equation for calculating the reinforcement rate is: $Node/Edge_{rate} = \frac{Information\, linger\, time(secs)}{Total\, number\, of\, nodes/edges} \times \frac{WM_{scale}}{10}$ where WM_{scale} shares the same range (i.e. 1 to 10) with the strength factor. Therefore, if the working memory currently contains 4 nodes and 3 edges and this information resides for 12 seconds when WM_{scale} equals to 10, then each node and edge will get its strength factor increased by 3 and 4 accordingly. This procedure can be interpreted as: a few items lingering for a long time in working memory would have their strength factor values higher than those of many items lingering for a short time. This design decision is inspired by findings in [28] in which the author found that memory reactive time increases in a linear fashion when the number of items in the memory increases. Though here we don't have a reactive time associated with each node and edge, we believe the strength factor can act as an indicator manifesting memory retrieving difficulties. Through this approach, a user can control the amount of information being processed in the working memory, achieving a similar effect to setting a working memory duration and processing strategy without concerns about individual memory differences and the graphical complexity of specific scenarios.

3.4 Long-term Memory

The final module is the *Long-term Memory* module. Intuitively, this module maintains an extensive number of concepts and maintains them for a longer duration. Unlike working memory, this component can contain multiple rather than one graph structure. While both $Node_{strength}$ and $Edge_{strength}$ will only get strengthened in the working memory, in long-term memory they will suffer from decay. Currently, the decaying activity occurs every 5 minutes and the decaying percentage roughly follows the classic Ebbinghaus forgetting curve [9]. In addition, while there is a great debate within the scientific community on whether memory elements are completely forgotten, from the engineering perspective, we have decided to remove any node and edge with $Node/Edge_{strength}$ below 1 and consider any element with $Node/Edge_{strength}$ higher than 10 as a permanent encoded concept. Last but not least, in developing memory models for virtual agents, the distinction between two types of long-term memory: the episodic memory and semantic memory, is not rare. However, in this work, we will not differentiate them for two reasons: firstly, we are targeting more general tasks in gaming environments rather than specific activities as discussed in the related work section. Secondly, in psychology studies, there exist evidence showing these two types of memories could transform into each other in certain forms. However, the process and detailed relationship between them has yet been resolved [2].

The general working pattern of above modules is as follows: firstly via perception, certain environmental cues are passed into the sensory memory. If the number of cues in a place are greater than $SM_{capacity}$, the cues will be random selected. These cues are then sent to the working memory, where they are formed into a strong connected graph with a minimum $Node/Edge_{strength}$. If a concept has additional properties such as color and material, only the concept and its properties would be linked together. Next, cues will be matched against long-term memory and elements (i.e. nodes and edges) above $Node_{threshold}$ and $Edge_{threshold}$ will be retrieved correctly while elements below this threshold would be subject to potential concept replacement. For example, if the correct concept "Coin_0" has a strength factor below $Node_{threshold}$, then it's possible it will be replaced by "Coin_1" under the meta-concept "Coin" (if "Coin_1" exists, otherwise it might be forgotten). In other words, a different coin might be incorrectly remembered. Currently, we only consider replacing concepts that are at the same level of the correct concept in our object ontology. This design decision is supported by the data we collected through a user study in which we found that most false memories involved players picking an object from under the same meta-concept instead of completely unrelated concepts. After successfully constructing a single graph structure in working memory, reinforcement will start according to the total number of nodes/edges, information linger time, and the value of WM_{scale}. Finally, memory material will transition to long-term memory with their various strength factors. The whole process is illustrated in Figure 2.

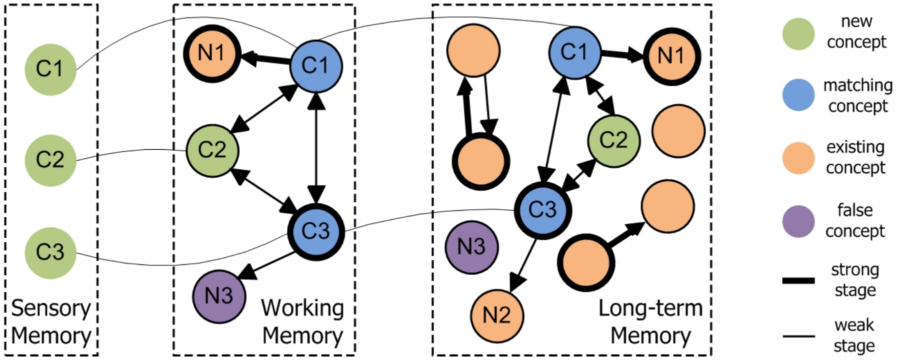


Fig. 2. The general working pattern of sensory memory, working memory and long-term memory. "C" stands for Cue and "N" stands for Node (not all nodes are labeled).

4 Example and Analysis

To further explore our memory model and its capabilities, we implemented a game (Seen in Figure 3). In this game, our heroes are tasked with saving their princess, who is locked away in a tower. The first person to find the magic gems to unlock the door and free her will win her everlasting love. Players (both human and a NPC we named Carl) begin by exploring the environment to find two magic gems (one red, one blue). As they explore different areas, they find many objects, but no magic gems. After a time, they encounter a villager NPC holding a red gem. Interacting with the villager, they discover he would be willing to trade the red gem for an iris. Players then have to try to remember where they might have seen an iris. If they cannot remember, they start looking around until they find it. While looking around, their memories of the environment are reinforced. A similar procedure is followed for the blue gem.

While our NPC Carl can certainly successfully complete his task, we are more interested in how his performance can approximate real human performance and what memory model parameters would be appropriate for different human player skill levels. In order to explore this, we conducted a user study. In total 31 subjects (15 female, 16 male) participated. Before playing, subjects took a simple memory test and completed a survey related to their experience with video games. Then each subject was asked to play the game solo eight times with different game level complexities. In the first four rounds, the game

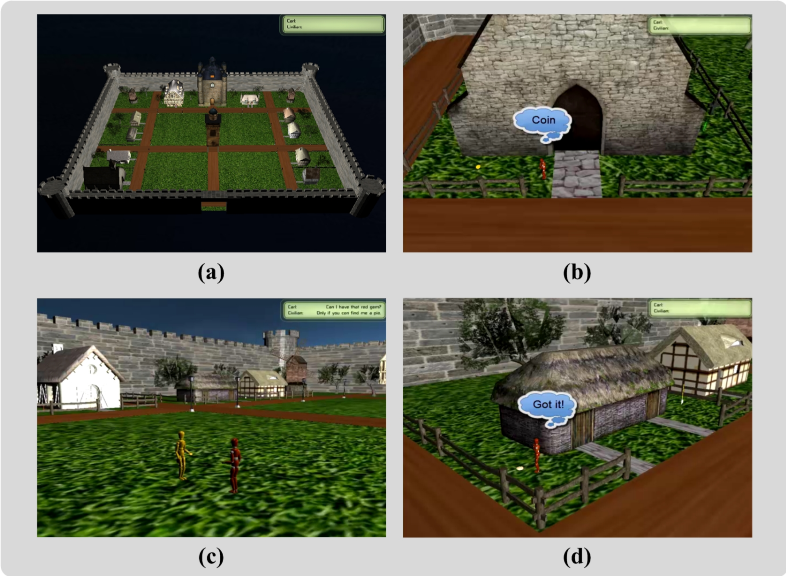


Fig. 3. (a) The game environment. (b) Carl is exploring the environment, trying to find the desired gems. (c) Carl talks to a civilian about trading his gem. (d) Carl uses his memory successfully to find the item for trading.

world contained only eight objects for the players to remember. The number of objects in a given area increased from a single object to four similar objects of different colors. In later rounds, more object models were included as opposed to differentiating by color. Results are shown in Figure 4 in which subjects are classified as having good, medium, or bad memories. We found game worlds containing more objects created more confusion, resulting in players forgetting or incorrectly remember object locations more often.

In particular, we have set Carl's $SM_{capacity} = 7$ and $Node/Edge_{threshold} = 5$. With gathered data, we were able to tune the parameters of our memory model to make our NPC achieve more human-like performances. By scaling WM_{scale}, which determines the reinforcement rate on nodes and edges in the working memory, we found when $WM_{scale} \geq 8$ Carl achieves similar performances to human player's with good memory; when $5 \leq WM_{scale} < 8$ medium memory performance is obtained and when $WM_{scale} < 5$ performance of human player with bad memory is reached. Furthermore, the data yielded some other interesting findings:

– Using all different object models has no significant improvement over using limited models with different colors (which was assumed to be more confusing) in terms of recalling their locations. This implies that creating more models in games may have limited function in helping players remember their locations.

– When the total number of objects in the environment was over 20, 90% of players forgot the desired item's location even when the item was among the last 6 items seen. This indicates that reinforcement is not strictly distributed to the latest concepts.
– People who play games more than 20 hours per week out-performed people who play games between 5 to 20 hours per week and those who play games less than 5 hours a week by 26% and 65% respectively, with no such increment in their memory capabilities (based on their self-reporting memory abilities and memory test results).

Besides the user study, in order to further provide users some guidance in tuning the parameters of our memory model and also to evaluate the sensitivity of these parameters, we conducted several analyses. In our first analysis, we examine the impact changing the values of $Node_{threshold}$ and $Edge_{threshold}$; how

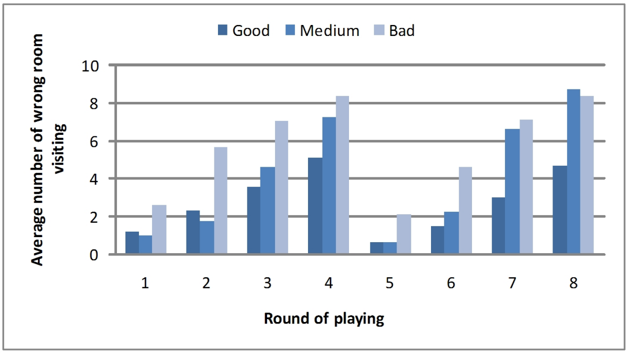
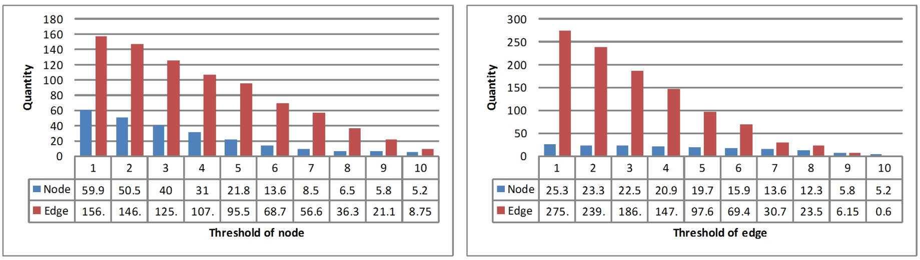


Fig. 4. Performance results of the user study: subjects are grouped as having good, medium, or bad memories according to the memory test. The first four (i.e. 1 to 4) rounds contain limited objects with different colors while the last four (i.e. 5 to 8) contain more object models in which influence of the color factor was dropped.



(a) **(b)**

Fig. 5. An experiment of influences of $Node_{threshold}$ and $Edge_{threshold}$ on the total number of strong nodes and strong edges retrieved from the long-term memory into the working memory. (Quantity of nodes is 100, of edges is 2,500; $SM_{capacity} = 5$, $Edge_{threshold}$ in (a) and $Node_{threshold}$ in (b) are both 5).

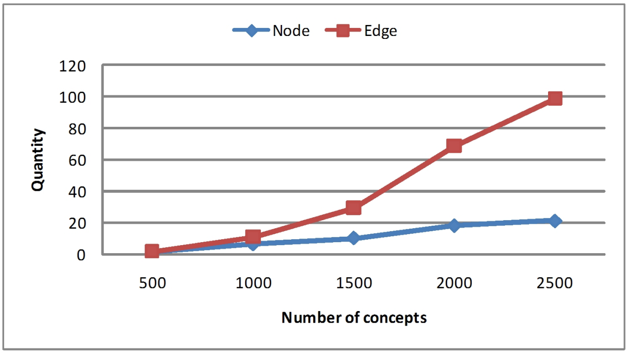


Fig. 6. An experiment of correlation between number of concepts and total number of nodes and edges in working memory ($SM_{capacity} = 5$, $Node/Edge_{threshold} = 5$)

many nodes and edges in their strong stages will get retrieved from the long-term memory into the working memory. These values play a vital role in determining agent memory capability. To carry out the experiment, we have chosen 100 nodes and random 2,500 edges with randomly assigned strength factors residing in the long-term memory. The first result is shown in Figure 5(a). In this case, the $Edge_{threshold}$ and $SM_{capacity}$ have been set to 5 while $Node_{threshold}$ increases from 1 to 10. As we can see, as $Node_{threshold}$ grows which indicates the range of node strong stage shrinks, the number of retrieved strong nodes and strong edges decreases. The result shown in Figure 5(b) has enlarged the difference between the two values. In this setting, both the $Node_{threshold}$ and $SM_{capacity}$ have been set to 5, while the $Edge_{threshold}$ increases from 1 to 10. This analysis indicates, given that in our memory model, concepts are represented by nodes, in terms of enhancing memory capability of the virtual characters, $Node_{threshold}$ has more influence over $Edge_{threshold}$ even though the later value decides how many possible links can be stretched out from a particular node during the memory retrieving process.

Based on the results of our first analysis, a second one was run testing the total number of retrieved strong nodes and edges in working memory when the edges were increased from 500 to 2500. The result is shown in Figure 6. In this case, the $Node_{threshold}$, $Edge_{threshold}$ and $SM_{capacity}$ have been all set to 5 for a consistent experiment. When 500 edges exist in the graph and also with random strength factors that could spread with value equal or higher than 5, we can see basically only starting nodes and nearly no edges would be activated. After that, the difference between the two values enlarges.

The final analysis focuses on $SM_{capacity}$. While the number of starting nodes controlled by $SM_{capacity}$ is assumed to have an impact on the number of strong nodes and strong edges retrieved into working memory, we found that increasing the value of $SM_{capacity}$ from 5 to 9 has a trivial effect. This is because while $SM_{capacity}$ is increasing from 5 to 9, making the starting nodes in the activation process increase, this ratio compared to overall nodes (i.e. 100 in this experiment) is still quite small. In addition, the sparse nature of the graph limits the effects.

5 Conclusion and Future Work

We designed and developed a synthetic, parameterizable memory model with the intent of creating more plausibly human NPCs. Our model embodies a sensory system, working memory model, and long-term memory model, which supports a variety of psychological activities such as forgetting concepts and creating false memories. This allows our virtual characters to perform complex, emergent, and believable behaviors. Additionally, our user study and analysis provides guidance and insights into potential uses for our memory model and it's effectiveness in setting up NPCs with skill levels comparable to human users.

Of course, our model is not complete. We would like to add a more refined perception model which includes modalities other than sight. Also, our future agenda includes extending the implementation of memory distortions to more than just a simple probability. This mechanism can potentially stimulate agent creativity and enable more spontaneous, emergent behaviors. Given our current detailed memory infrastructure, further development of an effective, plausible mental control module is also worth exploration. Finally, with some optimization, integration with an existing knowledge base could yield interesting results.

Acknowledgments. Partial support for this effort is gratefully acknowledged from the U.S. Army SUBTLE MURI W911NF-07-1-0216. We also appreciate donations from Autodesk.

References

1. Anderson, J.R.: How Can the Human Mind Exist in the Physical Universe? Oxford University Press (2001)
2. Baddeley, A., Eysenck, M.W., Anderson, M.C.: Memory. Psychology Press, Taylor & Francis Inc. (2012)
3. Brom, C., Pešková, K., Lukavský, J.: What does your actor remember? Towards characters with a full episodic memory. In: Cavazza, M., Donikian, S. (eds.) ICVS 2007. LNCS, vol. 4871, pp. 89–101. Springer, Heidelberg (2007)
4. Burke, R., Isla, D., Downie, M., Ivanov, Y., Blumberg, B.: Creature smarts: The art and architecture of a virtual brain. In: Proceedings of the Computer Game Developers Conference, pp. 147–166 (2001)
5. Cha, M., Cho, K., Um, K.: Design of memory architecture for autonomous virtual characters using visual attention and quad-graph. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, pp. 691–696 (2009)
6. Deutsch, T., Gruber, A., Lang, R., Velik, R.: Episodic memory for autonomous agents. In: Proceedings of the Fifth International Conference on Human Systems Interaction, pp. 621–626 (2008)

7. Dias, J., Ho, W.C., Vogt, T., Beeckman, N., Paiva, A., André, E.: I know what I did last summer: Autobiographic memory in synthetic characters. In: Paiva, A., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 606–617. Springer, Heidelberg (2007)
8. Dodd, W.: The design of procedural, semantic and episodic memory systems for a cognitive robot. Master thesis. Vanderbilt University. Nashville, TN, USA (2005)
9. Ebbinghaus, H.: Memory: A Contribution to Experimental Psychology (1964)
10. Franklin, S., Patterson Jr., F.G.: The lida architecture: adding new modes of learning to an intelligent, autonomous, software agents. In: LDPT 2006 Prceedings (Intergrated Design and Process Technology): Society for Design and Process Science (2006)
11. Gomes, P.F., Martinho, C., Paiva, A.: I've been here before! location and appraisal in memory retrieval. In: Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011, pp. 1039–1046 (2011)
12. Ho, W.C., Dias, J., Figueiredo, R., Paiva, A.: Agents that remember can tell stories: integrating autobiographic memory into emotional agents. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2007, pp. 10:1–10:3 (2007)
13. Kuffner Jr., J.J., Latombe, J.C.: Fast synthetic vision, memory, and learning models for virtual humans. In: Proceedings of Computer Animation 1999, pp. 118–127 (1999)
14. Laird, J.: The Soar Cognitive Architecture. The MIT Press (2012)
15. Laird, J.: Using a computer game to develop advanced ai. Computer, 70–75 (2001)
16. Li, W., Allbeck, J.M.: The virtual apprentice. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 15–27. Springer, Heidelberg (2012)
17. Lim, M.Y.: Memory models for intelligent social companions. In: Zacarias, M., de Oliveira, J.V. (eds.) Human-Computer Interaction. SCI, vol. 396, pp. 241–262. Springer, Heidelberg (2012)
18. Mateas, M.: Interactive drama, art and artificial intelligence. Ph.D. thesis. Carnegie Mellon University. Pittsburgh, PA, USA (2002)
19. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review 63(2), 81–97 (1956)
20. Minsky, M.: The Society of Mind. Simon and Schuster (1986)
21. Noser, H., Renault, O., Thalmann, D., Thalmann, N.M.: Navigation for digital actors based on synthetic vision, memory, and learning. Computers & Graphics 19(1), 7–19 (1995)
22. Nuxoll, A.M.: Enhancing intelligent agents with episodic memory. Ph.D thesis. University of Michigan, Ann Arbor, MI, USA (2007)
23. Pelechano, N., O'Brien, K., Silverman, B., Badler, N.I.: Crowd simulation incorporating agent psychological models, roles and communication. In: First International Workshop on Crowd Simulation, pp. 21–30 (2005)
24. Peters, C., O'sullivan, C.: A memory model for autonomous virtual humans. In: Proceedings of Eurographics Irish Chapter Workshop, EGIreland 2002 (2002)
25. Rickel, J., Johnson, W.L.: Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. Applied Artificial Intelligence 13(4-5), 343–382 (1999)
26. Ruttkay, Z., Reidsma, D., Nijholt, A.: Human computing, virtual humans and artificial imperfection. In: Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI), pp. 179–184 (2006)

27. Schacter, D.L.: The Seven Sins of Memory. Houghton Mifflin Company, New York (2001)
28. Sternberg, S.: High-speed scanning in human memory. Science 153, 652–654 (1966)
29. Sun, R.: The CLARION cognitive architecture: extending cognitive modeling to social simulation. In: Sun, R. (ed.) Cognition and Multi-Agent Interaction. Cambridge University Press (2006)
30. Thomas, R., Donikian, S.: A spatial cognitive map and a human-like memory model dedicated to pedestrian navigation in virtual urban environments. In: Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R. (eds.) Spatial Cognition V. LNCS (LNAI), vol. 4387, pp. 421–438. Springer, Heidelberg (2007)
31. Tulving, E.: Elements of episodic memory, vol. 2. Clarendon Press (1983)
32. Williams, H.L., Conway, M.A., Cohen, G.: Autobiographical memory. In: Cohen, G., Conway, M. (eds.) Memory in the Real World, 3rd edn., pp. 21–90 (2008)

DyBaNeM: Bayesian Episodic Memory Framework for Intelligent Virtual Agents

Rudolf Kadlec and Cyril Brom

Faculty of Mathematics and Physics
Charles University in Prague, Czech Republic
rudolf.kadlec@gmail.com
http://amis.mff.cuni.cz

Abstract. Episodic Memory (EM) abilities are important for many types of intelligent virtual agents (IVAs). However, the few IVA–EM systems implemented to date utilize indexed logs of events as the underlying memory representation, which makes it hard to model some crucial facets of human memory, including hierarchical organization of episodes, reconstructive memory retrieval, and encoding of episodes with respect to previously learnt schemata. Here, we present a new general framework for EM modeling, DyBaNeM, which capitalizes on bayesian representation and, consequently, enables modeling these (and other) features easily. By means of a proof-of-concept implementation, we demonstrate that our approach to EM modeling is promising, at least for domains of moderate complexity.

Keywords: Episodic Memory, Dynamic Bayes Networks, Dialog.

1 Introduction

Episodic memory (EM) [27] represents personal history of an entity. Episodic memories are related to particular places and moments, and are connected to subjective feelings and current goals. In the context of agent-based systems, episodic memory has been studied as a tool enhancing an agent's performance in simulated environments [23,26,13]. EM abilities can also increase *believability* of intelligent virtual agents (IVAs) in many applications, including role-playing games, serious games, interactive storytelling systems, and tutoring applications. Agents for a serious anti-bullying game were equipped by a simple EM for the purpose of debriefing [9]. Virtual guide Max uses EM to modify museum tours based on Max's previous experience [19,24]. Generic EM for virtual characters was proposed in [6]. Similarly EM can be used in cognitive robots [10]. At the same time, studies investigating how humans perceive IVAs with EM abilities started to be conducted. For instance, several results suggest that humans tend to prefer IVAs with imperfect memory [5,22]. Increased interest of users interacting with an agent presenting background stories possibly stored in EM in first person was shown in [3].[1]

[1] Some authors prefer the term autobiographic memory. In cognitive psychology, the meaning of the two terms differs but for the purpose of the present paper, we use them as synonyms. For more detailed review of EM agents, see [21].

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 15–28, 2013.

Here we present a new EM modelling framework, DyBaNeM. It has been created specifically with IVAs' needs in mind, that is, with the needs to model human EM in a more believable manner than in the past, for instance, to respect fallibility of human memory. The framework brings several key innovations. First, to our knowledge, none of the abovementioned systems enables *reconstruction* of hierarchical episode structure (i.e., episode — subepisode relationship) in cases where an observer IVA, let us call him Bob, equipped with the EM observes another IVA, say, Alice. Bob can see Alice's atomic actions but he has to reconstruct her high level goals if he wants to remember them. Only the model presented in [6] makes it possible for Bob to remember his own hierarchy of episodes but not that of Alice's. Second, our framework enables probabilistic reconstructive retrieval process that can result in reconstruction of events that have not happened at all, but they are sensible given the other stored memories. Third, the model remembers only some of the most salient events as opposed to most of the current models that use data structure resembling plain log of events such as [6,9]. While some models use emotions as a measure of saliency, e.g. [8,24], we use mathematically better rooted deviation from a statistical schema. Fourth, current models cannot express degree of belief in the recalled memory, they usually either return an episode or nothing. Our framework removes this restriction, recall in DyBaNeM results in multiple, more or less probable, possibilities. Fifth, the framework uses IVA's personal experience in encoding of the episodes. Two IVAs may remember the same episode in differently.

From the psychological point of view the framework is inspired by the Fuzzy-Trace Theory (FTT) [11]. FTT hypothesizes two parallel mechanisms that encode incoming information: *verbatim* and *gist*. While verbatim encodes the surface-form of the information in detail, gist encodes the meaning in a coarse-grained way [11], capitalizing on previously learnt *schemata* [2] of episodes and parts of episodes. From the computational point of view the framework uses Dynamic Bayesian Network (DBN) [18] as the underlying probabilistic model.

We illustrate possible use-cases of the framework on the following example. Imagine a MMORPG inhabited by hundreds or even thousands of non player characters (NPCs). Each NPC can be interviewed by a human player that may ask two basic types of questions: 1) "What were you doing yesterday?"; 2) "What was the player X doing yesterday?" The first question asks about the NPC's recollection of its own actions, the second asks about the NPC's memory for actions of a human controlled avatar (or a different NPC). It is clear that the NPC has to be equipped with an EM model to answer both of these questions. However, the second question also requires the model to be able to *interpret* the players' (NPCs') actions and infer his/her high level goal that are not directly observable in the environment. Our framework can do that. In addition, the model should be generic and applicable to different NPCs with minimal effort. In DyBaNeM, the model's parameters can be automatically adjusted for each type of NPC in the game by means of learning schemata of episodes by standard byesian methods.

Finally, while our IVAs are not equipped with a dialog generating system, our framework enables, in principle, the following features in the dialog between the player and a NPC :

1. The NPC can provide a high level summarization of an activity. For instance, when the player (P) asks: "What were you doing yesterday?", the NPC (N) equipped with our model can answer: "After getting up I went to work, in the afternoon, I visited my friends and then I returned home." instead of inadequately detailed "I got up, then I did my morning hygiene. I had a breakfast, I get dressed and ..."
2. The player can ask further clarifying questions. E.g., P: "How did you get to your friends?"; N: "I walked there."
3. The NPC can express degree of certainty for each recalled event. E.g., N: "Maybe I went there by a car. I'm not sure."
4. The NPC can make mistakes that are believable given the context. E.g., N: "I went to work by public transport." (Even though the NPC used a car.)
5. The memory weights interestingness of the events, thus it can highlight the most unusual memories. P: "What were you doing yesterday?"; N: "I saw a foreign army marching through the town, the rest of the day was as usual."
6. Personal experience can influence interpretation of others' activity. A worker NPC may think that the observed player is just visiting a museum, whereas a thief NPC may reveal that the player is preparing a theft.

We now detail these six above mentioned use cases and we sketch how DyBaNeM can be used to implement them. Then we describe DyBaNeM's core. In the end we present a prototype IVA simulated in a 3D environment equipped with DyBaNeM and experiments demonstrating applicability of the EM model.

2 How DyBaNeM Supports Rich Dialog with IVAs

First we will briefly summarize functions and processes of DyBaNeM EM, then we will detail how these can support the user's interaction with the IVA.

DyBaNeM uses *episodic schemata* learnt a priori to segment sequences of observations into meaningful *episodes* that can have hierarchical structure. Episodic schemata are parameters of the underlying probabilistic model used in several stages of DyBaNeM's working cycle. The probabilistic model is implemented by a DBN. First the schemata has to be specified by hand or learnt from labeled data. This happens "offline" before the model is deployed to the IVA. Later, in *encoding*, the model is presented with a sequence of observations to be stored. DyBaNeM deduces hierarchy of episodes represented by the observations and picks the most interesting facts called *mems* that will be stored (persisted in a long term store). These mems will become the internal representation of the observations, e.g. one day of the agent's activity. Interestingness is measured with the use of the episodic schemata. During *storage* some of the mems may be forgotten. In *retrieval* the mems, that is the exact memory of fragments of the past events, together with episodic schemata, are used to reconstruct the whole original episodes. This process may be imperfect, the more mems remain stored the better will be the match between the original and the recalled episodes. We may perceive this process as a lossy compression of the episodes.

Now we will show how the functions listed in Introduction are enabled by DyBaNeM. We will use an example dialog where a human user interviews IVA Bob about his observation of IVA Alice. Technical details of these six functions will be discussed later.

1. High level summarization is enabled by hierarchical nature of DyBaNeM. The recalled sequences contain not only atomic actions but also high level episodes that can be used as summarization of the sequence of actions. Thus if Bob has DyBaNeM with two levels of abstraction, the values reconstructed by the bayesian inference on the highest level can be used to provide a summary of the day. Fig. 1 shows an example of such situation.

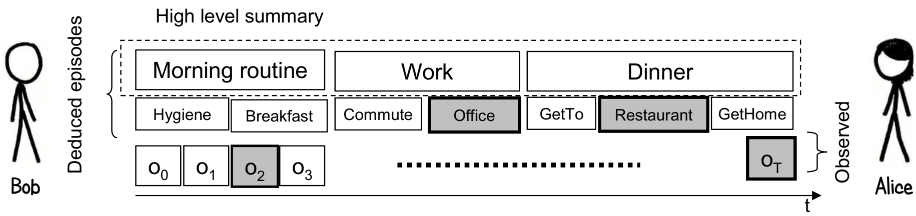


Fig. 1. Summarization example. Bob is a detective who monitors Alice. First, let Bob observe Alice's atomic actions $o_{0:T}$. Second, Bob can deduce the higher level episodes from this observation. Third, Bob encodes the whole day by the most salient events — these are the mems computed in the encoding algorithm. Mems are marked as the shaded boxes. When Bob is asked to summarize what Alice did yesterday he recalls the mems and reconstructs the rest with the use of the episodic schemata. In the end he responds by episodes in the highest level: "Morning routine, work and dinner."

2. Possibility of further clarifying questions is another useful feature of the hierarchical memory organization. When the user asks for details of an episode, Bob can reply by its sub-episodes as illustrated on Fig. 2a.

3. Expressing degree of certainty for recalled events is enabled by probabilistic nature of the framework. Each action/episode is represented by at least one random variable in the DBN. During reconstructive recall we obtain a probability mass function (PMF) for each variable that encodes probability of every action/episode at this point in time. When the probability of the most probable outcome dominates the other outcomes, we can say that the IVA is sure. However if there are two competing alternatives, the IVA can reflect this in the dialog. See Fig. 2b for an example.

4. Believable mistakes in recall can emerge as interplay of forgetting and reconstructive retrieval. When only a few mems remain stored then during the recall the forgotten events are reconstructed from the episodic schema. It can happen that the schema predicts an event that had not actually happen but it fits

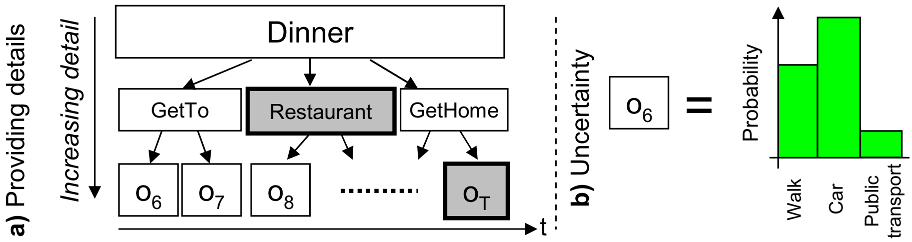


Fig. 2. **a)** When Bob is asked to say more about Alice's dinner, he will reply: "She left from work and went to the restaurant, she ate there and then she went back home." Shaded boxes represent mems, white represent reconstructed events. **b)** Further question can be: "How did she get to the restaurant?" which asks about recall of atomic actions represented by observations o_6 and o_7. In case of o_6 the associated PMF computed in recall assigns similar probability to both $Walk$ and Car. Thus Bob is not sure and he can reflect this in his answer: "She went by car or she just walked, I am not sure, sorry."

well to the way the episode usually unfolds. Different approach to this so called false memories phenomenon is discussed in our previous work [28]. Continuing in the example from Fig. 2, it may be the case that Alice used *Public transport* that day, but Bob does not remember this in a mem and his schema favors other options.

5. Measuring interestingness of events can be achieved by comparing the actual events to prediction from the schema. Imagine that 95 percents of days start by a sequence: *Get up, Brush teeth, Have a shower, Have a breakfast.* If the schema is expressive enough to capture this sequence, those events will become completely uninteresting. They are predictable, thus they do not distinguish one day from other. However meeting foreign soldiers marching through one's home town is much less probable. Thus it is the event that deserves more attention in the dialog than brushing teeth every morning again and again. The general notion is the lower the probability of an observed event given schemata the higher the surprise of observing it. We use Kullback-Leibler (KL) divergence [20] to measure how each observed event "diverges" from the prior prediction given solely by the schemata.

6. Influence of personal experience on interpretation of behavior of others is possible through a personalized set of episodic schemata for every IVA. Episodic schemata are parameters of the probabilistic model used in DyBaNeM, thus if Bob has a schema *theft preparation*, he may reveal that Alice was not visiting the gallery because of her interest in the new exhibition. Instead, he may conclude, she was examining the safety devices near the Da Vinci's painting. If the player asks IVA Cloe who does not have such schema, she would not know what Alice was planning.

3 DyBaNeM: Probabilistic EM Framework

We now describe DyBaNeM's computational core. We start with auxiliary definitions needed for description of the framework. Then we show how DBNs can be used for activity/episode recognition and how the episodic schemata are represented. We present the algorithms of encoding, storage and retrieval. Finally, we show how features 1-6 from Sec. 2 can be implemented in DyBaNeM. Additional details of DyBaNeM that are out of scope of this paper are available in [15,16]. DyBaNeM is available for download on its homepage[2].

Notation. Uppercase letters denote discrete random variables (e.g. X, Y) whereas lowercase letters denote their values (e.g. x, y). PMF of random variable X is denoted by $P(X)$. Domain of X is denoted as $D(X)$. Notation $X_{i:j}$ is a shorthand for sequence of variables $X_i, X_{i+1} \dots X_j$; analogically, $x_{i:j}$ is a sequence of values of those variables, the subscript denotes time. $\mathcal{M}$ will be a probabilistic model and $\mathcal{V}$ is a set of all random variables in the model.

Now we formalize representation of episodes and world state assumed by DyBaNeM.

Episode is a sequence (possibly of length 1) of observations or more fine-grained episodes (sub-episodes) that has a clear beginning and an end. Note that episodes may be hierarchically organized.

Episodic schema is a general pattern specifying how instances of episodes of the same class look like. For instance, an episodic schema (cf. the notion of script or memory organization packet [25]) might require every episode derivable from this schema to start by event a, then go either to event b or c and end by d.

Episodic trace $\epsilon_t^{0:n}$ is a tuple $\langle e_t^0, e_t^1 \dots e_t^n \rangle$ representing a hierarchy of episodes at time t; e_t^0 is the currently active lowest level episode, e_t^1 is its direct parent episode and e_t^n is the root episode in the hierarchy of depth n. Example of an episodic trace can be $\epsilon_0^{0:n} = \langle WALK, COMMUTE \rangle$ and $\epsilon_1^{0:n} = \langle GO\_BY\_BUS, COMMUTE \rangle$. The notation of episodic trace reflects hierarchical nature of agent behavior.

Our framework uses probabilistic representation, hence even if there is only one objectively valid episodic trace at each time step, input of the EM model will be a probability distribution. Let E_t^i denotes a random variable representing a belief about an episode on level i at time t. While the true value of E_t^i is, say, e_t^i, the PMF enables us to cope with possible uncertainty in perception and recall.

Probabilistic episodic trace $E_t^{0:n}$ is a tuple of random variables $\langle E_t^0 \dots E_t^n \rangle$ representing an agent's belief about what happened at time t. Analogically $E_{0:t}^{0:n}$ denotes probabilistic episodic trace over multiple time steps. The following data structure represents an agent's true perception of the environment state. Let ρ_t denotes **observable environmental properties** at time t.

For instance, ρ can hold atomic actions executed by an observed agent (and possibly other things too), e.g. $\rho_0 = STAND\_STILL$, $\rho_1 = GET\_TO\_BUS$. Analogically to $E_t^{0:n}$ and $\epsilon_t^{0:n}$, O_t is a random variable representing belief about observation ρ_t.

[2] DyBaNeM's homepage: `https://code.google.com/p/dybanem/`

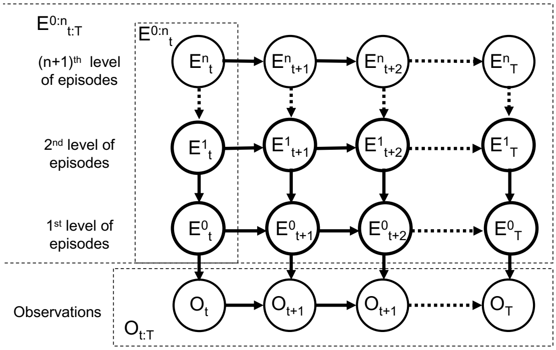


Fig. 3. An example of a DBN's structure called CHMM [4] together with our notation

Fig. 3 shows how these definitions translate to an example DBN structure used in this paper called Cascading Hidden Markov Model (CHMM) [4].

Surprise. In encoding, the framework works with quantity measuring difference between the expected real state of a random variable and its expected state given the remembered facts. We call this quantity surprise. In Bayesian framework surprise can be defined as "difference" between prior and posterior probability distributions. We adopt approach of [14] who propose to use KL divergence [20] to measure surprise. **KL divergence** of two PMFs $P(X)$ and $P(Y)$, where $D(X) = D(Y)$ is defined as:

$$KL(P(X) \to P(Y)) = \sum_{x \in D(X)} P(X = x) ln \frac{P(X=x)}{P(Y=x)}.$$

We use notation with $\to$ to stress directionality of KL divergence; note that it is not symmetrical.

Learning Schemata. Episodic schemata are represented by parameters $\hat{\theta}$ of a DBN. Expressiveness of schemata depends on the structure of a model at hand. We suppose that the DBN's topology is fixed. Thus learning schemata will reduce to well known parameter learning methods. Topologies without unobserved nodes including CHMM, are learnt by counting the sufficient statistics [18]. In our case examples of episodes that we want to use for schemata learning will be denoted by $\mathcal{D} = \{d_1, d_2 \ldots d_n\}$ where each d_i is one day of an agent's life; d_i itself is a sequence of examples c_t, that is, $d_i = \{c^i_0, c^i_1 \ldots c^i_{T_i}\}$. Each c^i_t is a tuple $\langle \epsilon^{0:n}_t, \rho_t \rangle$, it contains an episodic trace and observable state of the environment.

DBN Architecture. For computing probabilities, our framework makes it possible to use various DBN architectures. In this paper we use a CHMM [4] architecture which is a hierarchical extensions of a well known Midden Markov Model (HMM) (see Fig. 3). However more complex models, better suited for activity representation, like Abstract Hidden Markov Memory Model (AHMEM) [7], can be used. Downside of the AHMEM is its higher computational cost, thus we use simpler, but still sufficient CHMM. Experiments comparing AHMEM with CHMM in Dy-

BaNeM are presented in [15]. The schemata are represented by parameter $\hat{\theta}$, that is, by all conditional probability mass functions (CPMFs) of the DBN's nodes. Expressiveness of the schemata depends on the structure of DBN. In CHMM episodic schemata encode probability of an episode given previous episode on the same level in the hierarchy and also given its parent episode ($P(E_t^i | E_{t-1}^i, E_t^{i+1})$).

Encoding. The encoding algorithm computes a list of *mems* on the basis of the agent's perception, $Per_{0:T}$, of the situation to be remembered. $Per_{0:T}$ is a tuple of PMFs such that $Per_{0:T} = \{f_X : X \in \mathit{Observable}\}$, where f_X is PMF for each variable X of interest. In a case when Bob is going to encode Alice's activity (see Fig. 1), $\mathit{Observable} = O_{0:T}$. Alice's ϵ^{Alice} is hidden to Bob, nevertheless Bob perceives Alice's atomic actions that are contained in ρ^{Alice}.

Algorithm 1 is a skeleton of the encoding procedure. The input of the algorithm is $Per_{0:T}$, where the time window $0 : T$ is arbitrary. In our work we use time window of one day. The output is a list of mems encoding this interval.

Algorithm 1 General schema of encoding algorithm

Require: $Per_{0:T}$ — PMFs representing the agent's perception of the situation (i.e. smoothed observations)
Require: $\mathcal{M}$ — probabilistic model representing learned schemata

procedure ENCODING($Per_{0:T}, \mathcal{M}$)
  $mems \leftarrow empty$ ▷ List of mems is empty
  while $EncodingIsNotGoodEnough$ **do**
    $X \leftarrow GetMem(\mathcal{M}, Per_{0:T}, mems)$
    $x_{max} \leftarrow MLO_{P_{\mathcal{M}}}(X|mems)$
    $mems.add(X = x_{max})$
  end while
  return *mems*
end procedure

The algorithm terminates once the $EncodingIsNotGoodEnough$ function is false. We use stopping criterion $|mems| < K$ because this models limited memory for each day. In each cycle, the $GetMem$ function returns the variable X_t^i that will be remembered. The MLO function (most likely outcome) is defined as: $MLO_{P_{\mathcal{M}}}(X|evidence) \equiv \arg\max_{x \in D(X)} P(X = x|evidence)$. We get the most probable value for X and add this assignment to the list of mems. The $GetMem$ function is implemented in the following way. The idea is to look for a variable whose observed PMF and PMF in the constructed memory differs the most. This variable has the highest surprise and hence it should be useful to remember it. This memory creation strategy is retrospective, it assumes that the agent has all observations in a short term memory, and, e.g., at the end of the day, he retrospectively encodes the whole experience. The strategy memorizes the value of variable X such that: $X \leftarrow \arg\max_{Y \in VOI} KL\left(P_{\mathcal{M}}(Y|Per_{0:T}) \to P_{\mathcal{M}}(Y|mems)\right)$, where $P(Y|Per_{0:T}) \equiv P(Y|X = f_X : f_X \in Per_{0:T})$; we condition the probability on all observations. $VOI \subseteq \mathcal{V}$ is a set of random *variables of interest* whose value can be remembered by the model. In our implementation $VOI = E^{0:n}_{0:T} \cup O_{0:T}$. Note that we remember only the most probable value, including the time index.

Storage and Forgetting. During storage, the mems can undergo optional time decayed forgetting. The following equation shows relation between age t of the mem m, its initial strength S and its retention R: $R(m) = e^{-\frac{t}{S}}$ [1]. The initial strength S of mem m can be derived from the value of KL divergence computed in $GetMem$. Once $R(m)$ decreases under the threshold β_{forget}, m will be deleted from the list of mems and will not contribute to recall any more.

Retrieval. Retrieval is a simple process of combining the schemata with mems. We obtain the list of mems for search cue k, which can be, e.g., a given day. Then we use assignments in the mems list as an evidence for the probabilistic model. The resulting PMFs for all variables of interest are returned as a reconstructed memory for the cue k. Retrieval can be formalized as computing $P_{\mathcal{M}}(Y|mems)$ for each $Y \in VOI$.

Now we show how DyBaNeM's dialog supporting features are implemented.

1. High level summarization and 2. Further clarifying questions are possible because of the hierarchical structure of DBN used in both encoding and retrieval. Values of variables $E^n_{0:T}$ (see Fig. 3) can be used for summarization. If the user asks for details of time interval $\langle t_1, t_2 \rangle$, values of $E^{n-1}_{t_1:t_2}$ can be used to construct the answer (or $O_{t_1:t_2}$ when $n = 0$).

3. Expressing degree of certainty of recall is implemented by computing entropy of random variable corresponding to the action/episode. Entropy $H(X)$ of random variable X is defined as $H(X) = -\sum_{x \in D(X)} P(x) log_2 P(x)$. The higher the entropy is, the more uniform the PMF over X is. Thus there is more uncertainty since all outcomes of X seem similarly probable. On the other hand when entropy is close to zero there is only a little uncertainty about X's value.

4. Believable mistakes in recall result from forgetting and the inference process in retrieval. It can happen that there was an action a at time t' and during storage the mem for t' was forgotten. Later in retrieval, that is when computing PMF $f_{t'} = P_{\mathcal{M}}(O_{t'}|mems)$, the value had to be deduced from remembered mems and the probabilistic model $\mathcal{M}$ that includes the episodic schemata. If action b is more probable under this assumption ($P_{\mathcal{M}}(O_{t'} = b|mems) > P_{\mathcal{M}}(O_{t'} = a|mems)$), b will be recalled instead of a. There is no specific process for this feature, it is DyBaNeM's emergent property.

5. Interestingness of events is measured by KL divergence in the same way it is done by the encoding algorithm. The more different is a PMF predicted by the schemata from the recalled PMF the higher is the value of KL divergence. The first mem picked by the encoding algorithm is the one that deviates most from the prediction from schema. Subsequent mems contain less and less information. Thus if an IVA wants to communicate the interesting events first it can start with the first mem followed by the second and so on. If both the IVA and the human player have the same episodic schemata they will be able to reconstruct the same episodes. This is similar to function of lossy compression algorithms. DyBaNeM gets observed episode on the input, then it transforms the episode into a list of mems that is shorter than the original episode. With the use of the episodic schemata the mems can be used to reconstruct the episode. However some details of the episode might be changed due to forgetting and imperfect schemata. The difficulty in interpreting DBN*EM* as a compression algorithm is that not only mems but also the episodic schemata θ has to be stored (or transmitted). Since storage of θ requires far more space than storage of one mem this approach is feasible only if large number of episodes will have to be stored. On the other hand θ does not have to be transmitted if both parties, i.e. Bob and Alice, have already the same schemata.

6. Influence of personal experience follows from a different set of episodic schemata of each IVA. When Bob's schemata are trained on a different corpus of examples than Cloe's, the resulting probabilistic models will be also different. Thus inferences from these models may give different mems.

4 Prototype DyBaNeM Connection to an IVA

To demonstrate DyBaNeM's applicability to the domain of IVAs we connected it to an IVA whose behavior resembles a background character from a MMORPG. We show 1) that DyBaNeM can learn the schemata, store and recall one day of IVA's activity and 2) that it can support the dialog enhancing features discussed in Sec. 2. We also show 3) that the method has reasonable computational time requirements given domains of moderate complexity, even though the problem of exact inference in Bayesian Network is exponential in the network's treewidth.

Activity Dataset. As input for the EM model, we generated hierarchical activity dataset simulating 23 "days" of an IVA's life. The IVA was controlled by hierarchical decision making system (DMS) based on AND-OR trees formalism. An AND-OR tree describes decomposition of IVA behavior into goals and subgoals with possible alternatives of accomplishing each goal. The IVA's nondeterministic scheduling algorithm together with nondeterminism originating from the 3D simulation result in irregularities of stream of actions produced for each day. In our previous work we compared various statistical properties of the generated behavior to datasets of human behavior with reasonable match [17]. Our IVA is connected to a 3D virtual environment of Unreal Tournament 2004[3]. The agent was implemented in Java and the Pogamut platform [12] was used as a middleware for interfacing the IVA with the environment.

Every simulated day has a similar structure, the IVA gets up at home, he brushes teeth, washes face, goes to the toilet; then he usually goes to work; in the evening he may go to a theater or to a pub. He may also do shopping, clean the house and other activities resembling a normal life. In total the simulation contains 37 different types of atomic actions and 19 types of first level episodes. The generated stream of actions contains more levels of episodes but for this evaluation we use only the first level of episodes which is sufficient for

[3] Epic Games, 2004, [7.4.2013], `http://web.archive.org/web/20060615184746/http://www.unrealtournament2003.com/`

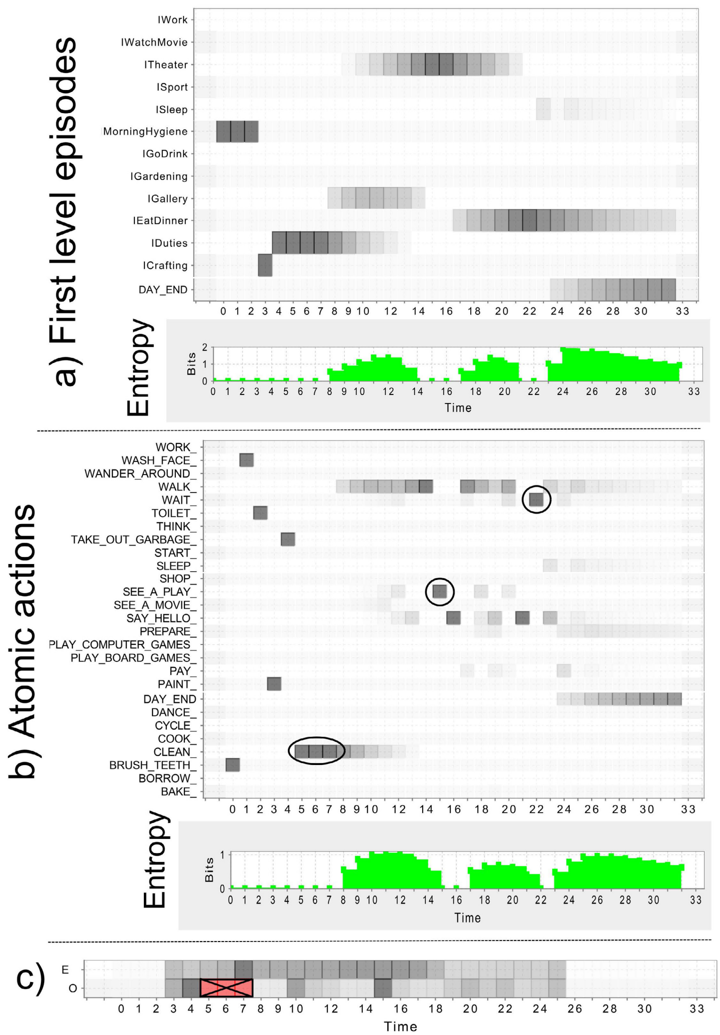


Fig. 4. Recall of the stored day when three mems are used for reconstruction. The mems are atomic actions (subfig. b), in ellipses). Subfig. **a)** shows selected high level episodes recalled for the day of interest. Level of gray indicates probability of each atomic action/episode at that time step. The darker the color is the more probable the action/episode is. This corresponds to the feature **1**. Entropy shows how certain the IVA is about his memory for those events (feature **3**). The more alternative memories are there the higher the entropy is. Subfig. **b)** shows probability and entropy of selected atomic actions. This is the second level of hierarchy that allows for clarifying questions (feature **2**). Subfig. **c)** shows KL divergence of all episodes (first line) and actions (second line) in the day compared to the prior episodic schema (feature **5**). The most interesting actions are marked by a cross (gray coding as in the case of probability). Most of the interesting actions/events become mems. Feature **4** is demonstrated by "fuzzy" transition around time 10: the model is not sure when exactly happened the switch from household duties to a gallery visit.

demonstrating all the features. There are different plans for working days and for weekends, that increases variability of the the IVA's episodic schemata. Not all days contained the same number of the atomic actions, the longest one has 33 actions. To make all days equal in size we added a sufficient number of padding actions *DAY_END* to the end of each day. Details of IVA's DMS are provided in [17] (sec 3.2).

Method. Twenty-two days were used for learning episodic schemata. The underlying probabilistic model CHMM was learned by counting the sufficient statistics [18]. The 23rd day was stored in DyBaNeM to demonstrate its abilities. The model was presented with the atomic actions from that day and it had to deduce high level episodes and store the observations. The result of the encoding process were 3 mems. This way we model Bob's EM for Alice's activity. For belief propagation in DyBaNeM's DBNs, SMILE[4] reasoning engine for graphical probabilistic models was used.

Results. When using only one mem to reconstruct the whole day, 52% of atomic actions were correctly recalled, with two mems it was 64%, and with three mems 73%. This means that when all three mems were used the most probable action in 73% of time steps matched the real action previously encoded. Recall of the stored day when all three mems were used is shown on Fig. 4. Learning the episodic schemata took 2 seconds, computing the first 3 mems for the stored day took 1.3 second on one core of P8600 2.4GHz, 1.5GB RAM.

Discussion. The evaluation indicates that computational cost is reasonable. Learning the schemata is done only once off-line and time necessary for encoding (1.3s) is also acceptable, though the domain is of moderate complexity only. Fig. 4 illustrates all the features 1-5. To demonstrate feature 6 we would need a second IVA with a different lifestyle that could be used to learn another set of episodic schemata. This extension is trivial but we omit it for space restrictions. The 73% recall accuracy is a reasonable starting point: it can be increased with more mems stored, and a user study, a future work, will indicate what accuracy is most welcomed by users.

Extending the IVA with ByDaNeM is a simple task that requires a developer only to: a) get logs of IVA's behavior that were used for episodic schemata learning, b) decide when to store episodes (e.g. at the end of the day) and c) decide when to recall the episode. Thus no advanced knowledge of DyBaNeM's internals is needed by the IVA developer.

5 Conclusion

We have demonstrated that bayesian approach to IVA—EM modeling, exemplified on our new DyBaNeM framework, is promising and it can be considered by developers of IVAs with EM abilities as a possible development method. To investigate scalability of this approach, we are presently experimenting also with

[4] SMILE was developed by the Decision Systems Laboratory of the University of Pittsburgh and is available at `http://genie.sis.pitt.edu/`.

larger domains, including human corpora, and different underlying DB representations. Our most recent evaluation data actually must be omitted for space limitations. A key future step is a user evaluation of the framework.

Acknowledgments. This research was partially supported by SVV project number 267 314 and by grant GACR P103/10/1287. We thank to xkcd.org for drawings of Alice and Bob.

References

1. Anderson, J.: A Spreading Activation Theory of Memory. Journal of Verbal Learning and Verbal Behavior 22, 261–295 (1983)
2. Bartlett, F.: Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge (1932)
3. Bickmore, T., Schulman, D., Yin, L.: Engagement vs. Deceit: Virtual Humans with Human Autobiographies. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 6–19. Springer, Heidelberg (2009)
4. Blaylock, N., Allen, J.: Fast Hierarchical Goal Schema Recognition. In: Proceedings of the National Conference on Artificial Intelligence (AAAI 2006), pp. 796–801 (2006)
5. Brom, C., Burkert, O., Kadlec, R.: Timing in Episodic Memory for Virtual Characters. In: Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, pp. 305–312 (2010)
6. Brom, C., Pešková, K., Lukavský, J.: What Does Your Actor Remember? Towards Characters with a Full Episodic Memory. In: Cavazza, M., Donikian, S. (eds.) ICVS 2007. LNCS, vol. 4871, pp. 89–101. Springer, Heidelberg (2007)
7. Bui, H.: A general model for online probabilistic plan recognition. In: International Joint Conference on Artificial Intelligence, pp. 1309–1318 (2003)
8. Deutsch, T., Gruber, A., Lang, R., Velik, R.: Episodic memory for autonomous agents. In: Human System Interactions (HSI 2008), pp. 621–626 (2008)
9. Dias, J., Ho, W.C., Vogt, T., Beeckman, N., Paiva, A., André, E.: I Know What I Did Last Summer: Autobiographic Memory in Synthetic Characters. In: Paiva, A., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 606–617. Springer, Heidelberg (2007)
10. Dodd, W., Gutierrez, R.: The Role of Episodic Memory and Emotion in a Cognitive Robot. In: IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2005, pp. 692–697 (2005)
11. Gallo, D.: Associative illusions of memory: False memory research in DRM and related tasks. Psychology Press (2006)
12. Gemrot, J., Kadlec, R., Bída, M., Burkert, O., Píbil, R., Havlíček, J., Zemčák, L., Šimlovič, J., Vansa, R., Štolba, M., Plch, T., Brom, C.: Pogamut 3 Can Assist Developers in Building AI (Not Only) for Their Videogame Agents. In: Dignum, F., Bradshaw, J., Silverman, B., van Doesburg, W. (eds.) Agents for Games and Simulations. LNCS, vol. 5920, pp. 1–15. Springer, Heidelberg (2009)
13. Ho, W.C., Dautenhahn, K., Nehaniv, C.: Computational Memory Architectures for Autobiographic Agents Interacting in a Complex Virtual Environment: A Working Model. Connection Science 20(1), 21–65 (2008)

14. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research 49(10), 1295–1306 (2009)
15. Kadlec, R., Brom, C.: DyBaNeM: Bayesian Framework for Episodic Memory Modelling. In: The 12th International Conference on Cognitive Modelling, ICCM (in press, 2013)
16. Kadlec, R., Brom, C.: Unifying episodic memory models for artificial agents with activity recognition problem and compression algorithms: review of recent work and prospects. Tech. rep. (2013)
17. Kadlec, R., Čermák, M., Behan, Z., Brom, C.: Generating Corpora of Activities of Daily Living and Towards Measuring the Corpora's Complexity. In: Dignum, F., Brom, C., Hindriks, K., Beer, M., Richards, D. (eds.) CAVE 2012. LNCS, vol. 7764, pp. 149–166. Springer, Heidelberg (2013)
18. Koller, D., Friedman, N.: Probabilistic graphical models: principles and techniques. The MIT Press (2009)
19. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 329–343. Springer, Heidelberg (2005)
20. Kullback, S.: Statistics and information theory (1959)
21. Lim, M.Y.: Memory models for intelligent social companions. In: Zacarias, M., de Oliveira, J.V. (eds.) Human-Computer Interaction. SCI, vol. 396, pp. 241–262. Springer, Heidelberg (2012)
22. Lim, M., Aylett, R., Enz, S., Ho, W.: Forgetting Through Generalisation — A Companion with Selective Memory. In: AAMAS 2011, pp. 1119–1120 (2011)
23. Nuxoll, A.M., Laird, J.E.: Enhancing intelligent agents with episodic memory. Cognitive Systems Research 17-18, 34–48 (2012)
24. Rabe, F., Wachsmuth, I.: Cognitively Motivated Episodic Memory for a Virtual Guide. In: ICAART, pp. 524–527 (2012)
25. Schank, R.C.: Dynamic memory revisited. Cambridge University Press (1999)
26. Subagdja, B., Wang, W., Tan, A.H., Tan, Y.S., Teow, L.N.: Memory Formation, Consolidation, and Forgetting in Learning Agents. In: AAMAS 2012, pp. 1007–1014 (2012)
27. Tulving, E.: Elements of Episodic Memory. Clarendon Press, Oxford (1983)
28. Čermák, M., Kadlec, R., Brom, C.: Towards modeling false memory using virtual characters: a position paper. In: Symposium on Human Memory for Artificial Agents, AISB, pp. 25–29 (2011)

Self-organizing Cognitive Models for Virtual Agents

Yilin Kang and Ah-Hwee Tan

School of Computer Engineering
Nanyang Technological University, Nanyang Avenue, Singapore 639798, kang0028
asahtan@ntu.edu.sg

Abstract. Three key requirements of realistic characters or agents in virtual world can be identified as autonomy, interactivity, and personification. Working towards these challenges, this paper proposes a brain inspired agent architecture that integrates goal-directed autonomy, natural language interaction and human-like personification. Based on self-organizing neural models, the agent architecture maintains explicit mental representation of desires, intention, personalities, self-awareness, situation awareness and user awareness. Autonomous behaviors are generated via evaluating the current situation with active goals and learning the most appropriate social or goal-directed rule from the available knowledge, in accordance with the personality of each individual agent. We have built and deployed realistic agents in an interactive 3D virtual environment. Through an empirical user study, the results show that the agents are able to exhibit realistic human-like behavior, in terms of actions and interaction with the users, and are able to improve user experience in virtual environment.

Keywords: Cognitive models, Virtual agents, Self-Organizing neural networks, Autonomy, Personality, Interactivity.

1 Introduction

Three key requirements of realistic characters or agents in virtual world can be identified as autonomy, interactivity, and personification [8]. However, most virtual worlds tend to constrain agents' actions to a very coarse level, dictated by hard coded rules [18,10]. In recent years, there has been growing interest in creating intelligent agents in virtual worlds that do not follow fixed scripts predefined by the developers, but instead react accordingly to actions performed by the players during their interaction. In order to achieve this objective, there have been approaches attempting to model the dynamic environments and user's immediate context [28,5,7]. However, they typically ignore a significant component of making the virtual world experience more intense and personalized for players, namely the capability for the agents to adapt over time to the environment and to the habits as well as eccentricity of a particular player.

Indeed, it has been a great challenge to develop intelligent learning agents that are able to adapt in real time and improve the interactivity and playability in virtual worlds. Learning in a virtual world, just like in the real world, poses many challenges for an agent, not addressed by traditional machine learning algorithms. In particular, learning in virtual world is typically unsupervised, without an explicit teacher to guide the agent in learning. Furthermore, it requires an interplay of a myriad of learning paradigms.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 29–43, 2013.

In this paper, we present a self-organizing neural model, named FALCON-X (Fusion Architecture for Learning and Cognition - eXtension), for creating intelligent learning agents in virtual worlds. By incorporating FALCON-X, an agent is able to learn from sensory and evaluative feedback signals received from the virtual environment. In this way, the agent needs neither an explicit teacher nor a perfect model to learn from. Performing reinforcement learning in real time, it is also able to adapt itself to the variations in the virtual environment and changes in the user behavior patterns.

The FALCON-X model is proposed based on an integration of the Adaptive Control of Thought (ACT-R) architecture [1] and the fusion Adaptive Resonance Theory (fusion ART) neural model [23]. Fusion ART is a generalization of self-organizing neural models known as Adaptive Resonance Theory (ART) [4]. By expanding the original ART model consisting of a single pattern field into a multi-channel architecture, fusion ART unifies a number of network designs supporting a myriad of learning paradigms, including unsupervised learning, supervised learning and reinforcement learning. A specific instantiation of fusion ART known as Temporal Difference-Fusion Architecture for Learning and Cognition (TD-FALCON) has shown to have competitive learning capabilities, compared with gradient descent based reinforcement learning systems [24]. While retaining the structure of the visual, manual, intentional and declarative modules of ACT-R, FALCON-X replaces the symbolic production system with a fusion ART neural network serving as the core inference area for fusing and updating the pattern activities in the four memory buffers. In addition, a critic channel is incorporated to regulate the attentional and learning processes of the core inference area.

FALCON-X may potentially be used to model a wide range of cognitive processes. In this paper, we describe how behavior models can be learned as sensory-motor mappings through reinforcement learning. We have developed learning personal agents using FALCON-X in a 3D virtual world called Co-Space. In this application, the learning personal agents are designed to befriend human users and proactively offer personalized services. Our experiments show that the agents are able to learn player models that evolve and adapt with player during run time. More importantly, the user study shows that the use of intelligent agents can improve user experience in the virtual world.

The rest of this paper is organized as follows: After a brief review of related work in section 2, we present the FALCON-X architecture in section 3. Section 4 describes the generic FALCON-X dynamics, followed by how it may be used to learn procedural knowledge and behaviour model in section 5. Section 6 presents the embodiment of FALCON-X in an integrated agent architecture. The evaluative experiments on the Co-Space simulated domain is reported in section 7. The final section concludes with a highlight of our future direction.

2 Related Work

2.1 Intelligent Virtual Agents

Intelligent agents have been popularly used for improving the interactivity and playability of virtual environment and games. However, most such agents are based on scripts or predefined rules. For example, in the Virtual Theater project, synthetic actors who portray fictive characters are provided by improvising their behaviors. The agents are based on a scripted social-psychological model which can define personality traits that depend on the values of moods and attitudes [18]. Agents in Metaverse, which was built using Active Worlds, are capable of taking tickets for rides, acting as shopkeepers or other tasks typically associated with humans. However, these agents are basically reactive agents which work in a hard-wired stimulus-response manner. Virtual psychotherapist ELIZA [26], although not even trying to understand its 'patients', often managed to make them feel taken care of, thus demonstrating the effects achievable with rule-based, adeptly modelled small talk. A conversational virtual agents Max has been developed as a guide to the HNF computer museum, where he interacts with visitors and provides them with information daily [10]. However, the design remains rule-based.

In view of the limitations of static agents, some researchers have adopted learning methods into agents in virtual environment. For example, Yoon et.al. present a Creature Kernel framework to build interactive synthetic characters in the project Sydney K9.0 [28]. Their agents can reflect the characters' past experience and allow individual personalization. But all the capabilities of the agents rely on past knowledge and couldn't adapt to user gradually during run time. To name the most elaborated one, ALICE [25] utilizes a knowledge base containing 40000 input response rules concerning general categories, augmented with knowledge modules for special domains like Artificial Intelligence. This approach has also been employed in other domains, e.g., to simulate co-present agents in a virtual gallery [5]. More recently, an embodied conversational agent that serves as a virtual tour guide in Second Life has been implemented by Jan [7]. Although it learns from past experience, it does not adapt over time according to the habits of a particular player or the changes in the environment.

All the work described above have developed a wide range of agents in virtual world with specific motivations. However, to the best of our knowledge, there have been very few, if any, agents that perform reinforcement learning in real time and can adapt their actions and behaviour during their interaction with the user and environment in virtual world. Our work is motivated by these considerations.

2.2 Cognitive Models

In the fields of artificial intelligence and cognitive science, there has been a debate over symbolic and sub-symbolic (connectionist) representation of human cognition [9], motivating two parallel streams of research directions. The symbolic field holds the view that, the human cognitive system uses symbols as a representation of knowledge and intelligence is through the processing of symbols and their respective constituents. Soar [11], ACT-R [1], and ICARUS [12], for example, are representative systems taking the symbolic approach.

On the other hand, the sub-symbolic camp argues that the human cognitive system uses a distributed representation of knowledge and is capable of processing this distributed representation of knowledge in a complex and meaningful way [6]. Sub-symbolic or connectionist systems are most generally associated with the metaphor of neural models, composing of neural circuits that operate in parallel. The key strengths of sub-symbolic systems lie in their learning abilities and allowance for massively parallel processing.

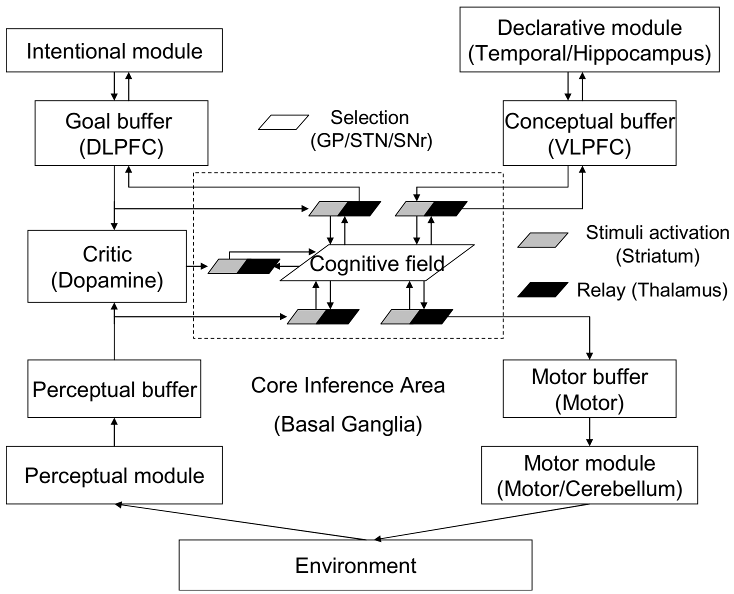


Fig. 1. The FALCON-X architecture

In view of their complementary strengths, there have been great interests in hybrid architectures that integrate high level symbolic systems with sub-symbolic massively parallel processes. Some examples are CLARION [20] and ACT-R with sequence learning [13]. Among the hybrid systems, temporal difference learning using gradient descent based function approximator has been commonly used. However, gradient descent methods typically learn by making small error corrections iteratively. In addition, instability may arise as learning of new patterns may erode the previously learned knowledge.

3 The FALCON-X Architecture

The FALCON-X architecture is presented herein, based on an integration of the ACT-R cognitive architecture and the fusion ART neural model (Figure 1). While retaining the structure of the visual, manual, intentional and declarative modules of ACT-R as the peripheral memory modules, the proposed architecture replaces the symbolic production system with a fusion ART neural network serving as the core inference area for fusing and regulating the pattern activities in the four memory buffers. Furthermore, the visual and manual modules are renamed as the perceptual and motor modules respectively, for the purpose of generality. As a key departure from ACT-R, an explicit critic module is also incorporated, which provides reward signals to the core inference area.

The roles and functions of the various peripheral modules are briefly elaborated as follows.

- The **Perceptual Module** receives input signals from the external environment. In actual applications, some preprocessing of the input signals may be necessary. The input signals are typically represented as a set of vectors of values in the perceptual buffer, taken from the sensors.
- The **Motor Module** receives and executes the actions, produced by a readout action from the core inference area. The actions are typically represented as a set of discrete values in the motor buffer, each of which denotes one of the possible actions.
- The **Intentional Module** consists of the task-relevant goals serving as the context. Each goal is represented as a target state vector in the goal buffer, representing the active goals of the agent.
- The **Declarative Module** consists of middle-term and long-term memories, relevant to the tasks. The memory can be represented in many ways. For example, it can be a look-up table or a neural network.
- The **Critic Module** computes reward signals that indicate the goodness of the actions taken. Generally, there can be two type of critics, namely, reward signals received from the external environment; and estimated payoff computed based on the current states and the target states.
- The **Core Inference Area** receives activations from the five memory modules and acts as a key driver of the inference process. In ACT-R, the production system operates in three processing steps: matching, selection and execution. In FALCON-X, the inference mechanism is realized via a five-step bottom-up and top-down neural processes, namely code activation, code competition, activity readout, template matching and template learning, described in the next section.

The design of FALCON-X is motivated by the neural anatomy of human brains. The core inference area of FALCON-X can be related to *basal ganglia* [16], which are a group of nuclei in the brain interconnected with the cerebral cortex and brainstem. Basal ganglia are important as they have been found to be associated with a variety of cognitive functions, including motor control, cognition, emotions and learning. The main components of basal ganglia includes striatum, globus pallidus (GP), subthalamic nucleus (STN), substantia nigra pars reticulata (SNr) and dopaminergic (DA) neurons.

The cognitive field in FALCON-X, employed for code selection, corresponds to the combined functionality of GP, STN and SNr, as supported in the literatures [17,3]. While ACT-R relates the pattern matching function of the production system to striatum, FALCON-X identifies striatum as the memory fields for stimuli presentation and pattern matching. While ACT-R associates thalamus to the execution function, thalamus is deemed to serve as a relay for motor commands in FALCON-X. Each pattern field of the FALCON is thus considered as a functional combination of striatum and thalamus. The neural substrates of the perceptual, motor, intentional and declarative modules have been discussed extensively in the context of ACT-R [1]. The new critic module in FALCON-X mirrors the dopamine neurons, whose phasic responses are observed when an unexpected reward is presented and depressed when expected reward is omitted [19].

4 The FALCON-X Dynamics

As a natural extension of ART, FALCON-X responds to incoming patterns in a continuous manner. In each inference cycle, the core inference area of FALCON-X receives input signals from the perceptual, intentional and declarative modules, and selects a cognitive node based on a bottom-up code activation and competition process. Whereas the intentional buffer maintains the active goals, the declarative module provides the relevant conceptual memory for code selection. The inference engine may also receive reward signals from the critic module. It is important to note that at any point in time, FALCON-X does not require input to be present in all the pattern channels. For those channels not receiving input, the input vectors are initialized to all 1s.

Upon activity readout, a template matching process takes place to ensure that the matched patterns in the four memory modules satisfy their respective criterion. If so, a state of resonance is obtained and the template learning process encodes the matched patterns using the selected cognitive node. Otherwise, a memory reset occurs, following which a search for another cognitive node begins. During prediction or action selection, the readout patterns typically include the actions to be executed in the motor module. In other cases, the conceptual memory buffer is updated and the goals may change as a result of inference.

The detailed dynamics of the inference cycle, consisting of the five key stages, namely code activation, code competition, activity readout, template matching, and template learning, are presented as follows.

Input vectors: Let $\mathbf{I}^{ck} = (I_1^{ck}, I_2^{ck}, ..., I_n^{ck})$ denote the input vector, where $I_i^{ck} \in [0,1]$ indicates the input i to channel ck. With complement coding, the input vector $\mathbf{I}^{ck}$ is augmented with a complement vector $\overline{\mathbf{I}}^{ck}$ such that $\overline{I}_i^{ck} = 1 - I_i^{ck}$.

Activity vectors: Let $\mathbf{x}^{ck}$ denote the F_1^{ck} activity vector for $k = 1, ..., K$. Let $\mathbf{y}$ denote the F_2 activity vector.

Weight vectors: Let $\mathbf{w}_j^{ck}$ denote the weight vector associated with the jth node in F_2 for learning the input patterns in F_1^{ck} for $k = 1, ..., K$. Initially, F_2 contains only one *uncommitted* node and its weight vectors contain all 1's.

Parameters: The fusion ART's dynamics is determined by choice parameters contribution parameters $\gamma^{ck} \in [0,1]$ and vigilance parameters $\rho^{ck} \in [0,1]$ for $k = 1, ..., K$.

Code activation: Given the activity vectors $\mathbf{I}^{c1}, ..., \mathbf{I}^{cK}$for each F_2 node j, the choice function T_j is computed as follows:

$$T_j = \sum_{k=1}^{K} \gamma^{ck} \frac{|\mathbf{I}^{ck} \wedge \mathbf{w}_j^{ck}|}{\alpha^{ck} + |\mathbf{w}_j^{ck}|}, \tag{1}$$

where the fuzzy AND operator $\wedge$ is defined by $(p \wedge q)_i \equiv \min(p_i, q_i)$, and the norm $|.|$ is defined by $|\mathbf{p}| \equiv \sum_i p_i$ for vectors $\mathbf{p}$ and $\mathbf{q}$.

Code competition: A code competition process follows under which the F_2 node with the highest choice function value is identified. The winner is indexed at J where

$$T_J = \max\{T_j : for\ all\ F_2\ node\ j\}. \tag{2}$$

When a category choice is made at node J, $y_J = 1$; and $y_j = 0$ for all $j \neq J$. This indicates a winner-take-all strategy.

Activity readout: The chosen F_2 node J performs a readout of its weight vectors to the input fields F_1^{ck} such that

$$\mathbf{x}^{ck} = \mathbf{I}^{ck} \wedge \mathbf{w}_J^{ck}. \tag{3}$$

Template matching: Before the activity readout is stabilized and node J can be used for learning, a template matching process checks that the weight templates of node J are sufficiently close to their respective input patterns. Specifically, resonance occurs if for each channel k, the *match function* m_J^{ck} of the chosen node J meets its vigilance criterion:

$$m_J^{ck} = \frac{|\mathbf{I}^{ck} \wedge \mathbf{w}_J^{ck}|}{|\mathbf{I}^{ck}|} \geq \rho^{ck}. \tag{4}$$

If any of the vigilance constraints is violated, mismatch reset occurs in which the search process selects another F_2 node J until a resonance is achieved.

Template learning: Once a resonance occurs, for each channel ck, the weight vector $\mathbf{w}_J^{ck}$ is modified by the following learning rule:

$$\mathbf{w}_J^{ck(new)} = (1 - \beta^{ck})\mathbf{w}_J^{ck(old)} + \beta^{ck}(\mathbf{I}^{ck} \wedge \mathbf{w}_J^{ck(old)}). \tag{5}$$

When an uncommitted node is selected for learning, it becomes *committed* and a new uncommitted node is added to the F_2 field. FALCON thus expands its network architecture dynamically in response to the input patterns.

5 Learning Procedural Knowledge

In this section, we illustrate how FALCON-X, specifically the core inference area together with the perceptual, motor and critic modules, can acquire procedural knowledge through reinforcement learning in a dynamic and real-time environment.

FALCON-X learns mappings simultaneously across multi-modal input patterns, involving states, actions, and rewards, in an online and incremental manner. Various strategies are available for learning in FALCON-like architectures. We highlight two specific methods, namely reactive learning and temporal difference learning as follows.

5.1 Reactive Learning

A reactive learning strategy, as used in the R-FALCON (Reactive FALCON) model [21], performs fast association between states and actions, based on reward signals. Given a reward signal (positive feedback) in the critic buffer, FALCON associates the current state in the perceptual buffer with the selected action represented in the motor buffer. If a penalty is received, it learns the mapping among current state, the complement pattern of the action taken and the complement value of the given reward.

Table 1. The TD-FALCON algorithm with direct code access

1. Initialize the FALCON network.
2. Sense the environment and formulate a state vector **S** based on the current state s.
3. Following an action selection policy, first make a choice between exploration and exploitation. If exploring, take a random action.
   If exploiting, identify the action a with the maximal $Q(s,a)$ value by presenting the state vector **S**, the action vector **A**=(1,...1), and the reward vector **R**=(1,0) to FALCON.
4. Perform the action a, observe the next state s', and receive a reward r (if any) from the environment.
5. Estimate the revised value function $Q(s,a)$ following a Temporal Difference formula such as $\Delta Q(s,a) = \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$.
6. Formulate action vector **A** based on action a and reward vector **R** based on $Q(s,a)$.
7. Present the corresponding state, action, and reward vectors **S**, **A**, and **R** to FALCON for learning.
8. Update the current state by s=s'.
9. Repeat from Step 2 until s is a terminal state.

5.2 Temporal Difference Learning

A key limitation of reactive learning is the reliance on the availability of immediate reward signals. TD-FALCON [24,22] is a variant of FALCON that incorporates Temporal Difference (TD) methods to estimate and learn value functions of action-state pairs $Q(s,a)$ that indicates the goodness for a learning system to take a certain action a in a given state s. Such value functions are then used in the action selection mechanism, also known as the *policy*, to select an action with the maximal payoff. The temporal difference learning algorithm is summarized in Table 1.

Given the current state s, TD-FALCON first decides between exploration and exploitation by following an action selection policy. For exploration, a random action is picked. For exploitation, TD-FALCON performs instantaneous searches for cognitive nodes that match with the current states and at the same time provide the highest reward values using a direct access procedure. Upon receiving a feedback from the environment after performing the action, a TD formula is used to compute a new estimate of the Q value of performing the chosen action in the current state. The new Q value is then used as the teaching signal for TD-FALCON to learn the association of the current state and the chosen action to the estimated Q value.

6 The Integrated Cognitive Agent Architecture

For modelling intelligent virtual agents, FALCON-X needs to be integrated with the necessary peripheral modules for interaction with the environment. As shown in Figure 2, the integrated agent architecture consists of a *Perception Module* receiving situational signals from the environment through a set of sensory APIs and an *Action Module* for performing actions through the various actuator APIs. If the sensory signals involve a text input, the *Chat Understanding Module* interprets the text for the player's intention. The outputs of *Situational Assessment* and *Chat Understanding Modules* then serve as part of the working memory content providing conditional attributes to the *Inference Engine*. The *Inference Engine* based on the FALCON-X model then identifies the most appropriate action, by tapping a diverse pool of knowledge, in accordance to the desire, intention and personality of the virtual agent. The knowledge learned and used by the Inference Engine include declarative knowledge of self, players, and environment, as well as procedural knowledge of goal-oriented rules, which guide an agent in fulfilling goals, and social rules, for generating socially appropriate behavior. The decision of the *Inference Engine* again forms part of the *Working Memory*, which throughout maintains the context of the interaction. For actions involving a verbal response, the *Natural Language Generation Module* translates the chosen response into natural text for presentation.

Consistent with the view in the state of the art [8], we outline three key characteristics of realistic characters in virtual worlds, namely autonomy, interactivity, and personification, described as follows.

Autonomy. Based on a family of self-organizing neural models known as fusion Adaptive Resonance Theory (ART) [23], the *Inference Engine* of the proposed agent architecture performs a myriad of cognitive functions, including recognition, prediction and learning, in response to a continual stream of input signals received from multiple pattern channels. As a result, an agent makes decisions not only based on the situational factors perceived from the environment but also her mental states characterized by desire, intention and personality. By modelling the internal states of individual agents explicitly, the virtual humans can live a more complete and realistic life in the virtual world.

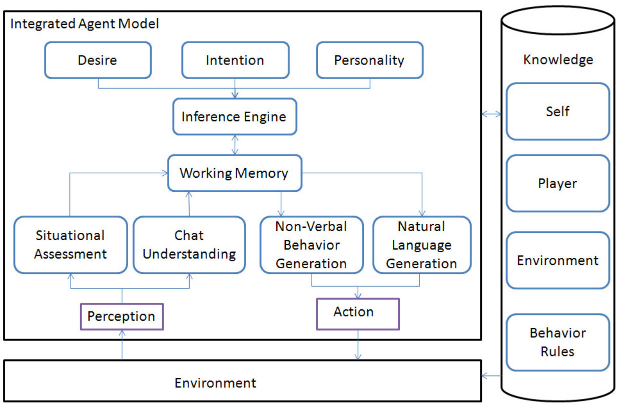


Fig. 2. A schematic of the integrated agent model

Interactivity. For interaction between the agents and the players, an intuitive user interface is provided, through which a player may ask typical questions and provide quick responses by button clicks. The player may also enter free-text sentences via the chat box. The dual communication mode provides the players both ease of use and flexibility. While interacting with player, the agent builds an internal model of the player, with his/her profile, interests and preferences. The player model in turns allows the agent to make intelligent conversation on topics relevant to the player.

Personification. For improving the believability of virtual humans, our agents adopt the Five Factor Model (FFM) [14], which characterizes personality in five trait dimensions. By giving a weighage to each dimension, a unique personality can be formed by a combination of the traits. Comparing with traditional pattern-matching-based conversational agents, our agents with strong *openness* and *extroversion* personality are warmer and friendlier as they do not stay idle and wait for input queries. Acting pro-actively, they approach the players, offer help, and make conversation.

7 Evaluative Experiments

7.1 Research Methodology

We developed three versions of NTU Co-Space, each with a distinct type of virtual agents in the form of Non-Player Characters (NPCs). The first environment (E1) provides the baseline control condition, wherein the NPCs are only able to display static messages but do not have the capability to interact with the users. The second environment (E2) is the first treatment condition, wherein the virtual humans are designed as embodied conversational agents using the Artificial Intelligence Mark-Up Language (AIML) [25]. AIML is an XML-compliant language that was developed by the Alicebot free software community. It is considered as a rule based repository of knowledge where the engine performs pattern matching to select the most appropriate output based on the input utterance. We have encoded as many AIML patterns as possible to enhance the conversational abilities of the agents. The third environment (E3) is the second treatment condition, wherein autonomous agents using our proposed fusion ART-based agent model are populated. Although the agents we described in the previous sections have different personalities, for the purpose of this study, we remove the variation in personality by deploying only friendly agents.

The subjects were recruited from an Introduction to Management Information Systems class at a large US university. The scenario given to the subjects is that they were looking for an overseas university for an exchange program and were visiting NTU Co-Space to help them in making the decision. Each subject was asked to complete a quest in the form of a mini-game, where they would experience the key places of the NTU campus through the quest. The quest involves finding five check-points on campus where the clue to each check-point was given at the previous check-point.

The objectives of the experiment are two-fold. First, we observe whether deploying virtual humans in the virtual world will benefit the player's experience. Second, we assess how virtual humans with different levels of intelligence may affect the player's experience, especially in terms of the following constructs, namely *Telepresence* (TP), *Social Presence* (SP), *Perceived Interactivity* (PI), *Perceived Usefulness* (PU), *Flow* (FLW), *Enjoyment* (ENJ) and *Behavioral Intention* (to return to NTU Co-Space) (BI).

Subjects participated in the experiment in a computer lab. They were asked to fill out a pre-questionnaire (used to assess the players' profile, covering demographics information and 3D virtual world skills) and then carry out the experiment by completing the quest given to them. For subjects in the E1 (control) condition, they completed the quest by using the map in the system to navigate the virtual world, check up information on different parts of campus and teleport to the respective checkpoints without receiving any help from NPCs. For subjects in the E2 (i.e., first treatment) condition, they were not only provided with the interactive map in the E1 condition, but they could also talk to the embodied conversational agents to ask for assistance before teleporting through the interactive map. For subjects in the E3 (i.e., second treatment) condition, in addition to being provided with the interactive map, they were also offered the assistance of fusion ART-based NPCs that have the ability of performing autonomous behaviors both in proactive and responsive ways; moreover, since these NPCs are embedded with a Natural Language Processing module, they can understand input sentences in a flexible way. Hence, subjects were able to interact with the intelligent autonomous NPCs to request for and obtain the information they needed. Because the NPCs are autonomous, they could even offer teleport service to the specific locations requested. After the subjects have completed the quest, they filled out a post-questionnaire which assessed their experience. The subjects were first asked for the acronym followed by the full name of the university to assess their level of recall. The main part of the questionnaire then captured the subjects' assessment of the seven constructs (as described earlier) related to NPC functions. In addition, 3D virtual world skills were also captured to examine their perceived improvement of skills after experiencing NTU Co-Space.

Among the various constructs, *Perceived Usefulness* could be objectively assessed through the time taken to complete the Amazing Quest. The less time spent to complete the Amazing Quest, the more useful the agents are. *Flow* was captured through providing the description of *Flow*, and then asking subjects to rate the degree of *Flow* they experienced and the frequency in which they experienced *Flow*. The other constructs were captured using measurement items. At least five items were used to measure each construct. The items used to measure *Telepresence*, *Enjoyment* and *Behavioral Intention* were derived from Nah [15]. The scale for measuring *Social Presence* and *Interactivity* were adopted from Animesh et al [2]. Given the limited space, we present a sample set of these items in Table 2.

Table 2. A sample set of Post-Questionnaire

Item	Measurement
TP 1	I forgot about my immediate surroundings when I was navigating in the virtual world.
SP 1	During the virtual tour, the interaction with the virtual humans were warm.
ENJ 1	I found the virtual tour to be fulfilling.
BI 1	I would consider visiting this virtual world site again.

7.2 Data Analysis

Overall Performance: Table 3 shows the overall performance in the three environments. We observe that subjects in all the three environments were able to complete the quest successfully. However, subjects in E3 spent the least amount of time and the percentage of subjects who could correctly recall the acronym of the campus is higher than those of the other two environments. This indicates that the autonomous agents deployed in E3 are more useful in helping the subjects than those of the other two environments.

Table 3. overall performance

Evaluation Measures	E1	E2	E3
% of players complete the quest	100%	100%	100%
Time to complete the quest	20m 31s	25m 32s	**16m 56s**
% of Players recall the acronym of the campus	44%	25%	**45%**

Descriptive Statistics: Table 4 shows the means, standard deviations (SD), and confidence intervals (CI, with a confidence level of 95%) of ratings in E1, E2, and E3 in terms of *Telepresence* (TP), *Social Presence* (SP), *Perceived Interactivity* (PI), *Enjoyment* (ENJ), *Flow* (FLW) and *Behavior Intention* (BI). All of the constructs were measured using the seven point Likert scale. From the table, we observe that for E3, the rating of *Telepresence*, *Social Presence*, *Perceived Interactivity*, *Flow* and *Behavior Intention* have better results than E1 and E2. This means by employing autonomous agents, these factors are perceived to be stronger than those in the environment with dummy and AIML based agents. However, we note that for enjoyment, the rating in E2 is the best of the three environments. Referring to the 3D virtual world skills assessed in the pre-questionnaire, subjects in E2 appear to have the best 3D virtual world skills compared to the other two. As prior work have shown that a higher level of skill is likely to enhance the feeling of enjoyment [27], we believe the rating obtained in E2 could be affected by the higher level of 3D virtual world skills.

Table 4. Descriptive Statistics

	E1			E2			E3		
Constructs	Mean	SD	CI	Mean	SD	CI	Mean	SD	CI
TP	3.86	0.34	3.64-4.08	4.08	0.49	3.76-4.40	**4.41**	0.42	4.14-4.68
SP	3.63	0.44	3.28-3.98	3.82	0.43	3.47-4.16	3.84	0.65	3.32-4.36
PI	4.66	0.67	4.13-5.19	4.37	0.58	3.90-4.83	**5.02**	0.66	4.49-5.54
ENJ	4.42	0.44	4.04-4.81	4.50	0.33	4.11-4.90	4.31	0.72	3.68-4.94
FLW	4.47	0.54	4.00-4.94	4.21	0.28	3.96-4.46	**4.58**	0.68	3.98-5.18
BI	4.23	0.71	3.61-4.56	4.51	0.30	4.24-4.78	4.58	0.40	4.23-4.93

Furthermore, a one-way analysis of variance (ANOVA) is used to analyze the results. Specifically, the F-test is used to evaluate the hypothesis of whether there are significant differences among the statistic data means for those constructs. The F values are calculated by the variances between conditions divided by the variance within the conditions. The p values, on the other hand, represent the probability of test statistic being different from the expected values and are directly derived from the F test. A small p value thus indicates a high confidence that the values of those constructs are different. A summary

Table 5. F-test result

	E1, E2 & E3		E1 & E3		E2 & E3	
Constructs	F	p	F	p	F	p
TP	**5.47**	**0.004**	11.22	0.001	4.55	0.034
SP	0.58	0.561	1.10	0.295	0.01	0.903
PI	**5.71**	**0.004**	4.00	0.047	12.12	0.001
ENJ	0.15	0.862	0.33	0.566	0.17	0.68
FLW	1.62	0.199	0.28	0.595	3.12	0.079
BI	1.23	0.294	2.40	0.120	0.10	0.751

of the F values and p values among E1, E2 and E3, between E1 and E3, and between E2 and E3 are given in Table 5.

This data analysis revealed the significant effects of the three kinds of virtual humans in virtual worlds on *Telepresence* and *Perceived Interactivity*: $F(2, 519) = 5.47, p < 0.01$ for *Telepresence* and $F(2, 345) = 5.71, p < 0.01$ for *Perceived Interactivity*, where the two parameters enclosed in parentheses after F indicate the degrees of freedom of the variances between and within conditions respectively. Consistent with the statistics in Table 4, the fusion ART-based virtual human generates higher levels of *Telepresence* and *Perceived Interactivity* than the other two types of virtual humans, with a mean of 4.41 for E3 (versus 3.86 for E1 and 4.08 for E2) for *Telepresence*, and a mean of 5.02 for E3 (versus 4.66 for E1 and 4.37 for E2) for *Perceived Interactivity*. Although the effect of E1, E2 and E3 on *Flow* is smaller than that of *Telepresence* and *Perceived Interactivity*, the difference in *Flow* is perceived to be marginally significant between E2 and E3, with $F(1, 198) = 3.12, p < 0.1$, and a mean of 4.58 for E3 (versus 4.21 for E2). This means the *Flow* experience perceived by subjects who interacted with the fusion ART-based virtual human is stronger than those interacting with the AIML based virtual humans. No significant difference was found for the rest of the constructs.

8 Conclusion

For creating realistic agents in virtual world, this paper has proposed a cognitive agent architecture that integrates goal-directed autonomy, natural language interaction and human-like personality. Extending from a family of self-organizing neural models, the agent architecture maintains explicit mental representation of desires, personalities, self-awareness, situation awareness and user awareness.

We have built and deployed realistic agents in an interactive 3D virtual environment. We have also carried out systematic empirical work on user study to assess whether the use of intelligent agents can improve user experience in the virtual world. Our user study has so far supported the validity of our agent systems. With the virtual characters befriending and providing personalized context-aware services, players generally found virtual world more fun and appealing. To the best of our knowledge, this is perhaps one of the few in-depth works on building and evaluating complete realistic agents in virtual worlds with autonomous behavior, natural interactivity and personification. Moving forward, we wish to extend our study by completing the agent architectures with more functionalities, such as emotion and facial expressions.

References

1. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111, 1036–1060 (2004)
2. Animesh, A., Pinsonneault, A., Yang, S.-B., Oh, W.: An odyssey into virtual worlds: Exploring the impacts of technological and spatial environments. MIS Quaterly 35, 789–810 (2011)
3. Bogacz, R., Gurney, K.: The basal ganglia and cortex implement optimal decision making between alternative actions. Neural Computation 19(2), 442–477 (2007)
4. Carpenter, G.A., Grossberg, S.: Adaptive Resonance Theory. In: The Handbook of Brain Theory and Neural Networks, pp. 87–90. MIT Press (2003)
5. Gerhard, M., Moore, D.J., Hobbs, D.J.: Embodiment and copresence in collaborative interfaces. Int. J. Hum.-Comput. Stud. 64(4), 453–480 (2004)
6. Haykin, S.: Neural Network: A Comprehensive Foundation. Prentice Hall (1999)
7. Jan, D., Roque, A., Leuski, A., Morie, J., Traum, D.: A virtual tour guide for virtual worlds. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 372–378. Springer, Heidelberg (2009)
8. Kasap, Z., Thalmann, N.: Intelligent virtual humans with autonomy and personality: State-of-the-art. Intelligent Decision Technologies 1, 3–15 (2007)
9. Kelley, T.D.: Symbolic and sub-symbolic representations in computational models of human cognition: What can be learned from biology? Theory and Psychology 13(6), 847–860 (2003)
10. Kopp, S., Gesellensetter, L., Krämer, N.C., Wachsmuth, I.: A conversational agent as museum guide – design and evaluation of a real-world application. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 329–343. Springer, Heidelberg (2005)
11. Laird, J.E., Newell, A., Rosenbloom, P.S.: Soar: An architecture for general intelligence. Artificial Intelligence 33, 1–64 (1987)
12. Langley, P., Choi, D.: A unified cognitive architecture for physical agents. In: Proceedings of 21st National Conference on Artificial Intelligence, pp. 1469–1474 (2006)
13. Lebiere, C., Wallach, D.: Sequence learning in the act-r cognitive architecture: Empirical analysis of a hybrid model. In: Sun, R., Giles, C.L. (eds.) Sequence Learning. LNCS (LNAI), vol. 1828, pp. 188–212. Springer, Heidelberg (2000)
14. Mccrae, R., Costa, P.: An introduction to the five-factor model and its applications. Journal of Personality 60, 172–215 (1992)
15. Nah, F., Eschenbrenner, B., DeWester, D.: Enhancing brand equity through flow and telepresence: A comparison of 2d and 3d virtual worlds. MIS Quaterly 35, 731–748 (2011)
16. O'Reilly, R.C., Frank, M.J.: Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation 18, 283–328 (2006)
17. Prescott, T.J., Gonzalez, F.M.M., Gurney, K., Humphries, M.D., Redgrave, P.: A robot model of the basal ganglia: Behavior and intrinsic processing. Neural Networks 19(1), 31–61 (2006)
18. Rousseau, D., Roth, B.: A social-psychological model for synthetic actors. In: Proceedings of 2nd International Conference on Autonomous Agents, pp. 165–172 (1997)
19. Schultz, W.: Getting formal with dopamine and reward. Neuron 36(2), 241–263 (2002)
20. Sun, R., Merrill, E., Peterson, T.: From implicit skills to explicit knowledge: a bottom-up model of skill learning. Cognitive Science 25(2), 203–244 (2001)
21. Tan, A.-H.: FALCON: A fusion architecture for learning, cognition, and navigation. In: Proceedings of International Joint Conference on Neural Networks, pp. 3297–3302 (2004)
22. Tan, A.-H.: Direct Code Access in Self-Organizing Neural Networks for Reinforcement Learning. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 1071–1076 (2007)

23. Tan, A.-H., Carpenter, G.A., Grossberg, S.: Intelligence through interaction: Towards a unified theory for learning. In: Liu, D., Fei, S., Hou, Z.-G., Zhang, H., Sun, C. (eds.) ISNN 2007, Part I. LNCS, vol. 4491, pp. 1094–1103. Springer, Heidelberg (2007)
24. Tan, A.-H., Lu, N., Xiao, D.: Integrating Temporal Difference Methods and Self-Organizing Neural Networks for Reinforcement Learning with Delayed Evaluative Feedback. IEEE Transactions on Neural Networks 9(2), 230–244 (2008)
25. Wallace, R.S.: The anatomy of A.L.I.C.E. Tech. report, ALICE AI Foundation (2000)
26. Weizenbaum, J.: ELIZA: a computer program for the study of natural language communication between men and machines. Communications of the ACM 9 (1996)
27. Yi, M.Y., Hwang, Y.: Predicting the use of web-based information systems: self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model. International Journal of Computer-Human Studies 59, 431–449 (2003)
28. Yoon, S., Burke, R.C., Blumberg, B.M., Schneider, G.E.: Interactive training for synthetic characters. In: AAAI, pp. 249–254 (2000)

Are You Thinking What I'm Thinking? An Evaluation of a Simplified Theory of Mind

David V. Pynadath[1], Ning Wang[2], and Stacy C. Marsella[1]

[1] University of Southern California, Institute for Creative Technologies
12015 E Waterfront Drive, Los Angeles, CA 90094 USA
[2] Curious Lab LLC
7430 Arizona Ave, Suite 35, Los Angeles, CA 90045 USA

Abstract. We examine the effectiveness of an agent's approximate theory of mind when interacting with human players in a wartime negotiation game. We first measure how accurately the agent's theory of mind captured the players' actual behavior. We observe significant overlap between the players' behavior and the agents' idealized expectations, but we also observe significant deviations. Forming an incorrect expectation about a person is not inherently damaging, so we then analyzed how different deviations affected the game outcomes. We observe that many classes of inaccuracy in the agent's theory of mind did not hurt the agent's performance and, in fact, some of them played to the agent's benefit. The results suggest potential advantages to giving an agent a computational model of theory of mind that is overly simplified, especially as a first step when investigating a domain with as much uncertainty as wartime negotiation.

Keywords: theory of mind, cognitive models, wartime negotiation, evaluation of formal models.

1 Introduction

Theory of mind is critical for success in social interaction [20]. Without it, people would not be able to understand each other's perspectives and desires. With it, people can form expectations of the others' behavior and choose their own behaviors informed by those expectations. Consequently, intelligent virtual agents (IVAs) also need theory of mind to successfully interact with people [5,9].

Researchers have used a variety of cognitive models to provide agents with this capability. A common approach for realizing theory of mind is for the agent to use its own reasoning mechanism as a model for the reasoning of others, after substituting the others' beliefs, goals, capabilities, etc. for its own. For example, an agent that uses partially observable Markov decision problems (POMDPs) [6] for its own decision-making can model other agents and people as using POMDPs of their own [1,4,10,14].

One of the challenges in implementing theory of mind within an IVA is its unavoidable uncertainty about the mental states of the others. For instance, even

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 44–57, 2013.

when playing a simple game like the Prisoner's Dilemma, an agent can never be sure what goals a human player is pursuing. An agent with a completely accurate theory of mind would have to capture whether players care about maximizing only their own payoff, or social welfare, fairness, and many other payoffs as well, not to mention the possible expectations they have about the agent's own behavior. In reality, a agent rarely has an accurate method for assessing the probability of all of these alternate hypotheses, so it is likely to make errors when making decisions based on these inaccurate assessments. Furthermore, there are computational costs associated with maintaining a richer, more accurate model, costs that can be hard to justify if the accuracy does not benefit the agent's own utility [15]. An IVA has to carefully balance this trade-off between complexity and accuracy when deciding how rich a theory of mind to use.

In this paper, we examine the effectiveness of an agent's theory of mind when interacting with people in a wartime negotiation game [16]. The game pits human players against agents implemented in PsychSim, a computational architecture for theory of mind that has been used to build IVAs in other negotiation domains [8,9,11]. The models use asymmetry of information to provide only one side (e.g., an agent) with complete information about the game's underlying likelihoods and costs, while leaving the other side (e.g., a human player) in uncertainty about those parameters. While the agent has the advantage of complete information, its performance is highly dependent on its ability to perform effective theory-of-mind reasoning about the human player's uncertainty. In our study, the agent's simplifying assumption that the player had no such uncertainty helped to achieve the initial goal of evaluating the coarse game-theoretic predictions of human behavior [16]. In this paper, we further analyze the data to study the impact the agent's approximate theory of mind had on its performance.

We begin our investigation of the agent's complexity/accuracy trade-off by measuring the accuracy of the agent's theory of mind. We observe significant overlap between the human players' actual behavior and the agents' idealized expectations, but we also observe significant deviations. Forming an incorrect expectation about a person is not *inherently* damaging, so we then analyzed how different deviations affected the game outcomes. We observe that many classes of inaccuracy in the agent's theory of mind did not hurt the agent's performance and, in fact, some of them played to the agent's *benefit*. The results suggest potential advantages to giving an agent a computational model of theory of mind that is overly simplified, especially as a first step when investigating a domain with as much uncertainty as wartime negotiation.

2 Wartime Negotiation

A number of formal models in the political science literature represent war as a costly process embedded within a negotiation game. In these models, two sides are in a dispute over the division of a desirable resource, which we will illustrate as territory claimed by both sides. The game begins with some initial split of the territory. The game progresses round by round, with each round consisting of one side proposing a split of the territory, the other side responding to that proposal, and a possible battle between the two. The game ends with a final split achieved by either an agreement on the proposed split or a decisive military victory by one side on the battlefield.

We chose two models, *Powell* [13] and *Slantchev* [18], for this investigation, based on their impact on the field and their appropriateness for a human-agent game interaction. Both models assume fixed probabilities associated with the battlefield, so that one side's probability of winning does not change during the course of the game, regardless of previous military outcomes. The costs of a single battle are also fixed throughout the course of the game. In our study, we present these costs to the human players in terms of troops lost.

A critical property of these models is uncertainty about the costs to the other side and the likelihood of battlefield outcomes. If both sides had complete information about the costs and probabilities, they could do an exact cost-benefit analysis and immediately agree upon a territorial split. In both models we implemented, only one side has complete information and the other is uncertain about the probability and costs of battlefield outcomes. This asymmetry lends itself to our human participant study, as we can give the agent complete information about the game probabilities and costs, but withhold that information from the human player. Even with complete information about the *game* uncertainty, the agent still needs to model the *players'* uncertainty and how it affects their decisions.

2.1 The Powell Model

In the following *Powell* model [13], Player 1 is a human player and Player 2 is an agent:

1. Player 1 makes an offer of $x\%$ of the territory.
2. Player 2 decides to accept, reject, or attack.
   (a) If accept, Player 2 gets $x\%$, Player 1 gets $(100 - x)\%$, and game ends.
   (b) If attack, Players 1 and 2 lose c_1 and c_2 troops, respectively. Player 1 collapses with probability p_1 and Player 2 collapses with probability p_2.
      i. If only Player 1 collapses, Player 2 gets 100%, and game ends.
      ii. If only Player 2 collapses, Player 1 gets 100%, and game ends.
      iii. Otherwise, return to Step 1.
   (c) If reject, Player 1 decides whether or not to attack.
      i. If attack, go to Step 2b; otherwise, return to Step 1.

Player 1 (the human player) does not have prior knowledge of the probabilities of collapse (p_i) or the costs of war (c_i), but Player 2 (the agent) does. In the game-theoretic analysis of this model, Player 2 can compute the optimal threshold, where it should accept any offer from Player 1 that exceeds the threshold and reject or attack otherwise. This threshold is lower the higher its costs (c_2), the higher its probability of collapse (p_2), and the lower Player 1's probability of collapse (p_1). Because Player 1 does not know these values, it is uncertain about Player 2's threshold for accepting an offer. The equilibrium behavior can be

Table 1. Game features across the four experimental conditions

	Powell28	**Powell72**	**Slantchev28**	**Slantchev72**
Start	28%	72%	28%	72%
War	A single battle can end the war in a win/loss for the player		A single battle changes military position $\in [0, 10]$, with 10 (0) being a win (loss) for the player	
Battle	Battle occurs only if either side unilaterally initiates		Battle occurs every round, *not* initiated by either side	
Offers	Agent cannot counteroffer		Agent must counteroffer	

described as *screening*, where Player 1 will make a series of increasingly attractive offers, expecting weaker opponents (who have a lower threshold) to accept early in the process, thus screening them out before making the higher offers necessary to appease stronger opponents [13].

2.2 The Slantchev Model

Unlike the Powell model's probability of collapse, the *Slantchev* model includes an additional variable, *military position* ($k \in \{0, 1, 2, \ldots, N\}$), that represents gradual military progress toward or away from complete collapse [18]. We again have a human as Player 1 and an agent as Player 2:

1. The initiating player makes an offer of $x\%$ of the territory.
2. The responding player decides to accept or reject the offer.
   (a) If accept, the responding player gets $x\%$, the initiating player gets $(100 - x)\%$, and game ends.
   (b) If reject, continue to Step 3.
3. Battle occurs, and Players 1 and 2 lose c_1 and c_2 troops, respectively. Player 1 wins the battle with probability p, Player 2 with probability $1 - p$.
   (a) If Player 1 wins, $k \leftarrow k + 1$. If $k = N$, then Player 1 gets 100% and game ends.
   (b) If Player 2 wins, $k \leftarrow k - 1$. If $k = 0$, then Player 2 gets 100% and game ends.
4. Return to Step 1 with initiating and responding players reversed.

Like the Powell model, Player 1 does not know the battle probability (p) or costs (c_i), but Player 2 does. Thus, the agent can compute a threshold for acceptable offers, but this threshold is now a function of k, the current military position. This threshold increases (and the agent's counteroffers decrease) as k decreases, c_2 decreases, and p decreases. As in the Powell model, the human players are uncertain about this threshold because of their ignorance of the underlying probability and cost, so the equilibrium behavior again exhibits some screening.

Examining human behavior in these two games allows us to study the effectiveness of the agent's theory of mind. In addition to the variation between our two game models, we also vary the players' starting territory (28% vs. 72%) to possibly shift their reference points in the negotiation [7,12]. These four combinations produce the four experimental conditions summarized in Table 1.

3 PsychSim Agents in Wartime Negotiation

We implemented both the Powell and Slantchev games within PsychSim, a multiagent framework for social simulation [10,14]. PsychSim agents have their own goals, private beliefs, and mental models about other agents. They generate their beliefs and behaviors by solving POMDPs [6], whose quantitative transition probabilities and reward functions capture the game-theoretic dynamics of our chosen models of wartime negotiation as follows:

State: Territory (0–100%), number of troops, military position (Slantchev only)
Actions: Accept/reject offer, attack (Powell), offer player $x\%$ of territory (Slantchev)
Transition: The probability distribution of the effects of actions on the state
Observation: We assume that the agent has complete information
Reward: Linear in amount of territory and number of troops

The PsychSim agent's theory of mind expands this POMDP model to include the human players' POMDP models as well. The agent does not know what POMDP would make the best model of the human player. For example, the game does not reveal the probability of battlefield outcomes to the player's side, so it is not clear what transition function would best capture the player's expectations. Furthermore, while the player most likely wants to increase both territory and surviving troops, the agent has no way of knowing whether that desire fits a linear function like its own, let alone what weights to use even if it does. On top of this uncertainty about the player's model of the game, the agent's theory of mind must also capture the players' theory of mind about itself.

Rather than trying to capture all of this uncertainty within the agent's theory of mind (e.g., by giving it a distribution over multiple POMDP models of the player), we instead implemented an agent that has no such uncertainty. In particular, it models the human player as following a POMDP that has complete information just as its own does. This POMDP uses a linear reward function, just like the agent's, but increasing in the human player's territory and troops instead. Thus, the agent uses a fixed model of the player as following the optimal policy computed by solving this POMDP. The agents can then use that policy as an expectation within its own POMDP to compute an optimal policy for itself.

3.1 Powell Agent Behavior: Attacking

Before an agent can decide whether it is optimal to accept an offer or not under the Powell condition, it must first examine its subsequent choice of simply rejecting an offer, or else rejecting the offer and attacking the player[1]. If either side

[1] Under Slantchev, there is no such choice, as a battle occurs every round.

attacks, the agent will win the war with probability $p_1(1-p_2)$ and will lose the war with probability $p_2(1-p_1)$. For our study, we use $p_1 = p_2 = 10\%$, so both probabilities come out to 9%. If the agent attacks, it therefore expects to have 100% of the territory with 9% probability, 0% with 9% probability, and its original amount of territory with 82% probability. Regardless of the outcome, the agent will incur a fixed loss of troops that, in its reward function, is weighted the same as 2% of territory. Thus, the difference between an attack (by either side) and no attack is the 18% chance of a military resolution (with an expected split of either 0% or 100% for the agent) and the 2% cost in troops. This possibility is valued differently by the agent depending on its starting territory:

Powell28: When not accepting an offer, the agent does not attack
Powell72: When not accepting an offer, the agent always attacks.

In other words, the possible military resolution is appealing in the Powell72 game when the agent starts with only 28% of the territory. Furthermore, the agent's Powell72 policy implies that the players *never* get a chance to attack, as they get that choice only if the agent rejects without attacking.

3.2 Powell Agent Behavior: Accepting Offers

Because the agent will not attack in the Powell28 condition, it weighs the human player's offer against its current 72% of the territory. Under the Powell72 condition, on the other hand, the agent weighs the human player's offer against both its starting 28% territory *and* the possibility of earning a more favorable split when it attacks. Because the agent does not distinguish between earning territory at the negotiation table vs. on the battlefield, it is willing to hold out for a higher offer than just the status quo in Powell72. In particular, the POMDP solution generates the following thresholds for the agent:

Powell28: Accepts offers $\geq 71\%$ (rejects otherwise)
Powell72: Accepts offers $\geq 35\%$ (attacks otherwise)

3.3 Slantchev Agent Behavior: Accepting Offers

The agent performs a similar computation under Slantchev, except that battlefield expectations are now contingent on military position, $k \in \{0, \dots, 10\}$. We set the probability, p, of a player winning a battle to be 30%, allowing the agent to compute its chances of winning or losing the war, which happens when $k = 0$ or 10, respectively. Solving the POMDP gives the agent a policy of holding out for higher offers as it gets closer to winning (i.e., k is low):

Slantchev28

$$\text{Accepts offers} \begin{cases} \geq 92\% \text{ if } k = 1 \\ \geq 85\% \text{ if } k = 2 \\ \geq 72\% \text{ if } k \in [3,7] \\ \geq 66\% \text{ if } k = 8 \\ \geq 51\% \text{ if } k = 9 \end{cases}$$

Slantchev72

$$\text{Accepts offers} \begin{cases} \geq 79\% \text{ if } k = 1 \\ \geq 64\% \text{ if } k = 2 \\ \geq 30\% \text{ if } k \in [3,7] \\ \geq 27\% \text{ if } k = 8 \\ \geq 20\% \text{ if } k = 9 \end{cases}$$

3.4 Slantchev Agent Behavior: Making Offers

To decide what offers it should make in the Slantchev game (it makes no offers under Powell), the agent first repeats the computation of Section 3.3 from the player's perspective. Again, it assumes that the players know that $p = 30\%$, and will have lower thresholds than the agent does because of this military disadvantage. It also does not change its beliefs about the player's thresholds, leading it to adopt a fixed policy of making offers as follows:

Slantchev28: Agent offers 10%, unless close to losing ($k > 7$), then offers 30%

Slantchev72: Agent offers 70%, unless close to winning, then offers 20% if $k = 2$, or 10% if $k = 1$.

4 Method

We recruited 240 participants, of an average age of 35, via Amazon Mechanical Turk. 51% of the participants are female and 49% are male. 65% of the participants are from the United States, 29% from India and 6% from other countries. 12% of the participants have some high school or high school diploma; 63% have some college or college degree and 25% have some graduate school or graduate degree. 13% of the participants use a computer for 1-4 hours a day, 43% use one 5-8 hours a day and 44% use one more than 8 hours a day.

Each participant is first assigned an anonymous ID and then reads the information sheet about the study. Then the participant fills out a Background Survey. Next the participant plays the negotiation game four times, each time with a different agent from one of the four conditions (the order is randomized). The game interface presents the participants with the troops and territory they own, as well as the number of rounds left and the history of previous offers and battle outcomes. There is no implication in the instructions that the participant would be playing against another human player. During the negotiation, the participant fills out an In-Game Survey. Following each negotiation game, the participant fills out an Opinion Survey. The study is designed to be completed within an hour, although the average duration was 32 minutes in our data.

We measured outcomes based on the following survey and game results:

Background Survey asks questions about the participant's age, gender, nationality, education, computer experience, Attitude Towards War [3], Social Orientation [19] and attitude towards Inappropriate Negotiation (SINS, from [17]).

Opinion Survey contains questions regarding the participant's goals during the game and modified questions from the Subjective Value Index (SVI) survey that measures perceptions of the negotiation outcome, process, relationship and the negotiator themselves [2].

In-Game Survey asks the participant to estimate the opponent's response after he/she makes an offer, e.g. accept the offer, reject it or attack.

Game Logs capture the actions the participant takes, the agent's actions and the world states, e.g. amount of troops and territory each side has.

5 Accuracy of the Agents' Theory of Mind

From our study's 240 participants, we have 238 games in the Powell72 condition and 239 games each in the Powell28, Slantchev72, and Slantchev28 conditions. We analyzed the participants' decisions made in all the games they played and categorized these decisions based on whether the agent would have made the same decision if it were in the participants' position. From Table 2, we see that the degree of conformity varies widely across the different conditions and decision types (empty cells are conditions where the action was inapplicable).

Table 2. Percentage of participant decisions that matched agent's theory of mind

Player Action	Powell 28	Powell 72	Slantchev 28	Slantchev 72
Making Offers	84%	59%	21%	34%
Response to Offer	—	—	80.3%	75.8%
— Reject Offer	—	—	100%	100%
— Accept Offer	—	—	0%	0%
Decision to Attack	43%	—	—	—
— Attack	100%	—	—	—
— Not Attack	0%	—	—	—

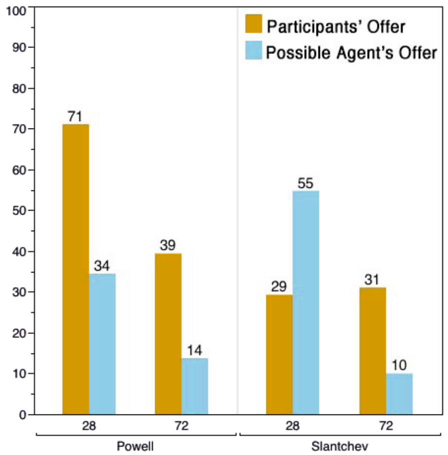
5.1 Participant Behavior: Attacking

From Section 3.1, we know that the agent will attack instead of reject in Powell72, preempting the participants' potential choice to do so themselves. In Powell28, the agent makes the opposite choice, and its theory of mind leads it to expect the participants to act as it did in the same situation (i.e., having only 28% of the territory) and attack instead of reject. The participants' matched that expectation and decided to attack only 43% of the time. One possible explanation for this deviation is that the participants may have placed a higher value on troops than the agent's reward function did. Alternatively, because the participants do not know the probabilities of collapse, their uncertainty may have led them to underestimate their expected gains from battle.

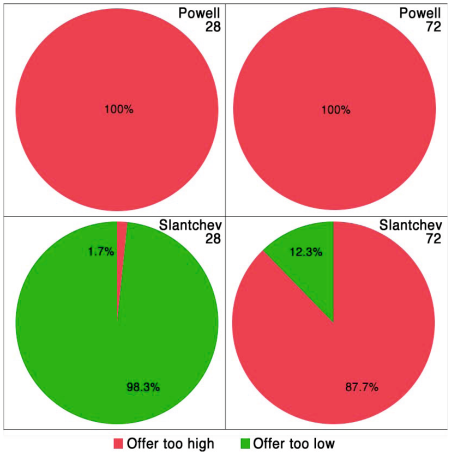
5.2 Participant Behavior: Making Offers

From Table 2, we see that the offers participants made did not always fall within the range of offers the agent would have made. Section 3.4 lays out the ranges of the agent's offers under Slantchev. In Powell games, the agent does not make offers, so we instead use the agent's offer-accepting policy (Section 3.2) as the range of offers the agent would have made. In particular, with a starting territory of 28% (for the agent), the agent will not offer more than 65%; with a starting territory of 72%, it will not offer more than 28%. One-way ANOVA tests show that, overall, the participants made offers consistent with the agent's policy more often under Powell than under Slantchev ($Mean_P = 77\%$, $Mean_S = 26\%$,

$p < .0001$). Within Powell games, as shown in Table 2, the participants' offers conformed more when starting with less territory ($p < .0001$). Within Slantchev games, however, the participants' offers conformed more when starting with more territory ($p < .0001$).



(a) Mean offer amounts.



(b) Percentage of participant offers that were higher/lower than the agent's.

Fig. 1. Offer amounts when participants deviated from the agent's policy

To see how the participants' offers quantitatively differed from the agent's model, we calculated the offers the agent would have made given the same game states the participants were in. As a crude approximation, we calculated the agent's offer as a uniform distribution within the offer range. In Powell, for example, we modeled an agent starting with 28% of territory as calculating its offers from a uniform distribution from 0 to 65%. Figure 1a shows the participants' and agents' offers when a deviation occurred. A total of 756 offers made by the participants were included in the analysis, including 227 in Powell28 condition, 157 in Powell72, 273 in Slantchev28 and 99 in Slantchev72. Paired-sample t-tests show significant differences between the participants' and the agent's offers ($p < .0001$) in all four conditions. In the Powell games, participants made higher offers than the agent would have. Section 5.1 gave multiple hypotheses for why the participants might be less likely to attack than the Powell agent, and the same causes would also lead them to be overly generous to minimize the risk of war. Alternatively, the participants may have a preference for achieving a resolution through negotiation than through war, so they may bias their offers to increase the chance of acceptance by the Powell agent. In the Slantchev games, the participants' deviating offers exceed that of the agent under Slantchev72, but it is *lower* under Slantchev28. The contrast of Slantchev28 is even starker in Figure 1b, where we see the vast majority of deviating offers are lower than the agent's. We do not have a good explanation for this outlying condition, although given the broad uncertainty facing the participants' they may be simply overestimating their probability of winning a war in this case.

5.3 Participant Behavior: Accepting Offers

The participants never receive offers to accept/reject under the Powell conditions, so we focus on only Slantchev games here. Table 2 shows that, when responding to an offer, the participants' responses match the agent's responses 80.3% of the time in Slantchev28 and 75.8% in Slantchev72. All of these deviations occur when the participants accepted an offer that the agent would not have. In other words, the agent would also have rejected any offer that the participants chose to do. This also implies that the agent never made offers that it would have accepted itself. Its theory of mind leads it to expect the participants to be more lenient given their military disadvantage, even though, in reality, the participants are initially unaware of that disadvantage.

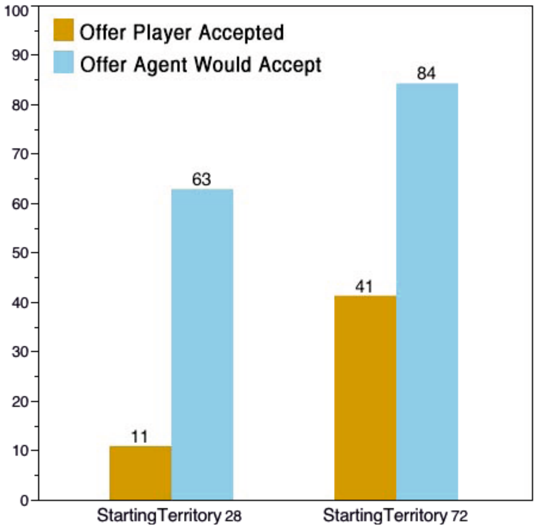


Fig. 2. Offers accepted by participants and agent in Slantchev games

Similarly, we compared the 145 offers participants accepted (91 from Slantchev28 and 54 from Slantchev72) and the ones the agents would have wanted in that same situation. We model the offers the agent would accept as a uniform distribution across each interval implied by the thresholds in the agent's policy in Section 3.3. Figure 2 shows that, regardless of the starting territory, the offers participants accepted are significantly lower than the ones the Slantchev agent would have wanted (paired-sample t-tests, $p < .0001$ for both pairs). This is consistent with the observation in Section 5.2 that the participants made higher offers, in that they were willing to achieve a negotiated settlement lower than what the agent would accept. It is also consistent with the hypothesis in Section 5.1 that the participants sought to avoid losing troops in the war more than the agent did, thus lowering their threshold for accepting an offer and bringing about agreement sooner.

6 Impact of Inaccuracies on Agent Performance

Although the participants did not always conform to the agent's model, it is not clear how much the inaccuracies in the agent's theory of mind affected its performance. There are two obvious measures of agent performance: territory owned and troops lost at the end of the game. We use territory as the primary measure because it is a zero-sum game outcome that identifies a "winner" between the agent and participant, whereas both players lose troops in every battle. Given that the agent attempted to maximize territory gains based on its theory of mind about the participants, we might expect its optimal policy to do worse when the participants deviate from those expectations. Indeed, when we examine the agent's performance in the Slantchev conditions, we see that it is negatively correlated with the percentage of participants' actions that deviated from the agent's policy ($r = -.463, p < .0001$). However, we did not observe such a correlation under the Powell conditions ($r = .080, p = .0813$) So while the Slantchev agent earned more territory the more accurate its model of the participant, we cannot make the same overall conclusion under the Powell condition.

6.1 Impact of Deviations in Attacking

Section 5.1 showed that deviations in the participants' attacking decisions occur in only the Powell28 condition, when they reject without attacking. The agent expects them to attack, because the potential gain from winning the war outweighs the 28% territory lost by losing the war. The data confirm this expectation. One-way ANOVA tests show that the agent ended up with more territory when it successfully avoided the war ($Mean_{War} = 57.7\%$, $Mean_{\neg War} = 76.5\%$, $N_{War} = 190$, $N_{\neg War} = 49$, $p = .0026$). Thus, the agent's belief that attacking is the optimal policy for the participants is borne out by the fact that they did worse when deviating, e.g. the agent successfully avoided the war. However, although this divergence between agent expectations and participant behavior works to the agent's benefit, it could do even *better* by holding out for higher offers if it knew that the participant might not attack in retaliation.

6.2 Impact of Deviations in Making Offers

We also analyzed the relationship between the percentage of the participants' "correct" offers and the territorial split at the end of the game. We broke down this analysis based on how the games ended, e.g. agent accepting an offer, winning/losing the war. The rationale is that making a "wrong" offer (e.g. overly high) may have no immediate impact on the game outcome when it gets rejected, but a "wrong" offer that gets accepted could possibly lead to a much worse outcome for the offeror. Indeed, within the 375 games where the offers were accepted by the agent, there is a negative correlation between the percentage of "correct" offers and territory the participants ended up with ($r = -.3234$, $p < .0001$). This also means that the more often the participants' offers conformed to the agent's policy, the more territory the agent ended up with, as we might expect from a more accurate theory of mind. On the other hand, in the 580 games that did *not* end with the participant making an acceptable offer to the agent (i.e., winning/losing the war, reaching the end of the game, or the participants' accepting an offer from the agent), there is a *positive* correlation between the participants' degree of conformity to the agent's model and their territorial outcome ($r = .3423$, $p < .0001$). In these games, the participants' offers are lower than what the agent wants or expects. Without an agreement that exceeds its threshold, the agent relies on a battlefield outcome or the status quo, which is not likely to favor it as much. This suggests that the agent would benefit by lowering its threshold when dealing with such less generous participants, as the current fixed threshold causes it to miss out on such opportunities.

6.3 Impact of Deviations in Accepting Offers

In the Slantchev games, all of the offers accepted by the participants deviated from the agent's policy, as seen in Section 3.3. To evaluate the impact of this deviation on performance, we compared the 145 games ending with participants accepting the agent's offer and the 333 games ending some other way, e.g. making an offer that the agent accepts or else somehow prolonging the war through all 15 rounds (with both sides keeping their original territory). One-way ANOVA tests show no significant impact on how much territory the participants ended up with ($Mean_{AcceptOffer} = 22.2\%$, $Mean_{Alternative} = 22.5\%$, $p = .9084$). This suggests that even though the participants accepted offers that were too low in the agent's estimation, this deviation had little impact on the game outcome.

7 Discussion

Despite its strong assumptions, the agent's theory of mind still allowed it to perform well (although not optimally), because most of the participants' deviations from expectations did not hurt the agent's performance. By examining the impact of the deviations on the agent's performance, we can prioritize the areas of its theory of mind that we need to improve. For example, by assuming that the participants have complete information, the agent misjudged the offers the participants would make and accept. The agent could instead bias its expectations of the participants to be more conservative given their lack of information. More generally, an agent operating in other domains with similarly uncertain human participants can also benefit from a theory of mind that perturbs an expectation of optimality with a degree of risk aversion.

Even when properly accounting for the participants' uncertainty, we also saw potential deviations due to unanticipated values. For example, the participants' SVI survey responses indicate that they placed a value on reaching an agreement with their negotiation partner. Our agent did not incorporate such a value, nor did it expect the participants to do so. By modifying our agent's theory of mind to include such a value on agreement, it could generate expectations that better account for the participants' willingness to give up territory in exchange for a quicker settlement. We could then repeat the methodology of this paper to evaluate the degree to which the participants' conformed to this modified theory of mind and the potential impact of this change on agent performance.

We can use this paper's methodology to evaluate completely different methods for theory of mind as well. For example, the original game-theoretic analyses of wartime negotiation prescribe that the offering side should start with a low amount and steadily increase it, to screen for the other side's acceptance threshold with minimal over-offering [13,18]. Even without implementing this model in our own agent, we can still measure the degree to which it matches our participants' behavior. Participants followed this strategy in 317 of our games, and violated it by either repeating or decreasing their offer in 286 games[2]. Overall, there is no significant impact of following the equilibrium strategy on the agent's performance. However, within the games ending with the agent accepting the participant's offer, the ones where the participants played consistently with the equilibrium strategy resulted in more territory for the agent ($Mean_{eq} = 34.6\%$, $Mean_{\neg eq} = 46.9\%$, $p < .0001$). Thus, while screening may be a best response in the theoretical setting, the participants' uncertainty leads them to over-offering in practice, to the agents' benefit.

It is clearly insufficient to evaluate only the accuracy of an IVA's theory of mind with respect to actual human behavior. Some classes of inaccuracy may not have any negative impact on the agent's performance, in which case it is unnecessary to enrich the model to remedy that inaccuracy. Furthermore, in a domain with multiple sources of uncertainty, even one as simplified as our wartime negotiation models, expanding theory of mind without a good model of that uncertainty can be even detrimental to agent performance. Theory of mind exists in service of the overall social interaction, and our analysis demonstrates that we should seek improvements to the modeling of others only when motivated by the subsequent improvements in that interaction.

Acknowledgments. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government. No official endorsement should be inferred.

[2] The remaining 352 games ended before the participant made a second offer.

References

1. Baker, C.L., Tenenbaum, J.B., Saxe, R.R.: Goal inference as inverse planning. In: CogSci (2007)
2. Curhan, J.R., Elfenbein, H.A., Xu, H.: What do people value when they negotiate? Mapping the domain of subjective value in negotiation. Journal of Personality and Social Psychology 91, 493–512 (2006)
3. Dupuis, E.C., Cohn, E.S.: A new scale to measure war attitudes: Construction and predictors. Journal of Psychological Arts and Sciences 3(1), 6–15 (2006)
4. Goodie, A.S., Doshi, P., Young, D.L.: Levels of theory-of-mind reasoning in competitive games. Journal of Behavioral Decision Making 25(1), 95–108 (2012)
5. Hoogendoorn, M., Soumokil, J.: Evaluation of virtual agents utilizing theory of mind in a real time action game. In: AAMAS, pp. 59–66 (2010)
6. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial Intelligence 101, 99–134 (1998)
7. Kahneman, D.: Reference points, anchors, norms, and mixed feelings. Organizational Behavior and Human Decision Processes 51(2), 296–312 (1992)
8. Kim, J.M., Hill Jr., R.W., Durlach, P.J., Lane, H.C., Forbell, E., Core, M., Marsella, S., Pynadath, D., Hart, J.: Bilat: A game-based environment for practicing negotiation in a cultural context. International Journal of Artificial Intelligence in Education 19(3), 289–308 (2009)
9. Klatt, J., Marsella, S., Krämer, N.C.: Negotiations in the context of AIDS prevention: An agent-based model using theory of mind. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 209–215. Springer, Heidelberg (2011)
10. Marsella, S.C., Pynadath, D.V., Read, S.J.: PsychSim: Agent-based modeling of social interactions and influence. In: ICCM, pp. 243–248 (2004)
11. Miller, L.C., Marsella, S., Dey, T., Appleby, P.R., Christensen, J.L., Klatt, J., Read, S.J.: Socially optimized learning in virtual environments (SOLVE). In: André, E. (ed.) ICIDS 2011. LNCS, vol. 7069, pp. 182–192. Springer, Heidelberg (2011)
12. Neale, M.A., Bazerman, M.H.: Cognition and rationality in negotiation. Free Press (1991)
13. Powell, R.: Bargaining and learning while fighting. American Journal of Political Science 48(2), 344–361 (2004)
14. Pynadath, D.V., Marsella, S.C.: PsychSim: Modeling theory of mind with decision-theoretic agents. In: IJCAI, pp. 1181–1186 (2005)
15. Pynadath, D.V., Marsella, S.C.: Minimal mental models. In: AAAI, pp. 1038–1046 (2007)
16. Pynadath, D.V., Marsella, S.C., Wang, N.: Computational models of human behavior in wartime negotiations. In: CogSci (to appear, 2013)
17. Robinson, R.J., Lewicki, R.J., Donahue, E.M.: Extending and testing a five factor model of ethical and unethical bargaining tactics: Introducing the sins scale. Journal of Organizational Behavior 21, 649–664 (2000)
18. Slantchev, B.L.: The principle of convergence in wartime negotiations. American Political Science Review 97, 621–632 (2003)
19. Van Lange, P.A.M., De Bruin, E.M.N., Otten, W., Joireman, J.A.: Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. Journal of Personality and Social Psychology 74(4), 733–746 (1997)
20. Whiten, A. (ed.): Natural Theories of Mind. Basil Blackwell, Oxford (1991)

Censys: A Model for Distributed Embodied Cognition

Tiago Ribeiro, Marco Vala, and Ana Paiva

INESC-ID & Instituto Superior Técnico, Technical University of Lisbon
{tiago.ribeiro,marco.vala,ana.paiva}@inesc-id.pt

Abstract. The role of the body in the generation of behavior is a topic that has sparked the attention of many fields from philosophy to science and more recently robotics. We address the question of how an embodied agent should be modeled in order to change the traditional dualist approach of creating embodied agents. By looking at behavior generation as a shared process between mind and body, we are able to create modules that generate and manage behavior, which are neither part of the body nor the mind, thus allowing a more flexible and natural control. A case study is presented to demonstrate and discuss our model.

Keywords: Embodied Agents, Embodied Cognition, Artificial Intelligence, Robotics.

1 Introduction

Embodiment as something of or related to the human body is an important field of research. Our physical bodies define how we stand in space and time, and our awareness is deeply influenced by the fact that we have a body. Over the years, philosophers, psychologists, cognitive scientists, and more recently computer scientists have looked at embodiment from different perspectives.

In Computer Science, there is the concept of embodied agent, which is a software agent that interacts with the surrounding environment through a body. Embodied agents can have actual physical bodies, like Robots [1], or they can have a graphical representation of their body, like Embodied Conversational Agents [2][3]. In both cases, a form of embodiment (physical or virtual) is a necessary condition to interact with the environment and with human beings [4].

One particular topic related with embodiment is the relationship between mind and body. The mind-body problem was famously addressed by Descartes in the 17th century when he proposed his dualist perspective. Cartesian Dualism assumes that the mental phenomena are essentially non-physical, and that mind and body are two separate things.

The traditional computational models to create embodied agents follow a dualist perspective. There is a clear separation between the "mind" of the agent and the "body" of the agent. The body is an interface to the environment through a set of sensors and effectors. The mind receives sensory information from the

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 58–67, 2013.

body, analyzes that information and activates the effectors. There is a continuous sense-reason-act loop in which the mind has full control over the body.

However, such approach has some implications. The mind, as a centralized decision-making system, has to cope with different levels of control at the same time, ranging from lower-level control of sensors and effectors to higher-level cognitive tasks that involve reasoning and deciding what the virtual agent or the robot should do next. Moreover, the level of abstraction provided by both sensors and effectors, and their ability to map sensory input into symbolic representations or turn symbolic representations into effector output, has a direct impact in what the agent can do. As a consequence, the mind usually ends up tightly coupled to a particular form of embodiment.

Human beings, on the other hand, have intermediate layers of control at different levels. Our bodies have regulation mechanisms that perform subconscious tasks in parallel with our higher-level cognitive tasks. Damásio [5] presents recent findings in neuroscience of how our bodies are important in shaping the conscious mind, and their role in key processes like the emotional phenomena. Pfeifer [6] also points out that, despite the breadth of the concept, whatever we view as intelligent is always compliant with the physical and social rules of the environment, and exploits these rules to create diverse behavior. Since our bodies define how we interact with the environment, we cannot dissociate intelligence from our body as a whole.

Therefore, we need a computational model that looks at mind and body as a continuum. In the Society of Mind [7], Minsky looks at the mind as a collection of cognitive processes each specialized to perform some type of function. A cognitive process is represented by a component and the internal composition of these components creates a network of complex behavior.

This paper looks at embodiment following the same approach. We look at the body as a sum of components that perform specialized functions. Mind and body can share the same space, because we don't look at them as separate processes. Instead, both the notions of mind and body emerge from the components that support them.

In the next section we will look at related work. Then we present our model and a case study that discusses our approach. Finally, we draw some conclusions and outline future work.

2 Related Work

Researchers working in the field of virtual and robotic agents have been exploring richer models for behavior generation in autonomous agents. Recently, there have been developments towards new frameworks and tools to create agents capable of generating and exhibiting complex multimodal behavior. A popular framework that defines a pipeline for abstract behavior generation is the SAIBA framework [8], illustrated in Figure 1.

Our architecture follows on [9], which uses a pipeline of components that are reusable and migrate across different forms of embodiment. These components

Fig. 1. The SAIBA framework [8]

were used to create mixed scenarios where agents can migrate between virtual and robotic bodies.

Moreover, in order to deal with the other side of the loop, Scherer et al. propose a Perception Markup Language (PML) that should work just like BML [10]. As agents perceive the external world through their bodies, it enables embodiments to create an abstraction of perceptual data in order to bring it up to the cognitive level.

The interaction between perceptual data and BML has also been explored. We have integrated perceptions into BML and had two different robots interact with each other while running on the same system [11]. The interactive behavior is abstract enough to drive two completely different embodiments.

Other approaches to build embodied agents use physiological models to create behavior in autonomous agents. For example, Cañamero uses a multi-agent approach where each physiological function is modeled using an agent [12]. The work models hormones which are the foundation of motivational and emotional processes that guide behavior selection in Robots.

From a point of view of engineering, the component-based approach is very common in Robotics. ROS - Robot Operating System is a popular middleware developed by Willow Garage [13] that provides a common communication layer to enable different kinds of sensors, motors and other components to send data between each other. ROS is module-based, meaning that a ROS-based robot actually runs several different modules, being each one of them responsible for controlling one or several components of the robot. The main advantage of this is that all these modules can be shared and reused throughout the community.

3 The Censys Model

The model we propose follows on the concepts we previously introduced in the first section, and is inspired in the component decomposition proposed by Minsky.

Censys is modeled as a distributed network of Modules, which can have any non-zero number of connectors. A Module is conceptualized as being a *black-box* which can have or not an internal state, and can react to data received on its sensors, through the use of its effectors. As an abstraction, a Module can actually be seen as a sub-agent that composes both the mind and the mind-body interface.

The Censys model does not enforce any specific typology for the network. It can therefore be built just like a traditional agent, as show in Figure 2. Sensors are illustrated as triangles pointed at the module, and actuators as triangles pointed away from the module.

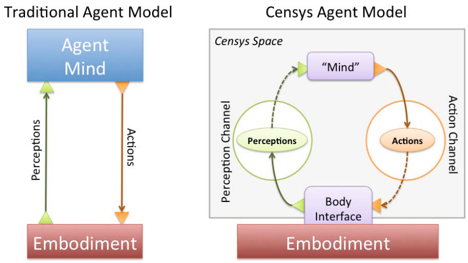


Fig. 2. How the Censys model fits into the Traditional model

What is generally viewed as the Agent Mind, providing deliberative, reactive or dialogue behavior, is now just a module that fits into the architecture. The Body is decomposed into a one or more Censys modules that serve as interfaces to the Embodiment.

A Module can use four types of connectors to sense and act:

PerceptionSensor subscribes to and receives perceptions P_T of a type T;
PerceptionEffector generates perceptions P_T of a type T;
ActionSensor subscribes to and receives actions A_T of a type T;
ActionEffector generates actions A_T of a type T;

The ActionEffectors and PerceptionEffectors generate actions and perceptions and place them in the perception and action channel respectively. The channels have the information of which ActionSensors and PerceptionSensors are subscribed to which type of actions and perceptions, so it can then transmit those to all of the subscribed sensors.

An example of how these modules can be composed into a Censys agent with a Body, Decision-Making (DM) module, and three other Modules is shown in Figure 3. This agent is merely an out of context example and as such its Modules do not have any specific meaning.

The three new modules in the example agent are able to receive some kind of perceptions and actions, process them, and generate other kinds of perceptions and actions.

Taking as example Module A, it subscribes to actions of type A_3 and converts them into actions of type A_1. The purpose of this module is thus to act as a high-to-low-level converter, decomposing high-level actions from the DM into lower-level actions that are more appropriate for the Body to manage and use without having to know how to interpret the high-level information produced by the DM.

Another example, Module C, has a dual purpose. One purpose is to act as a perception converter, by receiving lower level P_1 perceptions and turning them into higher level P_3 symbolic perceptions that are more appropriate for the DM. At the same time, it can also generate low-level reactions by generating A_2 actions depending on the Module's internal state and the data collected from the perceived P_1s.

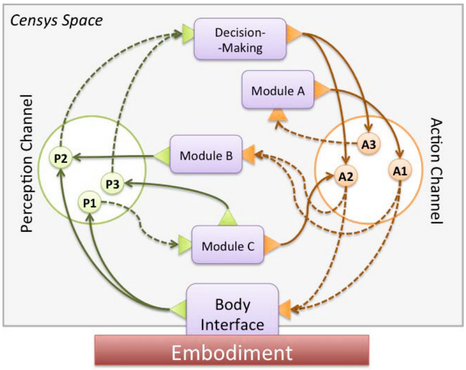


Fig. 3. An example (no context) Censys agent

The main advantage of using this architecture as that the DM does not explicitly need to know how to communicate with the Body and vice-versa.

In our example agent, we can see that the DM can only receive P_2 and P_3 perceptions and generate A_2 and A_3 actions. The body, however, can only generate P_1 and P_2 perceptions, and receive A_1 and A_2 actions, so the DM and Body used can natively only communicate through P_2 perceptions and A_2 actions. This means that if we removed the three Modules (thus having a traditional agent instead of a Censys agent), we would need to adapt both the DM and Body to be able to deal with all types of perceptions and actions.

In a Censys agent we just use Modules that can handle these types of perceptions and actions, and still use the DM and Body as they are. A direct advantage of this is that it makes it much easier to swap the traditional *Agent Mind* or the *Embodiment* with other ones that may not have support for all the used actions and perceptions, and still function together.

4 Case Study - A Component-Based Robot

In figure 4 we present a more complex case study in which a Censys agent is used to control a robot, based on the SAIBA framework. The Intention and Behavior Planning is done within the Decision-Making and Dialogue Manager modules (DMs), which generate only A_{BML} actions containing Behavior Markup Language (BML) blocks[8]. The DMs also receives only P_{PML} perceptions containing a Perception Markup Language (PML) blocks, which is a high-level representation of perceptual data [10].

This Embodiment is actually connected to the agent through three modules: the Speech, Audio and Body Interfaces. This case study thus also serves as an example of how a body is not necessarily one entity, but a coupling of several entities. Moreover, if we wanted to use a different Robot we could just switch its Body Interface and still keep all our behavior related to speech and audio.

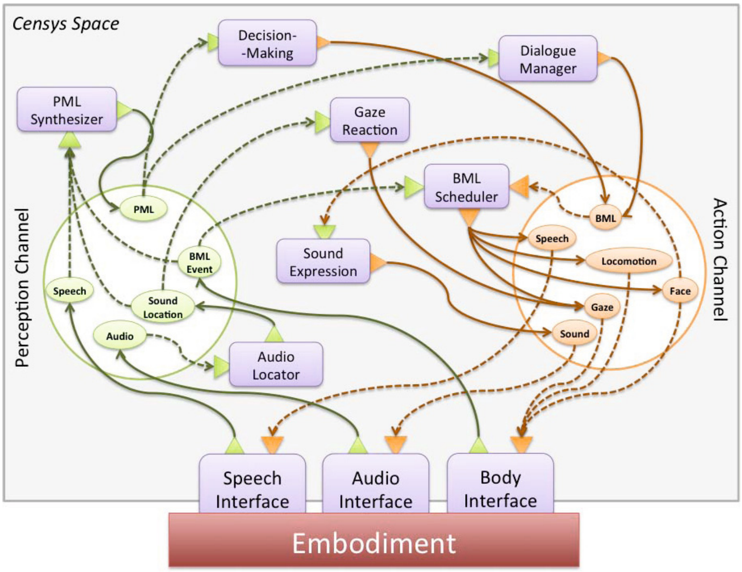


Fig. 4. Diagram of the Censys agent of the Cognitive/Reactive Robot case study

The **Body Interface** receives A_{Face}, A_{Gaze} and $A_{Locomotion}$ actions, and generates $P_{BMLEvent}$ perceptions. These perceptions contain feedback about the executing actions (when they physically started/finished/failed, etc.).

The **Audio Interface** receives A_{Sound} actions containing an audio signal to be output, and generates P_{Audio} containing audio signals whenever a sound is captured from the environment.

The **Speech Interface** acts as an interface to both a Text-to-Speech (TTS) engine, and to a Speech Recognition engine. In this case we assume that the Speech Recognition is already being fed with an audio signal, as such it produces P_{Speech} perceptions containing detected speech. This interface also receives A_{Speech} actions containing information that the TTS engine uses to generate speech.

We will now analyze each of the different modules that compose the behavior of this agent:

BML Scheduler deals with decomposing high-level behaviors, and scheduling and running the separate actions that compose such behavior. It subscribes to A_{BML} actions that correspond to the high-level behavior (a BML block) and generates actions of type A_{Face}, A_{Gaze}, $A_{Locomotion}$ and A_{Speech};

Sound Expression provides expressive redundancy to some of the agent's expressive behavior. In this case, every time a module produces an A_{Face} action, this module will generate an A_{Sound} action that contains an audio signal corresponding to the A_{Face} action;

Audio Locator serves as a converter module that takes as input P_{Audio} perceptions containing audio signals, and processes them in order to infer localization information. If the sound can be localized it generates a $P_{SoundLocation}$ perception with that information;

Gaze Reaction performs gaze reactions to sound events. Whenever it receives a $P_{SoundLocation}$ perception that is loud enough, it will generate an A_{Gaze} action that tells the body to gaze towards the direction of the sound;

PML Synthesizer acts as a low-to-high level converter, by receiving individual perceptions like $P_{BMLEvent}$, $P_{SoundLocation}$ or P_{Speech} and transforming them into high-level PML blocks.

4.1 Execution Example

Given the description of the case study, we now provide a description of how that scenario could actually run.

Let's start by assuming that we are using a NAO robot[1] for the embodiment, and that the DMs are very simple and currently only pursue the goal of travelling to a certain physical location that is located a couple of meters in front of the robot's current location. The Decision-Making module processes that goal and defines a simple plan consisting merely of walking forward.

That plan is transformed into a BML block containing a locomotion action and let's say, a speech action for *"I will be there in a minute!"*. The BML Scheduler receives this action and its scheduling of the two behaviors generates first an A_{Speech} action, and then an $A_{Locomotion}$ action. The A_{Speech} action is received by the Speech Interface, which unpacks it and sends it to the TTS, making the robot say *"I will be there in a minute!"*. The $A_{Locomotion}$ action is received by the Body Interface, which triggers the robot to start walking.

In the mean while, someone goes by NAO and speaks to him. The Audio Interface detects this sound and generates a P_{Audio} perception, which is sensed by the Audio Locator module which in turns calculates the offset angle at which the sound was detected. It then generates a $P_{SoundLocation}$ perception. This perception is received both by the PML Synthesizer which generates a P_{PML} perception based on its data, and also by the Gaze Reaction module. This one generates an A_{Gaze} action that tells the body to gaze at the angle of the detected sound. That A_{Gaze} action is then also received by the Body Interface, thus making NAO look at the direction of the person who spoke, while continuing to walk. The DMs also receive the P_{PML} perception containing the information about a sound perceived at certain angle, but as long as it does not interfere with its current state and goals, this perception does not trigger anything at this level.

That does not imply that the robot is unable to react to anything else. When an intense sound is located, gazing at it can have several benefits. We think of a functional one - being able to use vision recognition to analyze what has generated the sound; and an expressive one - if it was a person who made the

[1] `http://www.aldebaran-robotics.com`

sound, then having the robot gaze towards that direction helps to transmit a more sentient impression of the robot.

Besides these two benefits, this behavior is not encoded neither in the DMs nor in the embodiment, so that means it could be reused even if we switch to another robot.

4.2 Discussion

While the diagram presented in Figure 4 may seem complex, it actually portrays a very simple scenario. What do we gain from having such cross-connections and interrelationships?

We find this kind of model to be especially appropriate for modeling high-level autonomous behavior in robots, because we can create some behaviors that are body-independent (like gaze-reacting to a sound) and use them with different embodiments. Of course, another question may arise that is HOW does the robot implement that gaze. That is a problem that we consider to be body-dependent, as a robot that has a head may be able to gaze while walking, whereas a robot built on a two-wheeled self-balancing base (like the Segway technology [2]) will have to stop moving in order to turn and face another direction.

Another vantage point we find is that each module offers its own independent control. An example of this is the Sound Expression module. The Dialogue Manager does not even need to be aware of this module while it supports the expression of the robot. If we know that our robot has limited expression capabilities, we can just include this module without having to change anything else.

The BML Scheduler is also a complex control module as it manages the scheduling and composition of behaviors that the DMs decided to execute. However, the DMs do not need to know what the Body is currently doing - resource management is distributed along the Mind-Body space.

There is still another situation related to resource management, that is when for example both the BML Scheduler and the Gaze Reaction modules produce an A_{Gaze} action. In this case, and having no other module to play that role, the resource management is expected to be done at the Embodiment layer, so we cannot predict or define in our model what will be the resulting action.

Summing up this case study, by using a Censys agent we steer towards the ability to reuse behaviors that manage proper and natural interaction, even when working with different embodiments (virtual or robotic). As some of the agent's behavior does not need to be re-programmed, the development of new behaviors and contexts for robot usage should therefore become more accessible.

5 Conclusion

There are several aspects that we intend to approach with our model. First of all, we were deeply inspired by the fact that most traditional architectures

[2] `http://www.segway.com`

overlook the role of the body in the cognition process. This leads to placing a high computational load on the agent's mind and expecting it to be able to cope with both low and high-level processes.

That is the issue that generally leads to dependence between body and mind using the traditional approach. By breaking that hard link between body and mind, we are able to transfer some common behaviors into a more abstract functional space between the cognitive mind and the body, thus making it feasible to share behaviors and part of the cognition process even if we use different bodies.

One important thing on embodiment switching is when we switch between virtual and robotic bodies. Virtual embodiments are considered to be perfect and immediate, meaning that they have direct access to real information about their body and the world around. Robots, however, have imperfect sensors and effectors, meaning that there can be noise in the information and also measurement errors or other deviations caused by gravity, inertia, friction, etc. By including a space where we can create filters for sensors and actuators, we can relieve the higher-level processes from handling all those issues.

In the same sense, there are some processes that may require a continuous feedback loop between a specific controller and the embodiment. Again, such a controller can deal with feedback and adjust the behavior in real-time without having to interrupt the higher-level processes.

The bottom line is that Censys creates a space where we can define internal processes that run in parallel with the main control loop of the agent. This distributed control is the base to create agents that are able to display a more natural behavior, for example by enriching their primary behavior with involuntary movements or by reacting faster when something happens.

Therefore, embodied agents are able to rely on their bodies as much as they rely on the artificial minds that reason and decide for them. Mind and body work in parallel to generate behavior and continuously adapt to each other. Like humans do.

5.1 Future Work

The next step will be to implement some concrete scenarios using Censys. We want to compare the development of a Censys agent against the traditional approach. This comparison should yield results both about how easy and fast it is to develop a Censys agent, but also to what degree the Censys modules are actually interchangeable within different embodiments. It should also guide us towards defining the requirements on both the model and its modules in order for that interchangeability to remain valid.

Censys also enables the creation of an internal body model where we can explore concepts like physiological space, interoception, and proprioception. Usually these mechanisms are part of subconscious processes which are not that important for the actual behavior displayed by our agents. However, they might be important in the decision process that lead to those behaviors in the first place and we will be able to experiment with that.

There is also an open issue related to resource management (RM). Having distributed control implies having some concurrency and how to deal with shared resources, like the effectors. In this paper we have delegated that process to the actual embodiment. However, we have also hinted in our case study how the BML Scheduler also performs some RM. That make us to wonder if it may be possible to actually include better RM mechanisms in our model, and to what level can they perform to be shared amongst different types of embodiment.

Acknowledgments. This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013 and by the EU 7th Framework Program project EMOTE (FP7/2007-2013) under grant agreements no. 215554 and 317923.

References

1. Niewiadomski, R., Obaid, M., Bevacqua, E., Looser, J., Le, Q.A., Pelachaud, C.: Cross-media agent platform, vol. 1, pp. 11–20 (2011)
2. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: SIGCHI 1999, CHI 1999, pp. 520–527. ACM, New York (1999)
3. Poggi, I., Pelachaud, C., De Rosis, F., Carofiglio, V., De Carolis, B.: GRETA. A Believable Embodied Conversational Agent, vol. 27, pp. 1–23. Kluwer Academic Publishers (2005)
4. Dautenhahn, K.: Embodiment and Interaction in Socially Intelligent Life-Like Agents. In: Nehaniv, C.L. (ed.) CMAA 1998. LNCS (LNAI), vol. 1562, pp. 102–141. Springer, Heidelberg (1999)
5. Damasio, A.: Self Comes to Mind: Constructing the Conscious Brain. Random House (2011)
6. Pfeifer, R., Bongard, J., Grand, S.: How the Body Shapes the Way We Think: A New View of Intelligence. Bradford Books, Mit Press (2007)
7. Minsky, M.: Society of Mind. A Touchstone Book, Simon & Schuster (1988)
8. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
9. Kriegel, M., Aylett, R., Cuba, P., Vala, M., Paiva, A.: Robots meet IVAs: A mind-body interface for migrating artificial intelligent agents. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 282–295. Springer, Heidelberg (2011)
10. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A(S.), Morency, L.-P.: Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 455–463. Springer, Heidelberg (2012)
11. Ribeiro, T., Vala, M., Paiva, A.: Thalamus: Closing the mind-body loop in interactive embodied characters. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 189–195. Springer, Heidelberg (2012)
12. Canamero, D.: A hormonal model of emotions for behavior control. VUB AILab Memo 2006 (1997)
13. Quigley, M., Gerkey, B.: ROS: an open-source Robot Operating System (2009)

Automated Promotion of Technology Acceptance by Clinicians Using Relational Agents

Timothy Bickmore[1], Laura Vardoulakis[1], Brian Jack[2], and Michael Paasche-Orlow[2]

[1] College of Computer and Information Science, Northeastern University, Boston, MA, USA
{bickmore,laurap}@ccs.neu.edu

[2] Boston University School of Medicine, Boston Medical Center, Boston, MA, USA
{brian.jack,michael.paasche-orlow}@bmc.org

Abstract. Professionals are often resistant to the introduction of technology and can feel threatened if they perceive the technology as replacing some aspect of their jobs. We anticipated some of these problems in the process of introducing a bedside patient education system to a hospital, especially given that the system presents itself as a "virtual discharge nurse" in which an animated nurse agent interacts with patients using simulated face-to-face conversation. To increase acceptance by nursing staff we created a version of the character designed to build trust and rapport through a personalized conversation with them. In a randomized trial, we compared responses after 15 minute in-service briefings on the technology versus responses to the same briefings plus a personalized conversation with the agent. We found that the nurses who participated in briefings that included the personalized conversation had significantly greater acceptance of and lower feelings of being threatened by the agent.

Keywords: Relational agent, embodied conversational agent, technology acceptance, medical informatics.

1 Introduction

Despite billions in recent US federal spending targeted at clinical information systems, significant barriers to adoption of health information technology (IT) remain [1]. Some of the most significant barriers center on acceptance of the technologies by the clinicians who use them. Although financial incentives may motivate proprietors, it remains to be seen how effective these incentives will be at changing attitudes and actual use behavior by clinicians on the front lines of care who do not stand to personally profit from the incentives, especially if they feel threatened by the technology being introduced.

A significant amount of research has been conducted over the last two decades on the factors that lead to acceptance of a new technology. Attitudes towards a new technology are important even in environments in which use is mandatory, and there are many documented cases of underuse, workarounds, and abandonment of health IT [2]. Knowledge of the factors that affect technology acceptance could enable organizations to manipulate these factors to increase acceptance and use of health IT.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 68–78, 2013.

The most widely used framework for studying technology acceptance is the Technology Acceptance Model (TAM), which posits that an individual's actual use of a technology can be predicted from their stated intention to use the technology, their attitude towards the technology (overall satisfaction), and perceptions of the technology's ease of use and usefulness [3]. Figure 1 shows the originally hypothesized relationships among these factors. The TAM has received empirical support across a number of industries and technologies, typically accounting for 30% to 40% of IT acceptance.

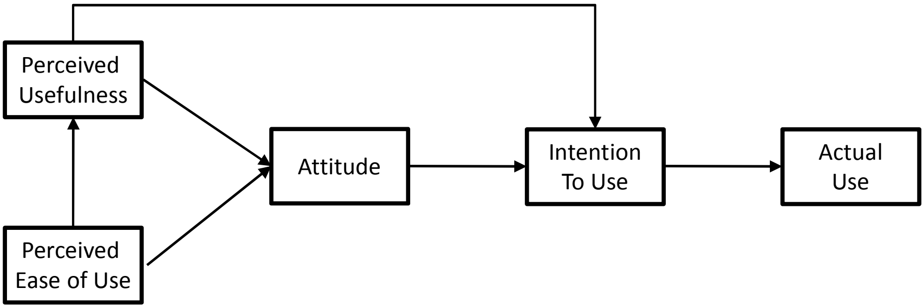


Fig. 1. Technology Acceptance Model

Very few studies of the TAM have focused on approaches for promoting acceptance using external influences (such as social influences or persuasive messages), partly due to the wide variety of messages that could be employed, the differential impact these messages have on different user groups, and the lack of temporal models explaining the time course of influences [4]. One framework that has been used widely in studies of attitude change and persuasion is the elaboration likelihood model, in which attitude change is hypothesized to be effected through either "central route" cues (argument quality, appealing to reason) or "peripheral route" cues (meta-information about a message or source, such as the recipient's trust in the source)[5]. This model has been used in studies of persuasion to improve technology acceptance, demonstrating that central route messages can significantly influence perceived usefulness and peripheral route cues can significantly influence both perceived usefulness and attitude towards the technology [4].

In the current study we sought to introduce a new hospital bedside patient education system into the general medical service at Boston Medical Center as part of an intervention to reduce unnecessary re-hospitalizations (Figure 2) [6, 7]. Boston Medical Center is a 547 bed safety net hospital that serves an urban, 84% minority, traditionally underserved population. Approximately 58% of the patients hospitalized at Boston Medical Center have inadequate health literacy. Although the new system was designed to offload a routine task from nurses and improve patient care by augmenting patient education regarding post-hospital discharge self-care, we were concerned that the nursing staff may feel threatened by the technology, since it could be seen as a "virtual discharge nurse" intended to replace certain traditional nursing duties. In order to increase acceptance of the technology we designed a version of the patient education system especially tailored for the nursing staff that not only explained how the system worked (central route messages), but attempted to directly establish trust and rapport with the nurses and decrease any feelings of being threatened by the technology (peripheral route cues).

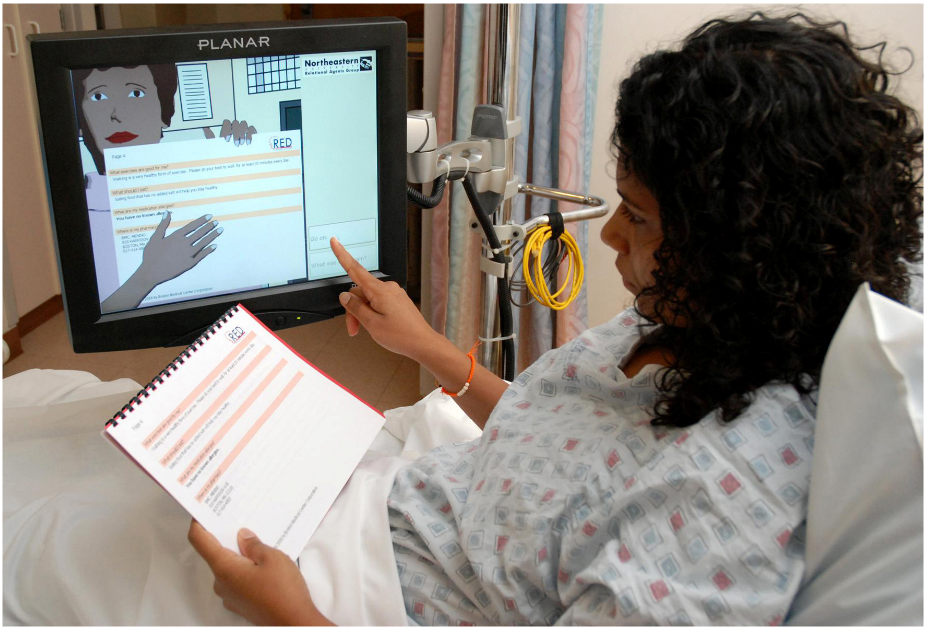


Fig. 2. Patient Interacting with the Relational Agent Patient Education System

Intelligent virtual agents that are explicitly designed to build rapport, trust, therapeutic alliance and other forms of working relationships with people have been referred to as “relational agents”, and incorporate many of the same behaviors that people use for this purpose, including social chat, expressions of caring and empathy, agreement on values and beliefs, and nonverbal behaviors such as smiling, facial and gestural animation, and close proximity [8]. Relational agents have been used with diverse patient populations for a range of health education and behavior change interventions [6, 9], and are also used in the hospital bedside patient education system (Figure 2). To date, however, relational agents have not been used with professionals for the purpose of increasing technology acceptance in the workplace.

The TAM is appropriate for assessing nurse acceptance even though nurses are not the end users of the system. When the patient education system is deployed in a hospital for routine care, then the nurses on the floor will ultimately determine whether a particular patient gets to use the system or not. We feel this is true in practice, regardless of whether the hospital mandates across-the-board use of the system, or a physician “prescribes” the system for a particular patient. The purpose of the current study is to see what we could do to motivate a nurse to want to wheel the system into the hospital room so their patient could use it. In our pilots and clinical trials we have not had one patient out of over 400 who refused to use this system once the study nurse wheeled it bedside and explained how it should be used, so technology acceptance has not been an issue with patients. The TAM Attitude (satisfaction) and Intent to Use measures are very appropriate in this situation with nurses, and the Ease of Use (difficulty for the nurse to set up and difficulty for the patient to use) and Usefulness (both for the patient and in offloading nurse effort) are relevant as well.

In the rest of this paper we describe the design and evaluation of the bedside patient education system, the version we created for the nursing staff, and the results of a pilot evaluation in which we had the nursing staff interact with the tailored system to promote acceptance.

2 Related Work

There have been over 20 studies applying the Technology Acceptance Model to clinician acceptance of health IT, demonstrating its applicability in the healthcare context [2]. These studies have involved technologies from electronic medical record systems to telemedicine and PDA-based systems, and investigated acceptance among physicians, nurses and other clinicians. However, we are not aware of any acceptance interventions that used persuasive messages or agents to promote acceptance among target users, including health professionals.

3 The Virtual Nurse System

The hospital discharge virtual agent system teaches patients about their post-discharge self-care regimen while they are still in their hospital beds. In order to make the system as acceptable and effective as possible with patients who had a wide range of health and computer literacy, we designed the interface to incorporate an animated virtual nurse who embodies best practices in health communication for patients with inadequate health literacy as well as incorporate the relational behaviors described above [9]. The agent is deployed on a wheeled kiosk with a touch screen display attached to an articulated arm that can be positioned in front of patients while they are in bed (Figure 2). The agent speaks using synthetic speech, driven by a hierarchical transition network-based dialogue engine with template-based text generation, and conversational nonverbal behavior is generated using the BEAT text-to-embodied-speech translation engine [10]. The agent's repertoire of nonverbal behavior includes a range of hand gestures (beats, emblematics, and deictics [11], including deictics at a document held by the agent), headnods, eyebrow raises, gazing at and away from the user, posture shifts [12], visemes, and facial displays of affect. User contributions to the conversation are made via a multiple-choice menu of utterances dynamically updated at each turn of the conversation [9].

Patients spend approximately half an hour with the RA, reviewing the layout and contents of an "After Hospital Care Plan" booklet that contains their personal medical information designed to assist them in transitioning from the hospital to post-hospital care. The paper booklet is given to patients before their conversation with the agent, and the agent reviews a digital version of the booklet in the interface, so that patients can follow along with the system's explanation of their paper booklets.

Pilot evaluations of the agent show that it is well liked and accepted by patients, and 74% said they prefer receiving their discharge instructions from the agent compared to their human doctors or nurses [6]. We have also found significantly higher levels of satisfaction among patients with inadequate health literacy [13], low computer literacy, and depressive symptoms [14].

We developed a dialogue script to support a personalized one-on-one conversation between a member of the nursing staff and the agent, which was especially tailored to promote acceptance. Figure 3 shows a fragment of a sample interaction that typically lasts 10 minutes. The dialogue was authored to support three major objectives. First, the agent explains how it works, demonstrating how it simulates face-to-face conversation using animated conversational behavior (lines 12-17). Second, the agent explains how it was designed to provide benefits to the patients it interacts with as well as the nurses and the hospital (lines 6-8, 14, 18). Finally, a wide range of agent relational behavior was designed into the conversation to establish trust and rapport with nurses, following work that used this same behavior in automated health counselor agents to establish therapeutic alliance relationships with patients [9]. This included: close forms of greeting and farewell (calling the nurse by his or her given name, lines 1, 20); expressing happiness to see the nurse (line 3) and desire to work together with the nurse (line 8); social chat (lines 10-11); complimenting the nurse (lines 4-5); and desire to talk to the nurse again (lines 20). In addition, the simple act of conducting a sample conversation with the agent provided nurses with first-hand knowledge of how the interaction worked and a sample of the patient experience.

4 Evaluation

To evaluate whether one-on-one conversations with the virtual agent promoted acceptance of the system among nurses, we conducted a randomized controlled trial of how the agent was introduced to the nursing staff. All introductions were conducted at regularly scheduled nursing in-service meetings. At these meetings, all nurses received a 15-minute presentation accompanied by PowerPoint slides that described the agent based intervention designed to reduce 30 day all cause re-hospitalizations. This presentation included a description of the study purpose and protocol, and a brief description of the agent including a static image of the agent and the interface. Following this presentation, volunteers were recruited for the study, told it would take another 30 minutes of their time, and offered a $10 gift certificate in compensation for their participation. The study was approved by the Boston Medical Center IRB, and all participants provided informed consent.

1. A: Hello there! My name is Elizabeth. And you are Mary, is that right?
2. U: Yes, that's right.
3. A: It's great to meet you Mary!
4. A: The researchers and clinicians I have been working with have told me how hard you all work, and you all deserve a pat on the back!
5. U: Thank you!
6. A: I am here at Boston Medical Center to help patients learn about their discharge instructions. I would like to help the hospital become more efficient and modernized, to provide a better environment for staff, patients, and their families.
7. U: Sounds good.
8. A: I am hoping that we can work together and keep BMC one of the greatest hospitals in Boston!
9. U: Great!
10. A: Speaking of Boston, how about those Red Sox?
11. U: They are really something!

    …
12. A: So by now, you are probably wondering how I work?
13. U: Yes, actually I was.
14. A: I am glad you want to know about me. Well, first I want to tell you more about who I am. I am a virtual computer agent that was created by computer scientists at Northeastern University. These scientists are interested in using computer characters like me, to help patients in the hospital understand their medications, follow-up appointments, and diagnosis, and feel confident in knowing all of this information when they leave the hospital, in order to help reduce the chance of a re-hospitalization.

    …
15. A: Now that you know how I speak, you will understand how I am able to move. I am also programmed to move or gesture in certain ways, to seem more human like, and to emphasize specific topics. My creators have found this helps with patients' acceptance of me.
16. A: For example, if I am happy, I can smile like this.
17. A: If I am concerned, I will look like this.

    …
18. A: My greatest advantage is that I can explain things to the patient as many times as they like. Since I am a computer, I never get tired or bored. I can also make sure that the patient understands their discharge instructions, by giving them a quiz, on the important aspects of their medications and appointments. If they do not get the answer correct, I can go back over their booklet, and explain it in a different way, to make sure that they understand.
19. U: Okay.

    …
20. A: Well, I won't take up anymore of your time today. Thank you for talking with me. It was great to meet you Mary, and I hope we can see each other again sometime. Have a great day and I hope to see you again soon! Bye!
21. U: It was great to meet you too!

Fig. 3. Excerpt from Acceptance Promotion Dialogue Lines beginning with “A:” are agent utterances; “U:” are User utterances

4.1 Participants

Eighteen nurses volunteered to participate in the study. Participants had zero to 30 years of service (average 10.8) and scored at the midpoint on a measure of computer literacy (average 4.2 on a composite scale of 1=no experience to 7=expert) and moderately high on attitude towards computers (average 5.4 on a composite scale of 1=negative attitude to 7=positive attitude) [15].

4.2 Measures

Primary self-report outcome measures were adapted from prior TAM studies for Perceived Usefulness, Ease of Use, Attitude, and Use Intention [3]. In addition, we used scale self-report items to assess how threatened the nurses were by the agent, and questions about nurses' professional and social relationship with the agent. All items are shown in Table 1.

4.3 Procedure

Recruited nurses were randomized into intervention and control groups. Intervention participants were taken to another room where they each conducted a personalized acceptance promotion conversation with the virtual agent as described above. Prior to each conversation, a research assistant entered the participant's given name into the system so that the agent could address the participant by name. agent conversations were conducted on the same mobile touch screen computer cart used for hospital bedside patient education (Figure 2). Immediately following this conversation, intervention participants were asked to fill out the self-report questionnaire for all measures.

In contrast, control group participants were asked to fill out the same questionnaire immediately following recruitment (i.e., shortly after the in-service presentation).

Approximately 30 days after the in-service briefing an attempt was made to reach all study participants and have them fill out the outcome measures portion of the questionnaire again.

4.4 Results

Exactly half of the 18 participants were randomized into each group. There were no significant differences between intervention and control groups on years of service, computer literacy or attitudes towards computers.

Outcome results are presented in Table 1. Intervention group participants scored the RA significantly higher on Perceived Usefulness, Ease of Use, and Attitude (overall satisfaction), and expressed a significantly greater intention to use the agent system, compared to those in the control group. Intervention group participants also expressed a significantly higher degree of trust in the agent character, felt that she was more a part of their "healthcare team", and felt significantly less threatened by the agent character, compared to those in the control group.

At 30 days, we were only successful in obtaining follow-up measures from eight participants. However, the trends on all measures remained consistent, and several differences between the groups remained statistically significant, including Ease of Use and Use Intention.

Table 1. Self-Report Measures and Outcomes Significance values from t-tests for independent means. "Elizabeth" is the name of the Virtual Agent.

	After In-Service (N=18)			After 30 Days (N=8)		
Questions	**Agent**	**Control**	**p**	**Agent**	**Control**	**p**
Perceived Usefulness (α=.95) (1=disagree completely … 7=agree completely) (1) Using the Elizabeth patient education system enhances my effectiveness on the job. (2) Using the Elizabeth patient education system allows me to get my work done more quickly. (3) The Elizabeth patient education system is useful to me.	6.4	4.0	**.002**	5.5	3.8	.1
Ease of Use (α=.90) (1=disagree completely … 7=agree completely) (1) The Elizabeth patient education system is easy to use. (2) The Elizabeth patient education system is easy to operate. (3) It is easy to get the Elizabeth patient education system to do what I want it to do.	6.5	4.2	**.001**	6.3	3.3	**.013**
Attitude (1=not at all … 7=very satisfied) How satisfied are you with the Elizabeth patient education system?	6.6	4.5	**.002**	6.0	5.0	.388
Use Intention (α=.94) (1=disagree completely … 7=agree completely) (1) If I could, I would use the Elizabeth patient education system as a routine part of my job over the next year. (2) If I could, I would use the Elizabeth patient education system at every opportunity over the next year. (3) I would recommend the Elizabeth patient education system to my patients.	6.4	4.7	**.020**	6.9	4.2	**<.001**
I feel threatened by Elizabeth. (1=disagree completely … 7=agree completely)	1.4	3.4	**.025**	2.0	3.5	.308
I consider Elizabeth part of my healthcare team. (1=disagree completely … 7=agree completely)	6.6	3.5	**.002**	6.8	3.3	**.001**
How would characterize your relationship with Elizabeth? (1=complete stranger … 7=close friend)	5.0	3.5	.144	4.8	2.7	.059
How much do you trust Elizabeth? (1=not at all … 7=very much)	6.0	4.4	**.042**	6.0	4.0	.078

5 Conclusions and Future Work

The rapport and trust-building dialogue was successful in increasing trust in the virtual agent, and significantly increased nurses' intention to use the patient education system. Whether this effect was due to increases in Perceived Usefulness or Ease of Use (the primary determining factors according to the Technology Acceptance Model) cannot be determined, since they were both significantly higher in the intervention group. We also cannot determine whether the influence was achieved via central or peripheral route persuasion cues (per the elaboration likelihood model), since these were not manipulated separately: future studies that parametrically vary these factors are needed. However, the combination of a personalized conversation with the RA and the specific promotion dialogue together did lead to a significant reduction in nurses' feelings of being threatened by the virtual agent.

This study has several limitations, including the small sample size and lack of an actual behavioral measure of system use following the intervention. In our case we could not measure this since the decision to use the agent with a patient was always made by research staff and the nurses in this technology acceptance study never had any actual choice or say in the decision. However, previous TAM studies have demonstrated that self-reported Use Intention (as measured here) is a highly significant predictor of actual use of voluntary use systems. The sample size was also too small to verify the causal path relationships among the TAM factors. This was also one of the primary reasons why we chose to use the original, simpler TAM and focus on the essential elements of attitude and intent to use, rather than using one of the more recent and complex models (e.g. UTAUT [16]). In addition, some of the differences observed may have been simply due to the intervention group having direct experience with the system being evaluated whereas the control group only had indirect experience, although neither group experienced the hospital discharge interaction that was given to patients. We also did not control for the additional time and attention given to the intervention group nurses. Finally, our approach, as with any attempt at manipulating acceptance, will likely have little impact if the IT system at issue is inherently either extremely acceptable or unacceptable.

Relational behaviors have been successfully used in animated conversational agents to build therapeutic alliance relationships with patients to increase satisfaction, engagement and adherence to medication and other health behavior regimens [9]. However, to our knowledge, this is the first time that automated relationship-building behavior by an interface agent has been used with clinicians, and the first time it has been successfully used to promote technology acceptance.

Future work includes assessment of whether trust-building conversations such as the one presented here lead to actual increased use, whether ongoing relational agent conversations are effective at maintaining clinician acceptance over time, and whether this technology can be used to influence other aspects of clinician behavior. The techniques described here could also be used in many other areas within and beyond healthcare to promote technology acceptance among professionals.

Acknowledgments. Thanks to Julie O'Donnell, Shaula Forsythe, and Gail Burniske for their assistance with the study, and to the many members of the Re-Engineered Discharge (RED) project team at Boston Medical Center and Northeastern University. This work was funded by grant R01 HL081307 from the US NIH National Heart, Lung, and Blood Institute.

References

1. Blumenthal, D.: Stimulating the Adoption of Health Information Technology. The New England Journal of Medicine 360, 1477–1479 (2009)
2. Holden, R., Karsh, B.: The technology acceptance model: its past and its future in health care. Journal of Biomedical Informatics (2009)
3. Davis, F.D.: User acceptance of information technology: System characteristics, user perceptions and behavioral impacts. International Journal of Man-Machine Studies 38, 475–487 (1993)
4. Bhattacherjee, A., Sanford, C.: Influence Processes for Information Technology Acceptance: An Elaboration Likelihood Model. MIS Quarterly 30, 805–825 (2006)
5. Petty, R., Cacioppo, J.: Attitudes and Persuasion: Classic and Contemporary Approaches. Westview Press, Boulder (1996)
6. Bickmore, T., Pfeifer, L., Jack, B.W.: Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents. In: ACM SIGCHI Conference on Human Factors in Computing Systems, CHI (2009)
7. Bickmore, T.W., Pfeifer, L.M., Paasche-Orlow, M.K.: Health Document Explanation by Virtual Agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 183–196. Springer, Heidelberg (2007)
8. Bickmore, T., Picard, R.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer Human Interaction 12, 293–327 (2005)
9. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ. Couns. 59, 21–30 (2005)
10. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: SIGGRAPH 2001, Los Angeles, CA, pp. 477–486 (2001)
11. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. Cambridge University Press, Cambridge (1992)
12. Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., Rich, C.: Non-Verbal Cues for Discourse Structure, pp. 106–115. Association for Computational Linguistics, Toulouse (2001)
13. Bickmore, T., Pfeifer, L., Byron, D., Forsythe, S., Henault, L., Jack, B., Silliman, R., Paasche-Orlow, M.: Usability of Conversational Agents by Patients with Inadequate Health Literacy: Evidence from Two Clinical Trials. Journal of Health Communication 15, 197–210 (2010)

14. Bickmore, T., Mitchell, S., Jack, B., Paasche-Orlow, M., Pfeifer, L., O'Donnell, J.: Response to a Relational Agent by Hospital Patients with Depressive Symptoms. Interacting with Computers 22, 289–298 (2010)
15. Nickell, G.S., Pinto, J.N.: The computer attitude scale. Computers in Human Behavior 2, 301–306 (1986)
16. Venkatesh, V., Morris, M., Davis, G., Davis, F.: User acceptance of information technology: toward a unified view. MIS Quarterly 27, 425–478 (2003)

Virtual Agents as Daily Assistants for Elderly or Cognitively Impaired People

Studies on Acceptance and Interaction Feasibility

Ramin Yaghoubzadeh, Marcel Kramer, Karola Pitsch, and Stefan Kopp

CITEC, Bielefeld University,
P.O. Box 10 01 31, 33501 Bielefeld, Germany
{ryaghoubzadeh,marcel.kramer,karola.pitsch}@uni-bielefeld.de,
skopp@techfak.uni-bielefeld.de

Abstract. People with cognitive impairments have problems organizing their daily life autonomously. A virtual agent as daily calendar assistant could provide valuable support, but this requires that these special user groups accept such a system and can interact with it successfully. In this paper we present studies to elucidate these questions for elderly users as well as cognitively impaired users. Results from interviews and focus groups show that acceptance can be increased by way of a participatory design method. Actual interaction studies with a prototype demonstrate the feasibility of spoken-language interaction and reveal strategies to mitigate understanding problems.

Keywords: Assistive technology, Virtual assistants, Participatory design, Spoken dialogue robustness.

1 Introduction

The number of people in need of support has been growing considerably in the last decade, and will continue to do so in the future. This applies to the sub-population of older adults, but also to cognitively impaired [1] people with congenital or acquired brain defects, whose life expectancy is nowadays like those of everybody else. People from both groups can exhibit cognitive limitations, either due to age-related mental decay, dementia, or disability. Such limitations can range from serious forgetfulness to a complete lack of the sense of time, and they can make it hard for them to autonomously organize and maintain a daily life with regular activities like meals, medication or social events, as well as extraordinary events like doctor's appointments. However, enabling such people to stay autonomous in their home environment and to have a self-determined way of living for as long as possible has been marked as one of the most important societal goals of the future.

[1] In this paper, we adopt the terminology from the ACM SIGACCESS guidelines. http://www.sigaccess.org/community/writing_guidelines

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 79–91, 2013.

Many people are able to live autonomously and with relative physical and mental well-being when given support in organizing and following a structured daily schedule. However, employing professional care services raises growing economic problems and routine tasks such as monitoring and reminding of daily activities consume a lot of valuable work hours of the caregiving personnel – time that is often not available in ambulatory care or that is better spent in uninterrupted in-depth interactions. Technology could offer some of the needed support but often suffers from a lack of ease of use and, consequently, an acceptance barrier with these special user groups.

In this paper, we present the VASA project, a cooperation with one of Europe's largest health and social care providers for both elderly people and people with various disabilities. This project aims to explore how elderly people as well as people with cognitive disabilities can be enabled to autonomously maintain and follow a daily schedule with the help of a computer-driven assistant in the form of a virtual agent. We address this problem from the perspective of the users, i.e. we ask whether and under what conditions a virtual assistant can provide suitable assistantance to such people. We break this down into two research questions: (1) are virtual agents accepted as assistants by these user groups, and which system design would be particularly preferred?; (2) how can the interaction between the agent and such users be made feasible, i.e. sufficiently robust and effective?

In the next section we review related work before we present the results from two studies carried out to elucidate these questions. Section 3 reports results from interviews and focus groups conducted as part of a participatory design approach. Section 4 presents findings from an actual interaction study with a Wizard-of-Oz prototype of a virtual agent-based daily assistant, in which we wanted to know whether robust interaction is possible and, specifically, whether interaction problems can be spotted and repaired by these user groups. Both questions are explored for elderly people and for people with congenital or acquired cognitive disabilities. Section 5 discusses the results and concludes the paper.

2 Related Work

A substantial body of work exists on special requirements for human–computer interfaces for older people or people with impairments. Jian et al. [6] provide a concise overview of fundamental remedial design practices to counter the effects of perceptive, articulatory, motor, and cognitive decline as well as reduced concentration. Likewise, the GUIDE project [4] has identified sensory, motor, as well as cognitive faculties that are subject to progressive decline, and for which specific countermeasures and guidelines can be used in interactive system design. They advocate the use of suitable user models and provide a catalogue of criteria to classify users according to their physical and mental capacities [4] by means of questions regarding their abilities for daily tasks. Williamson et al. [16]

employed participatory design in focus groups to obtain a qualitative analysis of the preferences of older people regarding multimodal reminders. They identified potential reminder tasks and concluded that the participants generally prefer reminder interactions to initiate in a unimodal way and be continued in a multimodal fashion if more information was required.

A number of projects have explored whether embodied virtual agents and speech input or output can be of benefit. The GUIDE project [4], also using focus groups, found out that spoken language was by far the most preferred interaction modality for elderly users unfamiliar with technology, and that avatars should disappear in problem-free interaction but should automatically reappear when a task problem was detected. Beskow et al. [2] applied a user-centered design approach to develop a reminder system for two users with cognitive problems following surgery. They presented a prototype for multimodal schedule management using handwriting recognition and spoken dialogue with a virtual character. Regarding speech input, mild to medium-severity articulatory impairments, such as slightly disfluent or slurred speech, have received particular research focus in recent years [19], with solutions for reduced recognition error rates ranging from the application of acoustic models of elderly speakers, to multimodal fusion of speech input and typed input of initial letters for disambiguation [3]. Even for more heavily impaired patients, solutions exist that can provide fairly reliable environment control via single-word recognition [5].

With regard to virtual agents as potential social companions, Vardoulakis et al. [14] found high acceptance ratings in a Wizard-of-Oz evaluation of a relational agent dialogue system envisioned as a social companion for older people, who were free to choose topics for the conversations. Sakai et al. [11] evaluated a Flash-based virtual agent providing back-channel feedback to monologues produced by elderly patients suffering from dementia, whom they reported as being “willing to be engaged in conversation” with the system.

Compared to the user group of senior citizens, people with congenital or acquired cognitive impairments have received relatively less attention, possibly due to this user group not having exerted great demographic pressure so far. Nonetheless, there are certain parallels with elderly people in terms of cognitive symptoms such as reduced working memory and decreased capacity for concentration, suggesting similar system designs using dialogue or conversational agents. Moreover, this group of people often shows a greater likelihood for behavioral anomalies. These are, for the most part, under control by medication. Yet, it is argued that autonomous systems in actual unsupervised use by those users need to implement dedicated security measures and coping strategies [18].

In sum, related work has shown that spoken dialogue with a virtual agent may be a suitable interaction paradigm for the elderly, but there has been very little work on cognitively impaired users with congenital or acquired cognitive impairments so far. It seems reasonable to assume that virtual conversational agents can provide a suitable way of interacting with technology for those people too, who often are illiterate and can only use simple graphical symbols or icons to interact with their technical environment. In the present work, we wanted to explore (1) if elderly users as well as cognitively impaired people would accept a virtual agent as an assistant helping with scheduling daily activities, and how both groups compare in this respect. In addition, it is not yet well understood (2) how dialogue and multimodal conversational behavior must be structured and presented on the part of the agent, in order to maximize robustness and effectivity of the interaction between these user groups and a virtual agent. Since misunderstandings by the system are inevitable, especially with these user groups and when it comes to unrestricted spoken input in natural environments, a key sub-question is how such users react to misunderstandings of the assistant and whether specific interaction strategies of the agent can help them to spot and repair them in a way suited to their cognitive limitations. The following two sections present studies aimed to shed light on these two questions.

3 Study 1: Participatory Design and Focus Groups

Achieving an optimal level of social acceptance for technology is a key guideline of our project, hence both research and development focus on the actual needs of possible users as well as their opinions, suggestions and ideas. In lack of sufficient empirical findings and design guidelines, we employed a user-centered and participatory design approach to investigate the actual needs of the targeted user groups, their ideas and reservations about a daily assistant, and their specific suggestions for an acceptable system design. To this end, we conducted interviews as ethnographic fieldwork as well as focus groups with prospective users living in the facilities of our institutional care-giving partner [9].

3.1 Initial Interviews

In exploratory interviews with elderly people, ten participants (7 female, 3 male, aged 76 to 85) took part in structured interviews, gathering their opinions about and use of modern technologies, their needs of assistance in everyday life and their daily routines. Furthermore we wanted to examine how the integration of a virtual assistant into participants' daily structures might change possible everday tasks and routines. Participants mainly named physical problems to be relevant causes of a need for assistance. Arguably, gradually decreasing cognitive abilities are barely realized – or outright denied – by older people when speaking about themselves. Several participants were hesitant to delegate any task to any other party, possibly due to biographical issues. Lacking graspable imagery considering technological possibilities, participants could rarely utter concrete wishes.

3.2 Focus Groups

Based on the results from the interviews, we devised a step-wise procedure [8] consisting of (1) ascertaining potential users' personal needs, attitudes and technical affinity by discussing a fictional case in focus groups, (2) giving them a live presentation of a system prototype and discussing their first impressions, (3) letting them interact with the agent in a Wizard-of-Oz study, solving a simple task with spoken language, followed by an interview about usability, functionality and design, and (4) having a final focus group discussing impressions and experiences in order to guide the future design of the system. Note that multi-participant focus groups were not possible with the cognitively impaired people, due to peculiarities in terms of social behavior. We thus reverted to focus groups with the care-giving personnel and personal interviews with the actual users.

In an initial focus group of six older adults (4 female, 2 male, aged 73 to 85), we found different degrees of familiarity with technology, ranging from having no cellphone to using a computer and video telephony. Regarding support requirements for a fictional senior person, participants named reminders for food and fluid intake, doctor's appointments and birthdays. In this initial focus group, only two people envisioned a virtual agent to be a helpful assistant for themselves. In the final focus group, step 4 of our method taking place after actually interacting with a Wizard-of-Oz version of the system (see Study 2), this had increased to four people. Four participants stated that they would prefer the physical appearance of a young adult of the opposite gender for the agent, two participants liked the avatar's current more child-like design.

A detailed analysis of ethnographical data on the cognitively impaired users is still underway. However, in terms of commonalities and differences, illustration and large lettering were named as desirable for the calendar as well as different modes of entering appointments and intervals for reminders. While elderly participants suggested a monthly overview, the carers of people with cognitive impairments said a daily view and just-in-time reminders would be useful. Both groups preferred the system to only take initiative when issuing a pre-set reminder, initiated by a pleasant audio signal, proceeding to uttering their name as a second call to attention, and postponing the reminder if no reaction was present, and only activating a camera if an interaction had actually started.

4 Study 2: Feasibility of Successful Interaction

Based on the results of the interviews and focus groups, we have developed a first prototype version of the daily assistant. Following internal pilot studies with the conversational agent, Study 2 served to explore whether successful interaction with a conversational agent is feasible for our special user groups. In that, besides trying to achieve successful dialogues in real interactions, we were interested in what happens when interaction problems in the form of misunderstandings on the part of the system – inevitable in speech-based conversational agents – arise. We first describe the system prototype and its Wizard-of-Oz version, used in this study, before the procedure and the results are presented.

4.1 Prototype System and Experimental Setup

In the targeted final system, interactions between the virtual assistant and the human user shall be conducted via a computer or TV screen (optionally a touch screen) and standard speakers, and a directional microphone for automatic speech recognition on the input side. One function of the daily assistant will be to assist in organizing and following a day schedule, which is presented graphically next to the agent. Fig. 1 (left) shows the current version of the system, also used in Study 2.

The virtual agent acts as an assistant in managing appointments on the calendar. The current prototype uses an agent (Fig. 1, left) driven by the ACE architecture [7] for conversational agents, and currently being ported to the ASAP realizer system optimized for incremental dialogue [15]. NLU is realized by spotting keywords and simple grammatical structures (such as direct declaration of new appointments, questions about the schedule) in free-form speech data as acquired from n-gram dictation-style recognizers, delivered by a Windows Speech Recognition or a Nuance Dragon NaturallySpeaking backend. The NLU module is capable of providing incremental results from the parser to the dialogue manager, which is an independent implementation following the basic tenets of the Info-State approach [13]. The current version of the system is capable of going over the weekly schedule with the user in an interruptible fashion and inserting or removing appointments as the user instructs. Schedule data in discussion are provided by the system in a multimodal fashion, using iconic visualization and highlighting on the calendar board, speech synthesis, gaze, head movements, as well as pointing and deictic gestures by the agent. For laboratory setups, an eye tracker component tracks the subject's visual focus to ascertain their capacity to follow the dialogue, in particular whether and where they look at the schedule, and when they turn back to adress the agent.

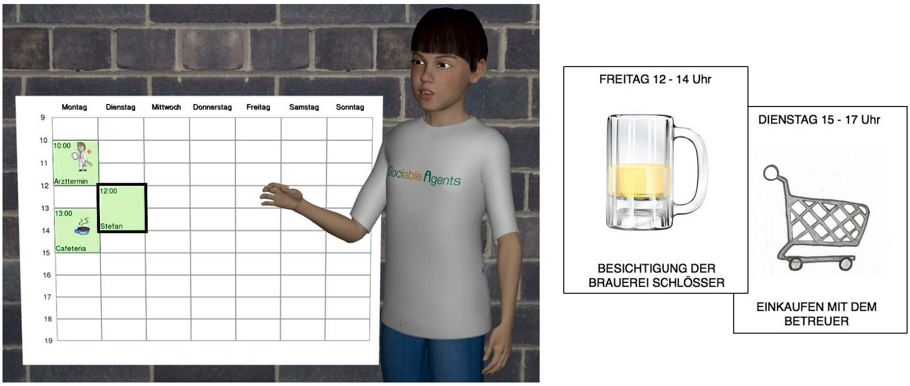


Fig. 1. Left: Daily assistant "Billie" presenting the user's appointments; **Right:** appointment cue cards used in the study

In the present study, we did not use the autonomous prototype system but a Wizard-of-Oz version in order to cancel out errors from accidental misinterpretation and circumvent the need to train speech recognizers for each participant beforehand. The Wizard listened to the user's spoken input and picked system responses from a graphical interface, which were synthesized in realtime by the character engine. The Wizard's GUI featured about 100 predetermined responses, ranging from generic feedback (e.g., "Yes", "Ok") to appointment-specific utterances like repetitions ("So, you have a dentist appointment tomorrow at 10") or requests ("And at what time will it take place?"). To cope with users' alterations of the order or contents of appointments, responses could also be flexibly configured manually by the wizard by typing abbreviated descriptions of, e.g., days, times, or activities, that were automatically expanded and sent to the character engine by the GUI.

The goal of Study 2 was twofold. First we wanted to see whether these user groups can interact with a conversational agent in a smooth and trouble-free way. Second, we also wanted to see what happens in cases when misunderstandings of the system occur. To this end, the following study was conducted with the two user groups, elderly users and cognitively impaired users.

4.2 Participants

Cognitively impaired participants were recruited from the clientele of an institution where people of all ages with various cognitive impairments could attend computer and photography courses; they thus had prior experience with interacting with computers via standard input devices. All participants (n=11, 7 male, 4 female, aged 24 to 57) had light to medium mental retardation (approximately F70–F71 on the APA DSM scale [1]), no unmedicated behavioral anomalies, and normal to slightly impaired speech capabilities. They were partly illiterate, but capable of reading at least isolated words, times, or days of the week.

The elderly participants (n=6, 4 female, 2 male, aged 76 to 85) were recruited from the focus groups in the retirement home. They were capable of independent living in their home environment and did not have cognitive impairments at a clinically relevant level.

4.3 Method

Participants were seated in front of the agent as shown in Fig. 3 (top). The system was explained as displaying an avatar able to engage in schedule-related dialog in normal spoken language. Participants were instructed to verbally request the system to insert appointments into their own calendar on their behalf. All appointments were given in form of cue cards (Fig. 1, right), designed to contain events from the typical week of the user group as well as a selection of recreational events. Participants were provided with seven cue cards on which the day, time, and the description of the event were given along with an iconic representation of the topic (recall that some of the cognitively impaired people were illiterate). All cards were explained individually beforehand. Participants were free to peruse the cards at any time during the interaction.

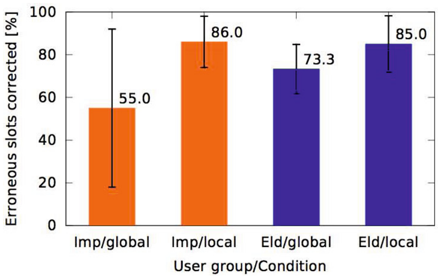
Upon giving the instructions, the staff left the room[2]. The wizard made the agent introduce itself and then verbally describe two items already entered in the fictional schedule, highlighting them on the virtual calendar board. The participants were then to start entering their own seven appointments.

During the actual interaction, when instructed by the user to enter an appointment, the agent would proceed with a predefined scheme of introducing errors at certain times: for cards 1, 4 and 5, no error would be introduced; for cards 2 and 7 the time would be misunderstood; for card 3, the topic would be altered to a similar-souding incorrect one; for card 6, both time and day would be altered. All items and errors were presented both verbally by the agent and graphically in the calendar in a small confirmation dialogue started by the agent after each insertion of a new appointment. For this we realized two different strategies: In the *global* condition, the agent summarized items in one coherent utterance ("So you will go shopping, Wednesday at 9?") and waited for feedback; in the *local* conditon slots were presented one-by-one ('So it is on Wednesday?', 'At 9 o'clock?', 'And you will go shopping.'), awaiting feedback after each step. The visual calendar either displayed the whole appointment at once or revealed it slot-by-slot (day, time, content), respectively. In both conditions, the final seventh appointment was summarized by displaying it on the visual calendar only ('So I will enter it like this?'), to determine whether subjects could detect errors that were apparent in visual form only. Whenever subjects spotted and corrected an error, the agent revised the corresponding slots and reiterated the clarification question until subjects signaled the appointment as being correct. The agent then proceeded to ask for more appointments. Whenever the participants did not take the initiative in idle conversation for some time, the agent asked them whether there was anything else to do. If participants repeatedly failed to react when seven topics had not been negotiated yet, or did not desire to add any more appointments, the wizard would proceed to valediction and terminate the interaction.

After the end of the interaction, participants were interviewed according to a structured interview plan. Its first part comprised nine questions about system usability and acceptability, each with a request to give a rating on a graded scale (1 = best, 6 = worst). Questions were: (1) How did you like to plan your week with Billie? (2) Did Billie always do what you wanted? (3) Did Billie provide sufficient help? (4) Could you understand Billie's language well? (5) Did Billie express himself in an easy way? (6) Did Billie understand you correctly? (7) How did you like the image-based calendar? (8) Do you think Billie could be of help to you? (9) Do you think Billie could be a good appointment assistant?

Since we expected impaired participants not to be able to handle abstract numbers consistently, we provided a chart which showed the grades along with appropriate emotional facial icons and asked participants to point at the most fitting one. The second part of the interview comprised open questions about design wishes and desirable additional support functions of the virtual assistant.

[2] One team member stayed in the room with the cognitively impaired people in case a participant would be overwhelmed or panic.



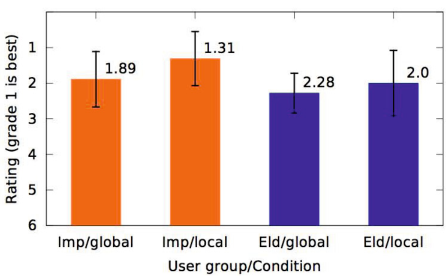


Fig. 2. Left: Mean error repair rates in *global* and *local* conditions of cognitively impaired users (abbreviated Imp), and elderly users (Eld). **Right:** Averaged ratings of the system (questions see text) by both user groups, in both conditions.

4.4 Quantitative Results: Cognitively Impaired Users

Ten of the cognitively impaired participants (n=11) succeded in entering their appointments, although one of them skipped the final two cards and stated there were no more appointments. One additional card was omitted due to a handling problem. One of the eleven participants did not enter any appointments even after repeated inquiry by the agent, and was disregarded for error repair data analysis since the interaction went from greeting to valediction directly. Recordings and log files from the interactions were subjected to an analysis of how often the users were able to spot and repair a misunderstanding by the agent. Fig. 2, left, shows the mean rates of successful error repairs for both user groups. In the group of cognitively impaired users, 75% of the introduced errors were spotted by the participants overall. More errors were noticed and repaired in the local condition (86%) than in the global condition (55%). Note that with five subjects per condition statistical power is too little for tests of significance, but the difference is considerable in absolute terms. In addition, the variance is much smaller in the local condition than in the global, i.e. inter-individual differences were less pronounced in the more successful condition. Errors in the visual-calendar-only items (not summarized verbally by the agent) were spotted in 66% of cases. The answers to the usability questions, on average, follow the same trend of slightly better ratings in the *local* conditon (*global*: $\mu = 1.89, \sigma = 0.78$; *local*: $\mu = 1.31, \sigma = 0.76$; see Fig. 2, right).

4.5 Quantitative Results: Elderly Users

All elderly participants (n=6) succeeded in interacting with the system and fulfilling the task; in one case the volume level had to be increased after the agent's introduction. The error repairs rates were 73.3% (*global*) and 85.7% (*local*), respectively. Errors in the visual-calendar-only items (without accompanying verbal summary of the schedule data) were spotted in all cases (100%). Compared to the cognitively impaired user group, the conditions did not noticeably influence usability ratings of the system (*global*: $\mu = 2.28, \sigma = 0.56$; *local*: $\mu = 2.0, \sigma = 0.92$). Further, the interaction styles were more varied than in the other user group, ranging from brief to verbose styles; one participant in particular selected a very verbose style and role-played a back story to the appointments, compatible with accounts of a propensity amongst elderly people for a "social" a-priori interaction style with virtual agents (cf. [17]).

4.6 Qualitative Analysis: Issues of Interactional Conduct

The advantage for the *local* repair elicitation strategy, where day, time and topic are presented one-by-one, was more prominent for the group of people with cognitive impairments – possibly they derive more benefit from this stepwise, simple-structured process. To further shed light on those dialogue structures that fostered or prevented error detection and correction in both conditions, we performed analyses of the subjects' interactional conduct and especially the repair attempts encountered. Video recordings of the interaction were analyzed for the collaborative processes between the parties, their verbal and nonverbal actions and the resulting sequential structures [9]. Initial investigation of the video data reveals the importance of the system's ability to invite the user to explicitly compare the tentative calendar entry with the original information. By analyzing the orientation or locus of attention (screen/cue card) of the users one can identify those moments at which they explicitly compared the entry in question with their original information.

For example, in the fragment shown in Fig. 3), the user (here, a cognitively impaired person) interacts with the system in the *global* condition. The user has successfully entered one appointment and now proceeds to the second one, containing the first introduced error. The user reads the appointment off the cue card (Fig. 3, line 01), the gaze alternating between the cue card and the screen. After the utterance, the user looks up to the agent (line 01), waiting for a reponse. The agent repeats the appointment naming the correct day and activity, but incorrect time information (line 02). The user seems to notice the error and looks down to the cue card, apparently checking the information. The user then gazes back to the screen (Fig. 3, top right) and utters an excuse (line 03), and, glancing at the cue card, initiates a self-repair (line 03) before redirecting their gaze to the agent. In subsequent turns, not shown here, an exchange of excuses takes places and the agent repeats the corrected information, which the user explicitely ratifies as being correct while double-checking the cue card.

This short fragment reveals that data entry, error detection and information comparison corresponded directly to gaze shifts between the screen and the card at sequentially relevant moments. The question arises how and at which precise moments users come to realize a problem, how the system can recognize this from the user's gaze behavior or to which extent it can provide orienting devices which invite the user to check for data correctness.


```
01   U:        dann von siebzehn uhr bis neunzehn uhr am dienstag (-) der kochkurs.|
               then from    5pm      to     7pm       on  tuesday     the cooking class
     U-gaze:   @cuecard----------@screen----------------@cuecard-----@screen---------
                                                                                  |#Img.1
02   A:        (1.6) also haben sie am dienstag um fünfzehn uhr kochkurs.|
                     so    you have  on tuesday  at       3pm     cooking class.
     U-gaze:   ----------------------------------------------------------|@cuecard------
                                                                          |#Img.2
03   U:        (1.1) äh=nein schuldigung. |(-) siebzehn bis neunzehn uhr.
                     um  no      sorry              5pm     to    7pm
     U-gaze:   -------@screen------------------@cuecard------@screen-------------------
                                          |#Img.3
```

Fig. 3. Transcript and anonymized snapshots from a repair attempt in Study 2. Transcription follows the GAT conventions [12]: pause length in parentheses, short pauses indicated by (-). Times of the three frames indicated by #Img.x.

5 Discussion and Conclusions

In this paper we have presented our first results of a project that explores if and how virtual agents can be employed to assist people with cognitive (and possibly other) limitations in managing their daily schedule and calendar. Our studies involve elderly users as well as, to the best of our knowledge for the first time, cognitively impaired users to address questions of (1) acceptability and (2) feasibility of symmetrical spoken-dialogue interaction. As for the first question, elderly people were found to be more reluctant to use such a system than the (younger) cognitively impaired people. In the early steps of the participatory process, elderly participants tended to recognize the usefulness of the assistive system mostly for third persons, but not for themselves. Notably, interviews, focus groups and encounters with a system prototype that actively engaged the prospective users in the system design, helped to lower this acceptability barrier – that is, the participatory design process itself helped to enhance the acceptance of the technology. Furthermore, such means of participatory design also resulted in specific design suggestions like the preferred form of reminders or a preference for spoken language interaction with the agent. As for the second question, our results show that both groups are also willing and, in principle, able to engage in a spoken-language conversational interaction with the agent. Besides confirming the well-known social effects of a virtual humanoid agent, e.g. eliciting story-telling from some elderly people, we could also show that both user groups are in fact able to handle the common interaction problems speech-based systems bring along, but that this needs to be – and can be – supported by the dialogue strategies employed by the system. Our results suggest that the optimal system behavior comprises explicit confirmation dialogues with information presentation in small chunks or installments. The importance of simple information presentation for elderly people and people with disabilities is widely reported in literature (e.g. [6], [4]). However, in contrast to other studies where users prefered low-verbosity system responses [10], the repetitive and verbose small-chunks condition did not lead to negative consequences for the evaluation of the system or the agent by the users, though future work needs to address the question whether this strategy might turn out unnecessary and bothersome in long-term interaction. Another crucial next step, currently underway, is to see if the autonomous virtual agent can also prove to be a suitable assistant in settings like the one explored here.

Acknowledgements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence 'Cognitive Interaction Technology' (CITEC). Additionally, Karola Pitsch acknowledges the financial support from the Volkswagen Foundation (Dilthey Fellowship). We thank the project members from v. Bodelschwinghsche Stiftungen Bethel and all the participants in our studies.

References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR, 4th edn. American Psychiatric Publ., Arlington (2000)
2. Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., Tobiasson, H.: The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, pp. 300–303 (2009)
3. Fager, S.K., Beukelman, D.R., Jakobs, T., Hosom, J.-P.: Evaluation of a Speech Recognition Prototype for Speakers with Moderate and Severe Dysarthria: A Preliminary Report. Augmentative and Alternative Comm. 26(4), 267–277 (2010)
4. GUIDE Consortium: User Interaction & Application Requirements - Deliverable D2.1 (2011)
5. Hawley, M.S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., Parker, M., Hatzis, A., O'Neill, P., Palmer, R.: A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics 29(5), 586–593 (2007)
6. Jian, C., Sasse, N., von Steinbüchel-Rheinwall, N., Schafmeister, F., Shi, H., Rachuy, C., Schmidt, H.: Towards effective, efficient and elderly-friendly multimodal interaction. In: Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2011), article 45, pp. 1–8. ACM, New York (2011)
7. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1), 39–52 (2004)
8. Kramer, M., Pitsch, K.: "Gibt es noch irgendwas, das Sie in der Woche machen wollen?" - Terminplanung mit dem virtuellen Agenten Billie. A peer-reviewed abstract. Presented at the 51st Meeting of Arbeitskreis Angewandte Gesprächsforschung, Dortmund (2012)

9. Kramer, M., Yaghoubzadeh, R., Kopp, S., Pitsch, K.: A Conversational Virtual Human as Autonomous Assistant for Elderly and Cognitively Impaired Users? Social Acceptability and Design Considerations. In: Proceedings of the INFORMATIK 2013 Workshop "Who is Afraid of Autonomous Machines? Transgressions Between Body, Mind, and Technology", Koblenz, Germany (in press, 2013)
10. Okato, Y., Kato, K., Yamamoto, M., Itahashi, S.: System-user interaction and response strategy in spoken dialogue system. In: 5th Int. Conf. on Spoken Language Processing/7th Australian Int. Speech Science and Tech. Conf., Sydney (1998)
11. Sakai, Y., Nonaka, Y., Yasuda, K., Nakano, Y.I.: Listener agent for elderly people with dementia. In: Proceedings of the 7th Annual ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI 2012), pp. 199–200. ACM, New York (2012)
12. Selting, M., et al.: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion 10, 353–402 (2009)
13. Traum, D., Larsson, S.: The Information State Approach to Dialogue Management. In: Smith, Kuppevelt (eds.) Current and New Directions in Discourse & Dialogue, pp. 325–353. Kluwer Academic Publishers, Dordrecht (2003)
14. Vardoulakis, L.P., Ring, L., Barry, B., Sidner, C.L., Bickmore, T.: Designing Relational Agents as Long Term Social Companions for Older Adults. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 289–302. Springer, Heidelberg (2012)
15. van Welbergen, H., Reidsma, D., Kopp, S.: An Incremental Multimodal Realizer for Behavior Co-Articulation and Coordination. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 175–188. Springer, Heidelberg (2012)
16. Williamson, J.R., McGee-Lennon, M., Brewster, S.: Designing multimodal reminders for the home: pairing content with presentation. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI 2012), pp. 445–448. ACM, New York (2012)
17. Wolters, M., Kallirroi Georgila, K., Moore, J.D., MacPherson, S.E.: Being Old Doesn't Mean Acting Old: How Older Users Interact with Spoken Dialog Systems. ACM Trans. Access. Comput. 2(1), Article 2 (2009)
18. Yaghoubzadeh, R., Kopp, S.: Toward a virtual assistant for vulnerable users: designing careful interaction. In: Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments (SMIAE 2012), pp. 13–17. Association for Computational Linguistics, Stroudsburg (2012)
19. Young, V., Mihailidis, A.: Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review. Assistive Technology 22(2), 99–112 (2010)

Modeling Brief Alcohol Intervention Dialogue with MDPs for Delivery by ECAs

Ugan Yasavur, Christine Lisetti, and Napthali Rishe

School of Computing and Information Sciences
Florida International University
Miami, FL, 33199, USA
ugan.yasavur@fiu.edu, {lisetti,rishen}@cs.fiu.edu

Abstract. This paper describes the design of a multimodal spoken dialogue system using Markov Decision Processes (MDPs) to enable embodied conversational virtual health coach agents to deliver brief interventions for lifestyle behavior change - in particular excessive alcohol consumption. Its contribution is two fold. First, it is the first attempt to-date to study stochastic dialogue policy optimization techniques in the health dialogue domain. Second, it provides a model for longer branching dialogues (in terms of number of dialogue turns and number of slots) than the usual slot filling dialogue interactions currently available (e.g. tourist information domain). In addition, the model forms the basis for the generation of a richly annotated dialogue corpus, which is essential for applying optimization methods based on reinforcement learning.

Keywords: spoken dialogue system, markov decision Processes, reinforcement learning, embodied conversational agent (ECA), intelligent virtual agents, brief intervention, behavior change, alcoholism, at-risk drinking.

1 Introduction

Excessive alcohol use, with approximately 85,000 of directly or indirectly attributable deaths per year, is the 3rd leading lifestyle-related cause of death in the United States [1]. In 2006, there were more than 1.2 million emergency room visits and 2.7 million physician office visits due to excessive drinking [2]. The economic costs of excessive alcohol consumption in 2006 were approximately \$223.5 billion [2].

Brief interventions (BI) are short, well structured, one-on-one counseling sessions, focused on specific aspects of problematic lifestyle behavior, and are ideally suited for people who drink in ways that are harmful or abusive. BIs can be delivered in 3-5 minutes [3] and (for alcohol consumption as a target) aim to moderate a person's alcohol consumption to reasonable levels and to eliminate harmful drinking behaviors. BIs are the top ranked out of 87 treatment styles in terms of efficiency [4]. It is reported that even a few minutes of advice and discussion about behavioral problems can be as effective as more extended

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 92–105, 2013.

counseling [5]. Many challenges are involved in delivering BIs to people in need, such as finding the time to administer them in busy doctors' offices, obtaining the extra training that helps staff become comfortable providing these interventions, and managing the cost of delivering the interventions [6]. Patients are often encouraged to use computer programs developed based on BI content in the doctor's waiting room or at home, or to access the intervention through the Internet, which not only offers privacy but also the ability to complete the program anywhere, any time of the day [7–9]. Although computer-based interventions adapted from one-on-one brief interventions are reported to have positive effect on reducing patients' drinking level [7, 8, 10], these programs interact with patients with menu-based text-only user interfaces [10, 11], and are less attractive to some users than one-on-one interventions. We posit that these challenges on the adoption of BIs delivered by computers can be overcome by delivering these interventions with spoken dialogue systems (SDS) integrated with multimodal interfaces and embodied conversational agents (ECAs) [12].

ECAs are animated anthropomorphic characters which is an emerging technology in multi-modal interfaces [13] that have become increasingly interesting user interfaces for a wide range of applications, such as tutoring systems [14], health behavior change systems [15, 16], and health applications [17]. ECAs can provide users with a natural anthropomorphic interface which can deliver verbal and nonverbal modalities similar to those found in face-to-face human interaction (e.g., facial expressions, hand and body gestures). The presence of non-verbal communication is shown to have different types of positive effects such as greater feelings of rapport [18] and greater feelings of trustworthiness [19] about the agent. In our current system, we use an ECA system (discussed in section 3) which can convey basic non-verbal behaviors with facial animation and lip-synchronization. However, because we currently focus on verbal communication performed by the spoken dialogue system, we do not exercise the option of controlling its non-verbal behaviors such as facial animations (i.e. its default animation engine generates facial expressions with lip-synchronization) and body gestures.

We have concentrated on the specific brief intervention which is prepared by National Institute on Alcohol Abuse and Alcoholism (NIAAA) [20] for alcohol screening and intervention. In this article, we survey related research on techniques used to-date to develop dialogue systems; we discuss the overview of our dialogue system and its integration with an Embodied Conversational Agent (ECA) and a multimodal interface; we describe our approach to modeling dialogue for brief interventions based on Markov Decision Processes (MDPs), our state-based unoptimized baseline system, and the nature of our annotated dialogue corpus that our system generates.

2 Related Research

Although, there exist no spoken dialogue systems (SDS) for the alcohol consumption domain, there has been growing interest to develop multimodal SDS which can converse, guide, assist or motivate users for different health related topics [21, 22, 17]. Dialogue management for health-related dialogue systems have so far been mostly designed based on *finite state* dialogue management mechanisms such as hierarchical transition networks [16, 21, 23]. These systems usually do not have speech recognition integration. Interaction is conducted based on menu-based choices but the system utterances is delivered vocally via text-to-speech or prerecorded voice.

Plan-based [24, 25] and *Information State Update (ISU)* based [26, 27] approaches are also employed in health-related dialogue systems. Dialogue management adapted from the existing plan-based TRIPS framework [24] has been used in the personal health assistance domain to help users with heart failure related problems [22]. SimCoach, designed to provide support and healthcare information about post-traumatic stress disorder, incorporates traditional information-state approach [26] with dialogue moves with assigned reward values [17]. While plan-based [24, 25] and ISU-based [26] approaches have been shown to provide a basis for flexible dialogue interaction, these approaches have a number of general limitations which stem from a design methodology based on the designer's intuition. These approaches require manual specification of update or inference rules which define an action for all possible dialogue situations. It is not practically possible, however, for the designer to anticipate all the possible situations of a dynamic dialogue environment. The main drawback of ISU-based approaches is that it is difficult for the dialogue designer to track the combined effect of sequentially applied updates to the information state. Since plan-based approaches highly depend on domain-dependent empirical design approach, system development can become opaque, and have high development and deployment costs.

In our system, we model brief interventions as *Markov Decision Processes (MDPs)*, which provide a stochastic data-driven framework for optimizing dialogue strategies. Optimization of dialogue strategies is usually performed by applying reinforcement learning algorithms [28]. Potential advantages of this approach in dialogue management are: **1)** data-driven development cycle, **2)** provably optimal dialogue actions, **3)** precise mathematical model for action selection, **4)** possibilities for generalization to unseen states, **5)** reduced development and deployment costs [29].

The *Reinforcement Learning (RL)* paradigm, in conjunction with fully observable and partially observable MDP dialogue models, are usually used in dialogue systems which use speech as communication medium [30–32], or which involve learning and optimization under noisy environments [31, 33]. Experiments showed that RL-based optimized approaches outperforms handcrafted dialogue management approaches [31]. So far, they have been mostly used in the tourist information domain, e.g. finding fun things to do in New Jersey [30], finding out about restaurants, hotels and bars [34], serving as a museum guide [32], with few exceptions such a system in the tutoring domain [35]. The main reason which limits the usage of RL-based dialogue management in different domains is the lack of *training dialogue corpus* for different domains. Most of the current work developed is based on annotated human-machine spoken dialogues corpora called

Communicator [36] which is used in user simulations to learn dialogue strategies [37, 38]. Versions of the *Communicator* corpora have been used by many researchers and have led to new technologies for speech and language processing. Therefore, annotated dialogue corpus is essential for performant RL-based systems.

Alternative to user simulation-based learning and using existing corpora is the *model-based learning* approach via collecting data from real user interactions [30]. In model-based approaches, the RL agent learns partial strategies from exploratory data generated by dialogues with real users. In model-based approaches a model represents the dynamics of the dialogue to compute an approximate value of taking each action in a particular state. With a model, the problem of learning a good dialogue strategy is reduced to computing the optimal policy for choosing a dialogue action in a dialogue state. We follow the model-based approach based on fully observable MDPs with some differences from previous systems [30, 31]. Our model of the problem is represented by interconnected separate MDPs with local sub-goals and global goals (details discussed in section 4.5). The system in the tourist information domains, the model of the problem is represented by a single MDPs and the optimization is performed based on a single global goal (e.g. task completion). Our approach helps to compute local optimal dialogue strategies.

The computation of optimal dialogue strategy can be achieved with standard RL algorithms [28]. This approach requires to build initial training system which can deliver basic but unoptimized functionality, and to specify performance criteria and estimates of dialogue states.

3 System Overview

Although this article is focused on the *Dialog Manager* of our system, we give a brief overview of the overall system in which it operates (see Fig. 1). Our multimodal spoken dialogue system has a multimodal ECA-based interface where: user's speech is recognized by the *Automatic Speech Recognition (ASR)* engine[1], user's facial information processing is performed by the *Facial Processing* third-party facial processing service[2]. We use two outputs of the facial processing service for annotating our training dialogue corpus: user's gender and smile, along with a confidence value. According to the gender attribute, a brief intervention SDS can adapt its behavior because there exist different thresholds for males and females (e.g. recommended drinking limits). The history of smiles, on the other hand, can give important information about the user's experience with the system (e.g. engagement and enjoyment levels can be inferred). These two outputs do not currently have any effect on our system's behavior; they are used for annotation to create an exploratory data set in the current version of the system. Nonverbal communication is also important in delivering health interventions but the focus of this paper is verbal aspect of the interaction.

[1] Currently, Microsoft Speech API (SAPI).
[2] Currently, Sky Biometry `http://www.skybiometry.com/`

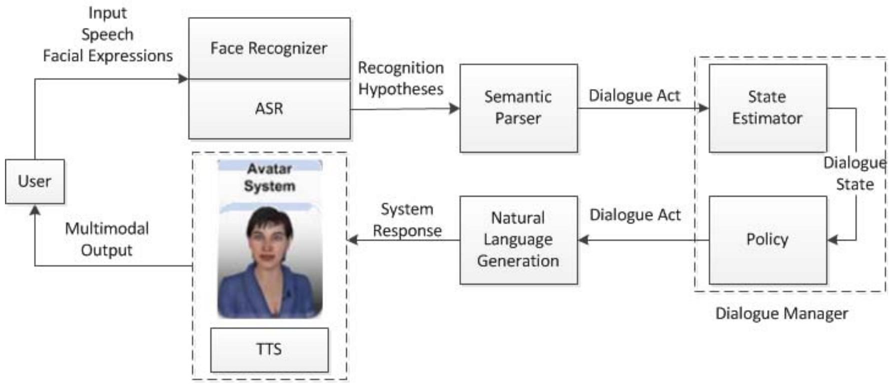


Fig. 1. Multimodal Dialogue System Architecture

Automatic Speech Recognition (ASR) hypotheses are parsed by the *Semantic Parser*[3] using context-free grammars and converted to dialogue acts. In addition to the parsing, we used a named-entity recognizer for behavioral health [40] to tag relevant entities such as alcoholic beverages. Semantic dialogue act output is passed to the *State Estimator*. The state estimator updates the *Dialogue State*, a random dialogue action is selected from the corresponding *Policy* table (state action mappings). The *Natural Language Generation* module uses a matching template for the dialogue act: if it is a question, it directly passes it to the *Text-To-Speech (TTS)* engine; if it is a confirmation, it fills the necessary parts in the template and passes it to the TTS engine. The TTS engine generates phonemes and the *Avatar System* automatically performs lip synchronization[4].

4 Approach

Compared to previous applications of RL-based dialogue systems, our brief intervention domain has several challenges: 1) the dialogue length and complexity; 2) the lack of a baseline system and of a dialogue corpus. The first challenge makes modeling harder and the dialogue size creates very large state-spaces which may cause data sparsity problems. The second challenge prevents us to optimize our dialogue policies and to evaluate optimized policies.

As we have discussed, we represent our problem with MDP framework which can be characterized by a tuple (S,A,T,R), where:

- S is a finite set of states
- A is a finite set of actions
- T is a state-transition function such that $T(s', a, s) = p(s' \mid s,a)$ which describes the probability of performing action **a** in state **s'** will lead to state **s**
- R(s, s') is a local reward function, and the objective of the SDS is to maximize the gained reward.

[3] Currently, Phoenix Parser [39].
[4] Currently, Haptek^{TM}.

4.1 Brief Intervention for Alcohol-related Health Problems

According to the clinician's guide for conducting brief interventions from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) [20], a brief intervention can be delivered in three sequential steps:

- Step 1: Asking About Alcohol Use
- Step 2: Assessing for Alcohol Use Disorders
  - Assessment of Abuse
  - Assessment of Dependence
- Step 3: Advising and Assisting according to degree of alcohol problem
  - At-risk drinkers
  - Drinkers with alcohol use disorder

To develop our dialogue content, we follow the brief intervention guide for alcohol prepared by NIAAA [6]. The goal of our dialogue system is to deliver alcohol screening and brief interventions based on the guide. We explain the first steps in detail in the following three sections (step 3 in less details for lack of space).

Brief intervention dialogue for alcohol problems can be modeled as slot-filling dialogue. However, the number of slots are larger than the applications discussed in the tourist information domain. Moreover, the number of slots that are needed to be filled by the system is not fixed. Another aspect which differs in brief intervention dialogue is that the strategy that the system needs to follow is not always constant, and needs to adapt according to inputs that the system receives from the user: there can be *no* fixed dialogue plan. According to the NIAAA guide, we identified the number of slots we need by minimizing the complexity of dialogue: the number of the slots for at-risk alcohol users is 11, and it is 9 for users with alcohol use disorder. The dialogue needs to branch according to user inputs. For the users who do not have harmful drinking patterns, interaction may end earlier. Therefore, the dialogue strategy needs to be adapted according to user's pattern of drinking.

For each dialogue state, the system has two options for dialogue action selection based on *initiative type*. One type is *user initiative* dialogue action which are usually open-ended questions. The second type is the *system initiative* dialogue action which are closed questions. According to the initiative type, the system uses different grammar types for automatic speech recognition (ASR). If the question type is user initiative (open-ended question), the system uses non-restrictive grammar. If the question type is the system imitative, the system uses restrictive grammar which only recognizes particular entities mentioned in the question (e.g. number of alcoholic beverages consumed).

4.2 Step 1: Asking about Alcohol Use

The system starts interaction by greeting and asking permission to talk about user's drinking. After receiving consent of a user, it asks single question about alcohol use (e.g. "Do you sometimes drink beer, wine, or other alcoholic beverages?"). If the client's answer is no, there is no need to continue to screening. If the client's answer is yes, the system asks about the amount of alcohol the client consumes to find out if the client is an at-risk drinker (e.g. "How many times in the past year have you had 5 or more drinks in a day?").

If a client is not an at-risk drinker, the system may finalize interaction by advising to maintain or lower drinking limits according to the situation and offer re-screening annually. If a client is an at-risk drinker, to get the complete picture of drinking, the system asks two more questions to query the drinking pattern of the client. We have demonstrated the sample dialogue for Step 1 in Table 1. The example dialogue actions to query pattern of drinking are performed in S4 and S5 dialogue turns. Since the questions asked in S4 and S5 are open-ended, the type of dialogue action is user initiative.

Table 1. Sample Dialogued during Step 1: Asking About Alcohol Use

S1:	Hi, I am [:::], Do you mind, if we talk about your drinking?
C1:	No, it is okay!
S2:	Do you sometimes drink, wine, or other alcoholic beverages?
C2:	Yes, I drink sometimes!
S3:	How many times in the past year have you had 5 or more drinks in a day?
C3:	I think at least once a week, I had around 5 drinks or more a day.
S4:	How frequently do you have an alcoholic beverage?
C4:	I think at least 3 days in a week.
S5:	On a typical drinking day, how many drinks you have?
C5:	I think 4 or 5 whiskeys.
S6:	Thanks for the information you have provided about your drinking. Next I will give you feedback about some important effects of your drinking.

In Step 1, there can be maximum 4 slots if the user is an at-risk drinker (see Table 1). The system continues to Step 2 only if a user is an at-risk drinker. Since the dialogue is branching, we have represented each distinct step or sub-step with a separate MDP (see Fig. 2). We have elicited a state-space for each MDP separately which greatly reduced state-space. We represented dialogue states in Step 1 with 5 features: **1)** Greet (G) indicates whether or not the system greeted the user and asked for permission to talk about client's drinking; **2)** Question (Q) indicates which question is being queried in the current state; **3)** Confidence (C) indicates confidence level of the speech recognizer (low, medium and high are represented by 0,1,2 respectively; confidence values 3 and 4 stand for confirmed and non-confirmed, respectively); **4)** Value (V) indicates, is the value was obtained or not for the current question; **5)** Grammar(Gram) indicates the type of grammar (restrictive or non-restrictive) used by the ASR.

For example, dialogue state *11210* indicates that the system greeted the user (G=1), the first question is queried (Q=1), the ASR confidence level is high (C=2), the type of grammar is restrictive (Gram=0). The Confidence (C), Value

(V) and Grammar (Gram) features are also used in state representations in Step 2 and Step 3. Since the Greet (G) and Question (Q) are not relevant to represent the state of the dialogue, if the system is performing Step 2 or Step 3, it is possible not to use them in order to reduce state-space. We used the same approach to reduce state spaces in each of the separate MDPs. In each step we only used relevant state features. The compact state representation helps to avoid the data-sparsity problem with limited number of training dialogues.

Table 2. Generating the dialogue for Step 1 shown in Table 1

States					Actions	Turn
G	**Q**	**C**	**V**	**Gram**		
0	0	0	0	0	GreetS	S1
1	0	2	1	0	NoConf	-
1	1	0	0	0	AskQ1S	S2
1	1	2	1	0	NoConf	-
1	2	0	0	0	AskQ2S	S3
1	2	2	1	0	NoConf	-
1	3	0	0	0	AskQ3U	S4
1	3	2	1	1	NoConf	-
1	4	0	0	0	AskQ4U	S5
1	4	2	1	1	NoConf	-
1	5	0	0	0	InformTrans1S	S6

As shown in Table 2, **dialogue actions** represent 2 types of actions for asking each question during the first time, according to the type of initiative (user or system). For each question, there are 2 types of actions to re-ask the question with user and system initiative types which are performed in the dialogue states when the system did not receive answer (i.e. Value feature of the state equals to 0 indicates that the answer was not obtained). There is also explicit confirmation action for each question to verify that the input was received, the system may also select not to confirm. If the system selects not to confirm action (i.e. showed as NoConf), it updates the dialogue state as input is received and continues with randomly selecting a dialogue action in the updated dialogue state.

A **dialogue policy** is a mapping between dialogue states and available dialogue actions in each state. In our training system, the dialogue action are randomly selected. This approach will create exploratory dialogue corpus for optimizing dialogue strategies with RL.

4.3 Step 2: Assessing for Alcohol Use Disorders

In Step 2, the system aims to determine whether or not there is a maladaptive pattern of alcohol use that is causing clinically significant impairment or distress. In this step, the system queries with 4 questions whether a client has alcohol abuse (e.g. risk of bodily harm, relationship trouble) and alcohol dependence (e.g. kept drinking despite problems, not been able to stick to drinking limits) problem. If a patient does not meet the criteria for alcohol abuse or dependence, the patient is still at-risk for developing alcohol related problems. If a patient has an alcohol use disorder (dependence or abuse), the next step (Step 3) will be different than at-risk drinkers.

Querying abuse and dependence are represented by two separate MDPs for this step. The dialogue state is represented by different features in addition to common features (C, V, Gram) for all states as discussed in Step 1. For *abuse*, we used two specific features. Question (Q) indicates which question is being queried. Since there are 4 questions (slots) for querying abuse, Q can take 1,2,3,4. The second feature specific to abuse is boolean feature Abuse (A). Since it is enough to elicit one abuse indicator, this feature is binary (0 or 1). Since it is enough to elicit one indicator, the system continues to the next step as soon as it elicits one abuse indicator. For *dependence*, a dialogue state is represented by 2 specific and 3 common features. The first specific feature is Question (range 1-7, since there are 7 questions for dependence). The second specific feature is Dependence (D) (range 0-3 which shows the number of indicators elicited: system may not elicit any dependence indicator and 3 dependence indicators is enough to elicit).

4.4 Step 3: Advising and Assisting According to Degree of Alcohol Problem

In Step 3, if the client is at-risk, the system states its conclusion according to the guideline [20] and recommends to the user to cut down his/her drinking. Then it tries to asses readiness to change based on readiness ruler approach (e.g. "On a scale of 1 to 10, how important is it for you to make a change?"). If the client is not ready to change, the system restates its concern for client's health, encourages reflection by asking positive versus negatives of drinking and reaffirms its willingness to help when the client is ready. If a client is ready to change, the system sets a goal (e.g. "How could I assist you in getting to a 7?"), agrees on a change plan and provides educational materials (e.g. pamphlets). In Step 3, for the clients who has alcohol abuse or dependence problems, the system states its conclusion, negotiates a drinking behavior goal and refers to an addiction specialist.

4.5 Modeling World with Interconnected MDPs

To address the data sparsity problem, we aimed at minimizing the number of system states used. Since the BI dialogue requires many dialogue turns between the system and a client, the number of available dialogue strategies is very large, and can make learning optimal policies infeasible with limited number of training data. To alleviate this problem, we used separate MDPs for each phase.

We represent each step or phase of the BI with one MDP with local goals and reward functions. This divided the problem into 5 interconnected MDPs (see Figure 2) but, in any interaction with the system, we use a maximum 4 MDPs,

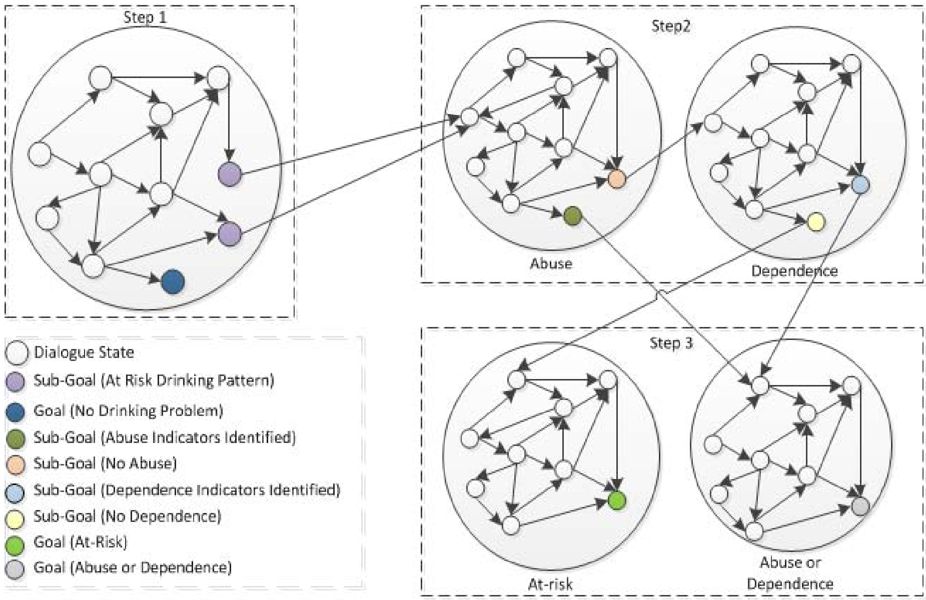


Fig. 2. Representation Of World Model With MDPs

i.e. 1) Step 1; 2) Abuse; 3) Dependence; and 4) one MDP from Step 3 based on Abuse or Dependence problem. This approach also reduced the number of required state features for each step, thus reducing the number of states required. For example for Step 1, there are 45 states with 2 action choices, which results in 2^{45} possible exploratory dialogue policies.

In *Step 1*, the reward can be awarded based on reaching one of the sub-goal states (i.e. there can be many sub-goals and goal states in each MDP) showed in Figure 2. Goal states, in the MDP representing Step 1, represent completion of the interaction, which means that a client does not have alcohol problem. Sub-goals represent identification of at-risk drinking patterns. If the system reaches one of the sub-goal states in Step 1, it is rewarded with a local reward function and the state is transited to the abuse assessment MDP in Step 2. The transitions between MDPs do not require to have stochastic transition model, thus they are deterministic.

Since there are two phases in *Step 2*, one for querying alcohol abuse and one for querying alcohol dependence, we represent Step 2 with two distinct MDPs (as shown in Figure 2), which greatly reduces number of exploratory policies without compromising fine-grained distinctions between dialogue strategies. Because the two phases are independent from each other, representing each phase with a separate MDP is appropriate. For each of the MDPs, the rewards are awarded based on reaching one of the local sub-goal states. The reward can be awarded in the stage of assessing alcohol abuse, as soon as eliciting one indicator of alcohol abuse, or completing the assessment with 4 questions without eliciting any indicator. The reward in the dependence stage is awarded based on reaching one of the sub-goal states in the MDP. To reach the goal state for dependence, the system needs to identify 3 indicators of the dependence or finish asking all of the questions without eliciting any indicator.

There are two separate MDPs for representing different phases in *Step 3*. One for representing the model for “At-risk” drinkers who does not have alcohol use disorder problems (i.e abuse and dependence). The reward is awarded upon reaching the goal state which is end of the intervention. For the client’s with abuse and dependence problems, the model is represented by MDP which is labeled with ”Abuse or Dependence”. The reward is awarded in the same way as for at-risk drinkers, although the dialogue actions are different.

In conclusion, the system is modeled with 5 MDPs. In each MDP, there are goals and/or sub-goals. Sub-goals represent that the system completed a step but that the interaction is not completed yet. Therefore each sub-goal deterministically transits dialogue state to start state of a successor MDP. At the same time, it awards the agent with the local reward. Local rewards shows how good is a dialogue policy selection for performed dialogue strategy. With this approach, learning the optimal dialogue strategy for an entire dialogue is reduced to learning optimal dialogue strategy for the each MDPs. Finally each goal represents that the interaction is completed and that there is no need to transition to another MDP. As discussed before this approach alleviates the data sparsity problem.

5 Components of the Corpus

We plan on creating a corpus from anonymized real user interactions with our training system, it will be used later for learning approximately optimal dialogue strategies. We want to make this as exploratory as possible by rich annotation. We are annotating each dialogue with several objective and subjective performance metrics. We annotate each interaction with best hypothesis of the ASR, ASR confidence scores, n-best hypothesis of ASR (with confidence), system prompts, dialogue acts from NLU, named entities, filled/confirmed slots, dialog context (speech act history), rewards and reward history, dialogue length, number of errors and confirmations.

The corpus represents the dialogues in hierarchical XML structure. Each interaction contains a sequence of turns which includes the system and client utterances, dialogue context (e.g. named entities, filled slots) and rewards. We also annotate subjective reward signals elicited from the user upon completion of the interaction by asking a few questions about ease of use, future intention to use, perceived task completion. We also annotate each dialogue with gender and smile labels with confidence value (for reasons described earlier).

6 Conclusion

In this paper, we demonstrated our approach to model relatively long and branching brief intervention dialogue with MDPs. We build an initial training system which can deliver basic unoptimized functionality. Using this training system, we will collect dialogue corpus to help solve optimization problems with RL. One of the largest obstacle building a system for a new domain is the lack of annotated data for training a model. In this project, we addressed the infrastructure needed to collect annotated data.

Acknowledgment. This material is based in part upon work supported by the National Science Foundation under Grant Nos. CNS-0821345, CNS-1126619, HRD-0833093, IIP-0829576, CNS-1057661, IIS-1052625, CNS-0959985, OISE-1157372, IIP-1237818, IIP-1215201, IIP-1230661, IIP-1026265, IIP-1058606, IIS-1213026, OISE-0730065, CCF-0938045, CNS-0747038, CNS-1018262, CCF-0937964.

References

1. Mokdad, A.H., Marks, J.S., Stroup, D.F., Gerberding, J.L.: Actual causes of death in the united states, 2000. JAMA 291(10), 1238–1245 (2000)
2. Bouchery, E.E., Harwood, H.J., Sacks, J.J., Simon, C.J., Brewer, R.D.: Economic costs of excessive alcohol consumption in the U.S., 2006. American Journal of Preventive Medicine 41(5), 516–524 (2011)
3. Moyer, A., Finney, J.W., Swearingen, C.E., Vergun, P.: Brief interventions for alcohol problems: a meta-analytic review of controlled investigations in treatment-seeking and non-treatment-seeking populations. Addiction 97(3), 279–292 (2002)
4. Miller, W.R., Wilbourne, P.L.: Mesa grande: a methodological analysis of clinical trials of treatments for alcohol use disorders. Addiction 97(3), 265–277 (2002)
5. Babor, T.F., Grant, M.: Programme on substance abuse: project on identification and management of alcohol-related problems. Report on phase ii, an randomized clinical trial of brief interventions in primary health care (1992)
6. NIAAA: NIAAA Alcohol Alert No. 66: Brief interventions (2006)
7. Riper, H., Spek, V., Boon, B., Conijn, B., Kramer, J., Martin-Abello, K., Smit, F.: Effectiveness of e-self-help interventions for curbing adult problem drinking: A meta-analysis. J. Med. Internet Res. 13(2), e42 (2011)
8. White, A., Kavanagh, D., Stallman, H., Klein, B., Kay-Lambkin, F., Proudfoot, J., Drennan, J., Connor, J., Baker, A., Hines, E., Young, R.: Online alcohol interventions: A systematic review. J. Med. Internet Res. 12(5), e62 (2010)
9. Portnoy, D.B., Scott-Sheldon, L.A.J., Johnson, B.T., Carey, M.P.: 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 (2008)
10. Hester, R.K., Squires, D.D., Delaney, H.D.: 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)
11. Mauriello, L.M., Gökbayrak, N.S., Marter, D.F.V., Paiva, A.L., Prochaska, J.M.: An Internet-Based Computer-Tailored Intervention to Promote Responsible Drinking: Findings from a Pilot Test with Employed Adults. Alcoholism Treatment 30(1), 1–15 (2011)
12. Lisetti, C., Yasavur, U., Leon, C.D., Amini, R., Rishe, N., Visser, U.: Building an On-demand Avatar-based Health Intervention for Behavior Change. FLAIRS (2012)

13. Cassell, J., Nakano, Y.I., Bickmore, T.W., Sidner, C.L., Rich, C.: Non-verbal cues for discourse structure. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL 2001, pp. 114–123 (2001)
14. Moridis, C.N., Economides, A.A.: Affective Learning: Empathetic Agents with Emotional Facial and Tone of Voice Expressions. IEEE Transactions on Affective Computing 3(3), 260–272 (2012)
15. Lisetti, C., Yasavur, U., Leon, C.D., Amini, R.: Building an On-Demand Avatar-Based Health Intervention for Behavior Change. In: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference (2012)
16. Schulman, D., Bickmore, T.W., Sidner, C.L.: An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change using Motivational Interviewing. In: Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series, pp. 61–64. Association for the Advancement of Artificial Intelligence (2011), http://www.aaai.org
17. Morbini, F., Forbell, E., DeVault, D., Sagae, K., Traum, D.R., Rizzo, A.A.: A mixed-initiative conversational dialogue system for healthcare. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2012, pp. 137–139. Association for Computational Linguistics, Stroudsburg (2012)
18. Wang, N., Gratch, J.: Don't just stare at me. In: CHI, Atlanta, GA, USA, pp. 1241–1249 (2010)
19. Kang, S.H., Gratch, J., Wang, N., Watt, J.: Does the contingency of agents' non-verbal feedback affect users' social anxiety? In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS, vol. 1, pp. 120–127. International Foundation for Autonomous Agents and Multiagent Systems (2008)
20. NIAAA: Helping Patients Who Drink Too Much, A Clinician's Guide (2007)
21. Bickmore, T.W., Schulman, D., Sidner, C.L.: A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. Journal of Biomedical Informatics 44(2), 183–197 (2011)
22. Ferguson, G., Quinn, J., Horwitz, C., Swift, M., Allen, J., Galescu, L.: Towards a Personal Health Management Assistant. Journal of Biomedical Informatics 43(5 suppl.), S13–S16 (2010)
23. Bickmore, T.W., Puskar, K., Schlenk, E.A., Pfeifer, L.M., Sereika, S.M.: Maintaining reality: Relational agents for antipsychotic medication adherence. Interacting with Computers 22(4), 276–288 (2010)
24. Ferguson, G., Allen, J., et al.: Trips: An integrated intelligent problem-solving assistant. In: Proceedings of the National Conference on Artificial Intelligence, pp. 567–573. John Wiley & Sons Ltd. (1998)
25. Bohus, D., Rudnicky, A.I.: The ravenclaw dialog management framework: Architecture and systems. Computer Speech & Language 23(3), 332–361 (2009)
26. Traum, D., Larsson, S.: The Information State Approach to Dialogue Management, pp. 325–353. Kluwer Academic Publishers (2003)
27. Bos, J., Klein, E., Lemon, O., Oka, T.: Dipper: Description and formalisation of an information-state update dialogue system architecture. In: 4th SIGdial Workshop on Discourse and Dialogue, pp. 115–124 (2003)
28. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction, vol. 1. Cambridge Univ. Press (1998)

29. Lemon, O., Pietquin, O., et al.: Machine learning for spoken dialogue systems. In: Proceedings of the European Conference on Speech Communication and Technologies (Interspeech 2007), pp. 2685–2688 (2007)
30. Singh, S., Litman, D., Kearns, M., Walker, M.: Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. Journal of Artificial Intelligence Research 16, 105–133 (2002)
31. Young, S., Gašić, M., Thomson, B., Williams, J.: Pomdp-based statistical spoken dialog systems: A review (2013)
32. Papangelis, A., Kouroupas, N., Karkaletsis, V., Makedon, F.: An adaptive dialogue system with online dialogue policy learning. In: Maglogiannis, I., Plagianakos, V., Vlahavas, I. (eds.) SETN 2012. LNCS, vol. 7297, pp. 323–330. Springer, Heidelberg (2012)
33. Frampton, M., Lemon, O.: Recent research advances in reinforcement learning in spoken dialogue systems. Knowledge Engineering Review 24(4), 375–408 (2009)
34. Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., Yu, K.: The hidden information state model: A practical framework for pomdp-based spoken dialogue management. Computer Speech & Language 24(2), 150–174 (2010)
35. Chi, M., VanLehn, K., Litman, D.: Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 224–234. Springer, Heidelberg (2010)
36. Walker, M.A., Passonneau, R., Boland, J.E.: Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, ACL 2001, pp. 515–522. Association for Computational Linguistics, Stroudsburg (2001)
37. Georgila, K., Lemon, O., Henderson, J.: Automatic annotation of communicator dialogue data for learning dialogue strategies and user simulations. In: Ninth Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL: DIALOR). Citeseer (2005)
38. Frampton, M., Lemon, O.: Learning more effective dialogue strategies using limited dialogue move features. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, pp. 185–192. Association for Computational Linguistics (2006)
39. Ward, W.: Understanding spontaneous speech: the Phoenix system. In: Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1991, pp. 365–367. IEEE Computer Society, Washington, DC (1991)
40. Yasavur, U., Amini, R., Lisetti, C., Rishe, N.: Ontology-based named entity recognizer for behavioral health. In: The Twenty-Sixth International FLAIRS Conference (2013)

Gendering the Machine: Preferred Virtual Assistant Gender and Realism in Self-Service

Jeunese Payne[1], Andrea Szymkowiak[1], Paul Robertson[1], and Graham Johnson[2]

[1] University of Abertay Dundee, Dundee DD1 1HG, UK
0501726@live.abertay.ac.uk,
{a.szymkowiak,p.robertson}@abertay.ac.uk
[2] NCR Discovery Centre, 3 Fulton Road, Dundee DD2 4SW, UK
graham.johnson@ncr.com

Abstract. A virtual agent is a human-like character that is designed to assist users in interactions with technology and virtual worlds. Research into the preferred visual characteristics of a virtual agent has focused on education-based agents, gaming avatars, and online help assistants. However, findings from these studies are not necessarily generalizable to other technologies, such as self-service checkouts (SSCO). This paper describes data from 578 participants, looking at the gender preferences of Virtual Assistants (VA) in a SSCO context and the impact of VA realism depending on user gender. Due to female participants' preference for female VAs, and an overall preference for three-dimensional characters, a realistic, female VA should be used in SSCO. The results are discussed in terms of similarity-attraction theory and social role theory.

Keywords: Interface Agent, Virtual Assistant, Gender, Visual Realism, Self-service, Psychology, Preference.

1 Introduction

Technology plays a central role in service delivery, providing several channels through which consumers can be reached, e.g., the Internet, mobile technologies and kiosks [1-2]. Many of these channels are self-service – technologies that enable users to direct the service they receive, e.g., cash withdrawals, shopping transactions etc., with minimal direct input from staff [3].

Self-service technology tends to be used less by people who value the social interaction available during face-to-face transactions [4]. This suggests that SSCO would benefit from social presence, i.e., the feeling of interacting with another person such that the perceived role of technology is reduced [5-7]. Reeves and Nass [8] argue that people already treat technologies like social actors – a tendency that could be enhanced in SSCOs by implementing a human-like interface agent in the role of Virtual Assistant (VA). Additionally, a visually present VA can improve performance by reducing the number of people that make mistakes during transactions [9].

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 106–115, 2013.

Most research into self-service technology has focused on the attitude towards and interaction with websites [4], and thus the impact of virtual characters has focused on their role in online applications [6], [10-15]. It is difficult to determine whether the effects of particular services investigated in one study can be generalized to other types of services [2], as users of different technology-based services have different motivations and rely on different types of guidance based on the context and task.

1.1 Background

Technology-based interaction with consumers reduces the capacity to meet the need for social interaction during service encounters, having a negative impact on the consumer's intention to use self-service technology [2]. An appropriate visual representation of an employee could reintroduce a positive social aspect. Conversely, a virtual agent that lacks perceived intelligence, competence or trustworthiness, or that appears eerie, unattractive, inappropriate or unhelpful, has the potential for negative outcomes such as frustration or unwillingness to use it [14], [16].

Enabling users to select an agent's appearance should provide a more veridical index of their preferences rather than requiring users to rate an agent based on critical features. This study looks at preferences for such an agent's gender and degree of realism. In this self-service context, we use the term 'VA' – a specific type of interface agent usually applied to e-commerce. An interface agent is, more generally, a computer-generated character designed to interact with users by simulating human appearance and behaviors through artificial intelligence. Dependent on their function, they are referred to as Virtual Customer Service Agent [15], Product Recommendation Agent [6], Embodied Conversational Agent [11], [17], Pedagogical Agent [18], [19-21], [31], or as a human-like or embodied agent/character [12-13], [23-27].

Research on user preferences for agent gender presents contradictory findings and trends, from a preference for same-sex virtual agents, to a preference for opposite sex virtual agents, and no gender preference at all [6], [17], [22], [28]. These differences may be due to user characteristics, but also context [28-29]. Lee [23] for example, found that participants were more open to suggestions made by a virtual character whose implied gender matched the gendered topic. In a gender-neutral task (sorting photographs) Cowell and Stanney [17] reported a trend in females towards a preference for a virtual character of the opposite sex. Conversely, Kim and Wei [22] found a preference for gender-matched virtual agents, even though the context – a computer-based algebra learning environment – may have been perceived as requiring male expertise. Baylor and Plant [28], however, revealed preferences for gender-matching in females dependent on the topic of the learning task; when the purpose of the agent was to aid users in a male-dominated topic (engineering), male agents were deemed more appropriate and similarity became less important than perceived expertise.

In online retail environments, Qiu and Benbasat [6] suggested that gender is less important to users than other demographics, finding no overt VA gender preference in an online shopping context, but rather a preference for ethnicity-matching, especially in females. McBreen et al. [11] also found little evidence of gender preference. VA gender preference may also be affected by other appearance-based differences.

McBreen & Jack [12], for example, reported that VA gender preference in an e-retail environment might be affected by its realism: participants preferred a realistic male over a realistic female agent, and a cartoonized female over a cartoonized male agent.

Virtual Assistant Characteristics. To systematically assess preferences for particular VA characteristics in SSCO, we invited participants to choose among VAs in a public technology exhibit. This helps to determine whether people prefer the idea of interacting with a male or female VA, and the extent to which he/she should resemble actual checkout staff, as opposed to idealized versions of the user, as tends to be the case for role/social models in learning environments [21], [28-30] and self-representations in shared virtual environments ('avatars') [29], [31-32].

Similarity Attraction Theory vs. Social Role Theory. There are two key theories relevant to VA characteristics. The first is similarity-attraction theory, which suggests that people are attracted to similar others [6], [20], [26-27], who tend to be perceived as more credible, resulting in increased liking [6], and in technology, reduced discomfort with the interface [33]. Similarity can be investigated on several levels, including global features such as culture, ethnicity and gender, local features such as hair color and face shape, social and personality factors such as levels of politeness or extraversion, and behavioral similarity [31]. We focus on gender and realism, since these can be directly compared to user characteristics. As it is necessary for human-like characters to be dressed, clothing was varied, but examined in an exploratory manner.

A VA created to give impressions of warmth and expertise may be deemed more appropriate in SSCOs. This is in line with the second theory of relevance, social role theory, which posits that social group roles are used to infer stereotypes about them, reflected in status and power differences [34], [35-36]. For instance, the disproportionate amount of domestic roles that women play relative to men results in the expectation that women should possess communal traits and behaviors such as friendliness. Deviations from gender norms tend to be met with surprise or moral disapproval [35]. In the example of SSCO, a warm (female) and competent (realistic) VA may be preferred, depending on stereotypical expectations of staff in such service encounters.

Gender. Similarity-attraction theory predicts preferences for gender-matching. This may be especially the case for females [6], [32] who have been found to prefer interacting with other females [37], not necessarily because they like females more than males, but because males, as a dominant group, pose more threat [38]. As a result of women's greater inter-dependence (relative to male independence), they tend to use similarity as the basis for social bonds [34]. Based on similarity-attraction theory, two hypotheses are derived with regards to VA and participant gender:

- Hypothesis 1a: Participants will more often choose VAs of the same implied gender to themselves.
- Hypothesis 1b: Female participants will show a greater preference for gender-matching than male participants.

Social role theory predicts a preference for VAs that represent typical checkout staff, given their role. Females are stereotyped as being caring/helpful [37], and thus, female VAs may be seen as more appropriate in SSCOs than male VAs, i.e., VAs by evoking stereotypes of helpfulness/approachability as deemed suitable for the role:

- Hypothesis 2: Participants will more often choose female VAs over male VAs.

Realism. Hypotheses are harder to derive in relation to realism because it is not possible to compare preferences of a "less realistic person" with preferences of a "more realistic person". It is also difficult to assess in relation to the two theories because a preference for the VA to be most like oneself or most like an actual employee would both result in a preference for high realism.

Unlike gaming worlds and computer-based learning environments, a VA in a SSCO may be more valued as task-oriented, as opposed to socially-oriented. This suggests a preference for realistic, three-dimensional VAs that are similar to users in terms of human-likeness and thus imply a similar level of intelligence and credibility. The reasoning is that while people are drawn to similar others, they do not need others to *be* or to represent themselves. Similarly, social role theory predicts a preference for realistic VAs because they are most like actual employees in human-likeness, implying more competence in their role. Based on both theories, we suggest:

- Hypothesis 3: Participants will more often choose more realistic VAs over cartoonized VAs.

Social role theory suggests that this may be especially true for males who value utilitarian aspects of human-computer interaction [6], making the potency of the VA role more important to this demographic of user. Baylor and Kim [39] found that males, but not females, learn more with realistic agents than cartoon agents. Similarly, Haake & Gulz [18] found that males, who preferred more detailed, three-dimensional agents over cartoon agents, also preferred more strictly task-oriented agents. Thus:

- Hypothesis 4: Male participants will show a greater preference for more realistic VAs than female participants.

2 Method

Setting. The impact of VA gender and realism was tested by asking participants to choose between VAs that differed on these aspects. Data were obtained from a SSCO simulation (Fig. 1) displayed at the Dundee Science Centre Exhibition titled "Robot: The Fantasy & the Reality" which was open to the general public. Amongst other attractions, such as robot models used in popular media and robots in science, visitors could interact with the touch-screen on a SSCO typically found in supermarkets. The exhibit stated that participant responses would be used for research and that individual data would be anonymous. After selecting a VA, users could then choose to engage in a produce-finding task. The preference data are the focus of this paper.

This method allowed us to collect data from a large participant sample, reducing the margin of sampling error. It may be argued that the sample of visitors to a science centre are not representative of the general population. However, a wide variety of individuals from the UK visited the Centre; the exhibition consisted of a variety of exhibits, from robots, to games and showcases of industry-specific technology. Those using the SSCO exhibit are likely to be self-selected – just as users of actual SSCOs are. It is also possible that some used the SSCO more than once or gave incorrect demographic information. The large sample size limits the effects of noise in the data. Also, there was no experimenter present while people interacted with the SSCO, reducing social response bias.

Fig. 1. The SSCO attraction in place at the Dundee Science Centre

Participants. Visitors attending the Dundee Science Centre in the first two weeks of June 2012 were the participant target group. Interested visitors could take part; none were specifically asked to do so. A total of 578 adults (220 male, 358 female) between the ages of 16 and 84 chose a VA (mean age 35.56 years).

Stimuli. Eight static VAs were presented in random order, varying in terms of gender, realism, and formality of dress (Fig. 2). Given the high number of permutations, randomization ensured that systematic effects of VA order were minimal. Formality of dress was an exploratory variable, yielding no significant impact on preference.

We collated the frequency of preferences for male versus female VAs, and the three-dimensional versus cartoonized VAs, and grouped these according to participant gender. Examples of the differences between VAs based on gender and realism can be seen in Fig. 2.

Procedure. After reading the start-up page providing information on the study, participants pressed the 'Begin' button, and eight VAs were displayed on the screen. Participants were asked to choose their preferred VA by touching its image (Fig. 2).

Participants were then asked to give demographic information, followed by the produce finding task. The frequency of VA choices was logged as the Dependent Variable.

Fig. 2. Eight VAs differing in gender, realism, and dress (exploratory)

Results - VA Gender. Addressing Hypotheses 1a, 1b, and 2, a 2x2 Chi-Square test revealed a relationship between the VA's implied gender and participant gender [$\chi2(1, N = 578) = 78.364, p < 0.001$] with a moderate association [$\Phi = 0.368$]. The majority (≈83%) of females chose a female VA, compared to nearly half (≈48%) of males (Table 1). Female VAs were preferred over male VAs with almost 70% of all participants choosing a female VA.

Table 1. The Count and Percentages of Male and Female Preferences for VA Gender

	Participant Female	**Participant Male**	**Total**
VA Female	296 (82.7%)	105 (47.7%)	401 (69.4%)
VA Male	62 (17.3%)	115 (52.3%)	177 (30.6%)
Total	358	220	578

Results - VA Realism. Addressing Hypotheses 3 and 4, a 2x2 Chi-Square test revealed a relationship between VA realism and participant gender [$\chi2(1, N = 578) = 8.093, p = 0.004$] with a weak association: $\Phi = 0.118$. The majority (≈63%) of females but more males (≈75%) chose three-dimensional, realistic options (Table 2).

Table 2. The Count and Percentages of Male and Female Preferences for VA Realism

	Participant Female	Participant Male	Total
Cartoonized VA	132 (36.9%)	56 (25.5%)	188 (32.5)
3-dimensional VA	226 (63.1)	164 (74.5%)	390 (67.5%)
Total	358	220	578

3 Discussion

Similarity-attraction theory predicts that participants would choose to gender-match. This was found for the majority of females, but not males, who chose male VAs almost as often as they chose female VAs, supporting Hypotheses 1a partially and 1b fully. Overall preference was for female VAs, but because this relies primarily on female participant data, the findings only partially support Hypothesis 2. The male gender-preference data were in line with neither similarity-attraction theory, nor social role theory. The findings also support Hypothesis 3, as most participants chose three-dimensional VAs, reinforcing the idea that VAs in a SSCO context serve a task-based function, with slightly more men than women preferring this level of realism (support for Hypothesis 4), lending support for social role theory.

While neither theory was fully supported, they provide insights and reasonable interpretations of the findings. Similarity-attraction theory suggests that identification with a social group enhances co-operative behavior, commitment, knowledge contribution, participation, and organizational citizenship behavior [7]. Females, having a stronger collective bond with other females than males have with other males [37], might have chosen to gender-match because it increases comfort during the interaction. There may also be positive performance-based outcomes for female users interacting with a same sex VA, as evidence suggests that women become more task-oriented when interacting in same-sex groups than in mixed-sex groups [33-34]. In contrast, males often like interacting with the opposite sex, due to positive attitudes towards maternal attributes [37]. Despite this, our results show no male preference for a female VA. The findings reflect a less pronounced within-sex bias in men, consistent with other research in human-human interaction [37].

The findings are also consistent with research suggesting that males are more concerned with utilitarian aspects of human-computer interaction [6], [39], since a higher percentage of males than females chose a realistic agent. Nevertheless, the majority of both males and females preferred three-dimensional, realistic VAs. Thus, both sexes seemed more concerned with utilitarian aspects of the SSCO simulation, and less concerned with hedonic aspects. While Haake and Gulz [18] found no preference for realism in males, they did find that males who preferred more detailed, three-dimensional agents over cartoon agents also preferred more strictly task-oriented agents – highlighting the link between realism and utilitarian goals.

More females preferred realistic VAs than cartoonized Vas, suggesting that human-likeness (similarity-attraction) and/or assistance-based role (social role) of VAs were more important to users over other factors, such as self-representation or enjoyment. The two possible explanations based on the two theories are not necessarily in opposition. First, in line with similarity-attraction theory, participants of both sexes may have been driven by the task-oriented nature of SSCO, resulting in a preference for VAs that were most like themselves in terms of realism. Second, in line with social role theory, participants may have been more concerned with the VA's role over its attractiveness, thus, choosing VAs that were closer visual representations of actual supermarket employees, i.e., three-dimensional. In SSCO, this involves displaying 'appropriate' checkout assistant characteristics such as warmth and approachability – traits which are more readily perceived in human-like females [36].

There are some limitations associated with this study. First, perceptions of the VA were not measured. These perceptions, and other possible reasons for choosing VAs, could be tackled in future studies, perhaps with more qualitative research. Second, there may have been other factors that made three-dimensional VAs more attractive, such as familiarity with realistic characters in computer games or the association between two-dimensional cartoons with children's TV programs. Realistic VAs could also have been more attractive because they were more (but not too) human-like, thus avoiding the "Uncanny Valley", which proposes that more human-like characters are preferred to less human-like characters, until they become so human-like that they are eerie [14], [16]. Finally, the context is less ecologically valid than an actual supermarket setting, but arguably more ecologically valid than a lab-based study.

Conclusions. Females chose to gender-match and to interact with a more realistic VA. Males exhibited little preference for either gender, and a greater preference than females for realistic VAs. Thus, where it is not feasible to gender-match in SSCO, the recommendation is to implement a realistic female VA.

Our findings raise the question of whether VA *choice* would be a viable option to engage consumers in SSCO, as this would go beyond mere acceptance of an existing agent, thus potentially alleviating the perpetuation of stereotypical assumptions.

Acknowledgments. We thank the Dundee Science Centre for exhibiting our SSCO. Thanks go to all participants for taking part, Rosie Henderson (Abertay University) for the creation of the VAs and Stuart Birse (NCR) for providing the pictures of the SSCO set-up. Finally, we thank NCR for funding this research.

References

1. Rust, R.T., Espinoza, F.: How Technology Advances Influence Business Research and Marketing Strategy. J. Bus. Res. 59, 1072–1078 (2006)
2. Simon, F., Usunier, J.-C.: Cognitive, Demographic, and Situational Determinants of Service Customer Preferences for Personnel-in-contact over Self-service Technology. International J. Res. in Market. 24, 163–173 (2007)

3. Meuter, M.L., Bitner, M.J., Ostrom, A.L., Brown, S.W.: Choosing among Alternative Service Delivery Modes: An Investigation of Customer Trial of Self-service Technologies. J. Market. 69, 61–83 (2005)
4. Lee, H.-J., Fairhurst, A., Cho, H.J.: Gender Differences in Consumer Evaluations of Service Quality: Self-Service Kiosks in Retail. The Serv. Industries J. 33, 248–265 (2013)
5. Abeele, M.V., Roe, K., Pandelaere, M.: Construct Validation of the Concepts Social Presence, Emotional Presence and Connectedness and an Application of Zajonc's Social Facilitation Theory to Social Presence Research. In: Proceedings of the 10th Annual International Workshop on Presence, pp. 215–224 (2007)
6. Qiu, L., Benbasat, I.: A study of Demographic Embodiment of Product Recommendation Agents in Electronic Commerce. Int. J. Hum.-Comp. Stud. 68, 669–688 (2010)
7. Shen, K.N., Yu, A.Y., Khalifa, M.: Knowledge Contribution in Virtual Communities: Accounting for Multiple Dimensions of Social Presence through Social Identity. Beh. Inf. Tech. 29, 337–348 (2009)
8. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. CSLI Publications, Cambridge (1996)
9. Payne, J.A., Johnson, G.I., Szymkowiak, A.: The Behavioural Impact of a Visually Represented Virtual Assistant in a Self-Service Checkout Context. In: Proceedings of the 25th BCS Conference on Human-Computer Interaction, pp. 58–63. British Computer Society (2011)
10. Heckman, C.E., Wobbrock, J.O.: Put Your Best Face Forward: Anthropomorphic Agents, E-Commerce Consumers, and the Law. In: Proceedings of the ACM Conference on Autonomous Agents, pp. 435–442 (2000)
11. McBreen, H., Anderson, J., Jack, M.: Evaluating 3D Embodied Conversational Agents in Contrasting VRML Retail Applications. In: Proceedings of International Conference of Autonomous Agents Workshop on Multimodal Communication and Context in Embodied Agents, pp. 83–87 (2001)
12. McBreen, H.M., Jack, M.A.: Evaluating Humanoid Synthetic Agents in E-Retail Applications. IEEE Transactions on Systems, Man, and Cybernetics 31, 394–405 (2001)
13. Brave, S., Nass, C., Hutchinson, K.: Computers That Care: Investigating the Effects of Orientation of Emotion Exhibited by an Embodied Computer Agent. Int. J. Hum.-Comp. Stud. 62, 161–178 (2005)
14. Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J.K., Robles, E.: Evaluating the Effects of Behavioral Realism in Embodied Agents. Int. J. Hum.-Comp. Stud. 67, 842–849 (2009)
15. Verhagen, T., van Nes, J., Feldberg, F., van Dolen, W.: Virtual customer service agents: Using social presence and personalization to shape online service encounters. Res. Memorand. 10 (2011)
16. MacDorman, K.F., Green, R.D., Ho, C.-C., Koch, C.T.: Too Real for Comfort? Uncanny Responses to Computer Generated Faces. Comp. Hum. Beh. 25, 695–710 (2009)
17. Cowell, A.J., Stanney, K.M.: Manipulation of Non-verbal Interaction Style and Demographic Embodiment to Increase Anthropomorphic Computer Character Credibility. Int. J. Hum.-Comp. Stud. 6, 281–306 (2005)
18. Haake, M., Gulz, A.: A Look at the Roles of Looks & Roles in Embodied Pedagogical Agents – A User Preference Perspective. Int. J. Art. Int. Ed. 19, 39–71 (2009)
19. Sahimi, A.M., Zain, F.M., Kamar, N.A.N., Samar, N., Rahman, Z.A., Majid, O., Atan, H., Fook, F.S.: The Pedagogical Agent in Online Learning: Effects of the Degree of Realism on Achievement in Terms of Gender. Contemp. Educ. Technol. 1, 175–185 (2010)

20. Veletsianos, G.: Contextually Relevant Pedagogical Agents: Visual Appearance, Stereotypes, and First Impressions and Their Impact on Learning. Comp. & Educ. 55, 576–585 (2010)
21. Baylor, A.L.: The Design of Motivational Agents and Avatars. Educ. Tech. Res. & Develop. 59, 291–300 (2011)
22. Kim, Y., Wei, Q.: The Impact of Learner Attributes and Learner Choice in an Agent-based Environment. Comp. Educ. 56, 505–514 (2011)
23. Lee, E.-J.: The Effects of “Gender” of the Computer on Informational Social Influence: The Moderating Role of Task Type. Int. J. Hum.-Comp. Stud. 58, 347–362 (2003)
24. Berry, D.C., Butler, L.T., de Rosis, F.: Evaluating a Realistic Agent in an Advice-giving Task. Int. J. Hum.-Comp. Stud. 63, 304–327 (2005)
25. Gong, L.: Is Happy Better than Sad even if They Are Both Non-Adaptive? Effects of Emotional Expressions of Talking-Head Interface Agents. Int. J. Hum.-Comp. Stud. 65, 183–191 (2007)
26. Pratt, J.A., Hauser, K., Ugray, Z., Patterson, O.: Looking at Human-computer Interface Design: Effects of Ethnicity in Computer Agents. Inter. Comp. 19, 512–523 (2007)
27. Van Vugt, H.V., Konijn, E.A., Hoorn, J.F., Keur, I., Eliëns, A.: Realism is Not All! User Engagement with Task-Related Interface Characters. Inter. Comp. 19, 267–280 (2007)
28. Baylor, A.L., Plant, E.A.: Pedagogical Agents as Social Models for Engineering: The Influence of Agent Appearance on Female Choice. Art. Intell. Ed. 125, 65–72 (2005)
29. Baylor, A.L.: Promoting Motivation with Virtual Agents and Avatars: Role of Visual Presence and Appearance. Phil. Trans. of the Royal Soc. B: Biol. Sciences 364, 3559–3565 (2009)
30. Rosenberg-Kima, R.B., Baylor, A.L., Plant, E.A., Doerr, C.E.: Interface Agents as Social Models for Female Students: The Effects of Agent Visual Presence and Appearance on Female Students’ Attitudes and Beliefs. Comp. Hum. Beh. 24, 2741–2756 (2008)
31. Bailenson, J.N., Yee, N.: Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. Psych. Sc. 16, 814–819 (2005)
32. Ducheneaut, N., Wen, M.-H., Yee, N., Wadley, G.: Body and Mind: A Study of Avatar Personalization in Three Virtual Worlds. In: Proceedings of CHI (2009)
33. Li, I., Forlizzi, J., Dey, A., Kiesler, S.: When the Interface is the User’s Face: Ideas for Research And Applications. In: CHI Workshop in HCI and the Face (2006)
34. Cross, S.E., Madson, L.: Models of the Self: Self-construals and Gender. Psych. Bull. 122, 5–37 (1997)
35. Eagly, A.H., Wood, W., Diekman, A.B.: Social Role Theory of Sex Difference and Similarities: A Current Appraisal. In: Eckes, T., Trautner, H.M. (eds.) Developmental Social Psychology of Gender. Lawrence Erlbaum Associates, Inc., New Jersey (2000)
36. Cuddy, A.J.C., Fiske, S.T., Glick, P.: The BIAS map: Behaviors from Intergroup Affect and Stereotypes. J. Pers. Soc. Psych. 92, 631–648 (2007)
37. Rudman, L.A., Goodwin, S.A.: Gender Differences in Automatic In-Group Bias: Why Do Women Like Women More than Men Like Men? J. Pers. Soc. Psych. 87, 494–509 (2004)
38. Vigil, J.M.: A socio-relational Framework of Sex Differences in the Expression of Emotion. Beh. Brain Sc. 32, 375–390 (2009)
39. Baylor, A.L., Kim, Y.: Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds.) ITS 2004. LNCS, vol. 3220, pp. 592–603. Springer, Heidelberg (2004)

Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training

Ligia Batrinca[1], Giota Stratou[2], Ari Shapiro[2],
Louis-Philippe Morency[2], and Stefan Scherer[2]

[1] FBK-IRST, Via Sommarive 18, 38050 Trento, Italy
batrinca@fbk.eu
[2] University of Southern California
Institute for Creative Technologies, Los Angeles, California
{stratou,shapiro,morency,scherer}@ict.usc.edu

Abstract. Public speaking performances are not only characterized by the presentation of the content, but also by the presenters' nonverbal behavior, such as gestures, tone of voice, vocal variety, and facial expressions. Within this work, we seek to identify automatic nonverbal behavior descriptors that correlate with expert-assessments of behaviors characteristic of good and bad public speaking performances. We present a novel multimodal corpus recorded with a virtual audience public speaking training platform. Lastly, we utilize the behavior descriptors to automatically approximate the overall assessment of the performance using support vector regression in a speaker-independent experiment and yield promising results approaching human performance.

Keywords: Virtual Reality, Behavioral Modification, Multimodal Perception, Public Speaking, Training.

1 Introduction

Public speaking is an essential skill for a large variety of professions and in everyday life. The quality of a presentation can greatly influences the presenters career development or the likelihood to close a deal. However, public speaking itself is not a skill that is innate to everyone, but can be mastered through extensive training[1]. Further, mild forms of public speaking anxiety may be controlled via frequent exposure to presentation scenarios (even virtual ones) [10,11]. The best form of training often is to present in familiar and forgiving environments and by receiving the audiences feedback during and after the presentation. Audiences provide indirect feedback during presentations by signaling nonverbal feedback, while they continuously rate and sense the presenters speaking style. While audiences show signs of high attention (e.g. mutual gaze or forward leaning posture) and cues of rapport (e.g. nodding or smiling) in presentations they enjoy, they often show no interest (e.g. averted gaze or lack of backchannel behavior) or

[1] http://www.toastmasters.org/tips.asp

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 116–128, 2013.

disagreement otherwise. By consciously perceiving and adapting ones speaking style with respect to these feedback behaviors the presenter can greatly improve.

Unfortunately, often no human audience is available or the fear of presenting in front of a human audience is difficult. The present work provides preliminary steps towards an artificial virtual audience capable of providing such nonverbal feedback during presentations, based on the perceived multimodal presenter's speaking style. Such a virtual audience would be available at any time and could help improve public speaking skills in an efficient and non-threatening way. Within this paper, we present our prototype virtual human platform for public speaking training, called Cicero. In the future, Cicero will serve as such a virtual audience that will provide users with helpful feedback.

We analyze a preliminary dataset of 14 subjects giving a 5-15 minute presentation in front of a virtual audience. Public speaking experts from the world-wide organization of Toastmasters, assess both public speaking related behaviors and observations and estimate the presenters' overall performance in a viewing study. We then correlate automatically observed multimodal nonverbal behaviors with expert assessments of the assessed behaviors and try to automatically approximate the experts' overall assessment of the presenters' performance in a speaker-independent regression task. In particular, our research goals for the present work are:

- **R1 Expert Assessment:** We aim to identify expert estimates of nonverbal behaviors, including flow of speech, clarity of intonation, correct use of gestures, and gaze patterns, that correlate with the experts' overall assessment of the presenter's performance.
- **R2 Automatic Behavior Descriptors:** We seek to identify basic automatic multimodal behavior descriptors that strongly correlate with the experts' assessment of the presenters' audiovisual nonverbal behavior. These automatic measures are extracted from three independent sensors and comprise basic estimates of speech characteristics, gestures, and gaze.
- **R3 Automatic Performance Assessment:** We further estimate the presenters' performance in a preliminary presenter-independent classification experiment using the automatically estimated nonverbal behaviors as input for the support vector regression.

The remainder of this paper is organized as follows: Section 2 discusses some related work on virtual audiences and speaking performance. Section 3 then introduces our experimental setup, the investigated dataset. In Section 4, we discuss the details of the expert assessment study. The automatic behavior descriptors using audiovisual information are then introduced and their correlations with expert opinions are studied in Section 5. In Section 5.4, we investigate how well the automatic behavior descriptors can be used to approximate the expert assessments. In Section 6, we discuss our findings and outline future paths. Finally, Section 7 concludes the paper.

2 Related Work

Virtual Audiences. Virtual audiences have been investigated in the past to treat public speaking anxiety. One of the first works on virtual reality (VR) used to treat public speaking anxiety was done by [10]. This study suggested that VR could indeed be useful in treating public speaking anxiety. At the end of the study, self-reported levels of anxiety were reduced. In a study by [11] participants were asked to give a 5 minutes long presentation to 3 different types of audiences: a neutral, a positive, and a negative audience. The virtual audience consisted of 8 virtual characters. The study showed that all three settings mentioned above, have an influence on the subject, generating anxiety in participants which scored high on the Personal Report of Confidence as a Public Speaker (PRCS). In the same year, another study, by [3], focused on university students with prominent public speaking anxiety. One group was exposed to Virtual reality Exposure Therapy (VRET) while another group were put in a wait-list control group. The results of this study are in line with previous findings: virtual reality treatment sessions are effective in reducing public speaking anxiety.

Although lately, more researchers have become aware of the importance and effectiveness of VR in treating anxiety-like behaviors when holding a talk in front of an audience, to the best of our knowledge this is the first study to directly address the scenario of giving a presentation in front of a virtual audience. Moreover, using non-invasive, state-of-the-art sensing technology capturing the presenters' nonverbal behavior patterns. Additionally, in this work we primarily focus on the quality of the performance itself rather than investigating possible treatment strategies of public speaking anxiety.

Public Speaking Performance. In general, excellent and persuasive public speaking performances, such as giving a presentation in front of an audience, are not only characterized by decisive arguments or a well structured train of thoughts, but also by the nonverbal characteristics of the presenter's performance, i.e. the facial expressions, gaze patterns, gestures, and acoustic characteristics. This has been investigated by several researchers in the past using political speakers' performances. In [17] for example researchers found that vocal variety, as measured by fundamental frequency (f_0) range and maximal f_0 of focused words are correlated with perceptual ratings of a good speaker within a dataset of Swedish parliamentarians. Further, manual annotations of disfluencies were identified to be negatively correlated with a positive rating.

In [13], the acoustic feature set, used in [17], was complemented by measures of pause timings and measures of tense voice qualities. The study shows that tense voice quality and reduced pause timings were correlated with overall good speaking performances. Further, the authors investigated visual cues, in particular motion energy, for the assessment of the speakers' performances. They found that motion energy is positively correlated with a positive perception of speakers. This effect is increased when only visual cues are presented to the raters.

The authors of [8] investigate more complex motion features, such as hand trajectories and identify correlates of gestures with ratings of personality. Again, they extract these motion features from videos of German politicians and present them as stick-figure representations to the raters. In all three above studies, non-experts assessed the performances of professional speakers (i.e. politicians), within this work we want to investigate features that are present in presentation performances by the general population and potentially untrained speakers. We further ask experts to rate their performances not only with respect to an overall assessment, but by utilizing a more fine grained questionnaire that disseminates the behaviors into multiple ratings.

3 Experimental Design and Dataset

In the following we provide details regarding the experimental setup in which the study took place and some details on how we setup the virtual audience for the experiments. Additionally, we detail the participant recruitment and the experimental procedure.

3.1 Experimental Setting

Figure 1 illustrates the room setup used in the study. As it can be seen, the lab was arranged to resemble a conference room. The experimenter initializes the virtual audience that is projected on the virtual audience screen. The characters approach life-size measures. The participant controls the presentation, which is projected on the presentation screen, with the help of a standard Logitech remote control. The nonverbal behaviors of the participant are captured using a Microsoft Kinect and 2 off-the-shelf webcams, mounted in front of the participant and to the side (cf. Figure 1). Acoustic information was collected using a lapel microphone.

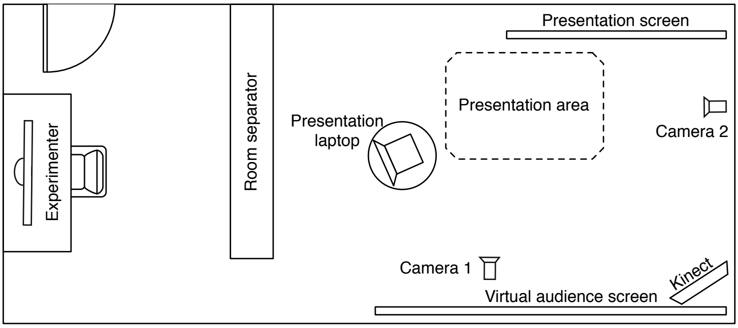


Fig. 1. Experimental setting of the Virtual Human Lab

3.2 The Virtual Audience

A high-performance desktop computer was used to project the virtual audience, animated using SmartBody character animation system [16] and VHToolkit [7]. In Figure 2 a snapshot of the virtual audience is provided. Each virtual human was able to change its posture (e.g. straight, relaxed, forward), head orientation (e.g. up, left, right, front) or eye-gaze. Eye-contact with the speaker was simulated by having the characters looking at the speaker. This enabled us to create an audience designed to give the impression of a real life audience. Within our study the virtual audience was modeled to display interest in the participant's presentation, which was accomplished by the proper combination of posture, head-orientation and eye-gaze directionality. It is important to mention that during the study, the attentive behavior of the virtual characters was not static and the audience remained attentive and lively.

Fig. 2. Snapshot of the virtual audience

3.3 Participants and Experimental Procedure

In our study we recorded 14 participants, 13 of which were recruited from craigslist and one participant was recruited at a university. The data set consisted of 11 females and 3 males, with an average age of 39 and standard deviation of 14.34. The participants were provided with two different presentations to choose from, three days in advance of the scheduled presentation. They were instructed to rehearse, before giving the presentation at our lab, as they would normally prepare for an important presentation.

Prior to the presentation, the participants filled in a series of questionnaires including a brief demographic assessment, the Big-Five Inventory short form (BFI SF)[12], the Personal Report of Confidence as a Public Speaker (PRCS)[9], and the Self-Statements During Public Speaking (SSPS) [4]. Immediately before the start of the experiment, the participants were introduced to the experimental setting of the lab. They were instructed not to look directly into the cameras, but at the audience. Markers on the floor were provided to give a guidance to the presenters where to stand while giving the presentation, to ensure optimal viewing angles for the cameras.

Participants then filled in post-session questionnaires, including Positive and Negative Affect Schedule (PANAS) [19], and a modified version of [21] virtual audience presence questionnaire and performance self-assessment questionnaires. At the end the participants were debriefed and received 25 USD for their efforts.

4 Expert Assessment

In order to obtain independent expert opinions on the participants' performances, we invited two senior Toastmasters, i.e. members of a worldwide organization devoted to improve speaking skills through exercise and critique. Both experts assessed their own experience level with the highest possible value on a seven point Likert scale. Additionally, both feel very comfortable presenting themselves and performed 10 or more times in the last two years. Lastly, they estimate their own public speaking skill to be clearly above average.

The experts viewed the presentations using the frontal camera view with the audio from the lapel microphone. They viewed each presentation only once and assessed the performances of the participants using two sets of questions, all of which were answered on a seven point Likert scale. The first set consists of typical behaviors and observable characteristics of public speaking performances and comprise assessments of the flow of speech, the presenter's pacing behavior

Table 1. Summarizes the correlations between the expert assessed behaviors from three sources (i.e. voice, body, and gaze) with the experts' opinion of the presenters' overall performance. Spearman's ρ values are reported, along with p-values of test if estimated correlation is significantly different from no correlation.

Source	Assessed behavior	Spearman's ρ	p-value
Voice	Flow of speech	**.477**	.010
	Clear intonation	**.436**	.021
	Interrupted speech	.016	.933
	Speaks too quietly	-.363	.057
	Vocal variety	**.471**	.013
Body	Paces too much	**.599**	$< .001$
	Gestures to emphasize	.354	.065
	Gestures to much	-.062	.764
Gaze	Gazes at audience	.166	.398
	Avoids audience	-.358	.062

on the stage, the posture's stiffness, the presenter's nervousness, and the observed amount of eye contact with the virtual audience. In total we assessed 21 characteristics for each speaker, a subset of these are presented in Table 1. Additionally, we assess the experts' perception of the overall performance. All expert annotations are z-score normalized in order to remove perception biases. The inter-expert agreement on the overall assessment results in Krippendorff α = .715, which corresponds to considerable agreement. The correlation between the overall rating for the performances is with Spearman's $\rho = .648$ quite strong and significantly different from zero with p = .012.

Table 1 summarizes the estimated correlations of some of the assessed behaviors and characteristics with the overall estimated presenters' performance.

5 Automatic Behavior Descriptors

Similarly to the above evaluation of correlations between assessed behaviors and the overall performance, we investigate automatic behavior descriptor-correlations with the expert-assessed behaviors and characteristics listed in Table 1. The behavior descriptors are automatically extracted from three audiovisual sensory inputs using the multimodal sensor fusion framework called *MultiSense* [15,14]. MultiSense is a flexible framework that is based on the Social Signal Interpretation framework (SSI) by [20] and it is created as a platform to integrate and fuse sensor technologies and develop probabilistic models for human behavior recognition. The modular setup of MultiSense allows us to integrate multiple sensing technologies for this analysis. We detail the extracted behavior descriptors in sections 5.1 and 5.2. The results of the correlation analysis are reported in Section 5.3 and Table 2. The automatic overall performance assessment evaluation is provided in Section 5.4.

5.1 Acoustic Nonverbal Behavior Descriptors

Using the lapel microphone recordings, we extracted several basic acoustic and prosodic features. The features are extracted with a sample rate of 100 Hz. Hesitations and pause fillers were counted by one of the experts and noted on the evaluation sheet for each presenter. The following sections detail each acoustic feature.

Energy in dB. The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms (i.e. 100Hz sample rate). This speech window $w(t)$ is filtered with a hamming window and the energy

$$e(t) = \sum_{i=1}^{|w(t)|} w_i(t)^2 \tag{1}$$

is calculated and converted to the dB-scale

$$e_{dB}(t) = 10 \cdot \log_{10}(e(t)). \tag{2}$$

Table 2. Summarizes the correlations between the expert assessed behaviors from three sources (i.e. voice, skeleton, and gaze) with automatic behavior descriptors extracted from the audiovisual data. Spearman's ρ values are reported, along with p-values of test if estimated correlation is significantly different from no correlation.

Source	Assessed behavior	Behavior descriptor	Spearman's ρ	p-value
Voice	Flow of speech	Num. pauses	-.469	.09
	Clear intonation	Avg. intensity	**.805**	.002
		Breathiness	**-.615**	.033
	Interrupted speech	Num. pause fillers	**.612**	.034
	Speaks too quietly	Avg. intensity	**-.842**	$< .001$
	Vocal variety	Std. f_0	**.709**	.010
		Spectral Stationarity	**-.586**	.045
Body	Paces too much	Leg movement	**.682**	.021
	Gestures to emphasize	Arm movement	**.710**	.014
	Gestures to much	Arm movement	.437	.179
Gaze	Gazes at audience	Face gaze towards	**.621**	.030
	Avoids audience	Face gaze towards	-.548	.065

Fundamental Frequency f_0. In [2], a method for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions, is introduced. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process removes strong influences of noise and vocal tract resonances. For each $r(t)$ the amplitude spectrum $E(f)$ is computed, showing peaks for the harmonics of f_0, the fundamental frequency. Then, the summation of residual harmonics (SRH) is computed as follows [2]:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)], \quad (3)$$

for $f \in [f_{0,\mathrm{min}}, f_{0,\mathrm{max}}]$, with $f_{0,min} = 50$ and $f_{0,max} = 300$. The frequency f for which $SRH(f)$ is maximal is considered the fundamental frequency of this frame. By using a simple threshold θ, the unvoiced frames are discarded as in [2].

Pause Timings. Pauses were considered as continuous segments of at least 300 ms in length with a signal strength of at least 25 dB below the 99th percentile of the recording. This implementation follows the same parameter setting and recommendations as in the standard Praat pause detection algorithm [1].

Spectral Stationarity. To characterize the range of the prosodic inventory used over utterances, we make use of the so called *spectral stationarity* measure

ss. This measurement was previously used in [18] as a way of modulating the transition cost used in the dynamic programming method used for f_0 tracking. Spectral stationarity, ss is measured with:

$$ss = \frac{0.2}{\mathrm{itakura}(f_i, f_{i-k}) - 0.8} \quad \in [0, 1], \tag{4}$$

where itakura(.) is the Itakura distortion measure [5] of the current speech frame f_i and f_{i-k} is the previous frame with $k = 1$. We use a relatively long frame length of 60 ms (with as shift of 10 ms; sampling rate 100Hz) and frames are windowed with a Hamming window function before measuring ss. The long frame length was used in the attempt to characterize relatively long periods of maintained vocal tract articulation. ss is close to 1 when the spectral characteristics of adjacent frames are very similar and goes closer to 0 if the frames show a high degree of difference.

Voice Tenseness Measured by $\mathbf{OQ}_{NN}$. In order to characterize the tenseness of the speaker's voice, we extract $\mathbf{OQ}_{NN}$ a novel parameter estimating the open quotient using standard Mel frequency cepstral coefficients and a trained neural network for open quotient approximation [6].

5.2 Visual Nonverbal Behavior Descriptors

Visual behavior descriptors were extracted from the tracked skeleton and face using information provided by the Kinect sensor and the frontal web-camera. Measures were extracted using a sample rate of 30 Hz. The following sections detail each visual feature.

Arm and Leg Movement. Based on the tracked skeletal information we calculate an overall intensity measure of the arm and leg movement respectively. We calculate movement by computing simple distances between consecutive frames of the tracked skeletal joints. These distances are summed up for the respective group (i.e. legs and arms) and normalized by the total length of the presentation.

Face Gaze Towards. We utilize the tracked face direction to assess the presenters' gaze. We track if the presenter looks towards the screen on which we present the virtual audience using the frontal webcam placed on a tripod (100 cm in height) facing the presenter. We consider a relatively wide range of degrees as facing towards the audience (i.e. +/- 45 degrees) as the audience is fairly large and close to the presenter. Additionally, we track the vertical gaze direction and consider angles above zero degrees as gazing towards the audience. Angles below zero are considered as looking at hand-held notes or the floor. We measure the gaze towards the audience as a ratio of the overall total duration of the presentation.

5.3 Automatic Behavior Descriptor Correlations

Here, we report results of the correlation analysis between automatic behavior descriptors and expert assessments of presenters' characteristics. We calculate Spearman's ρ for each behavior descriptor with the associated expert-assessed behavior and report p-values indicating if the observed correlation is significantly different from zero.

As seen in Table 2, we could identify a large number of basic behavior descriptors that correlate significantly with expert assessments for all investigated modalities. Based on the voice descriptors we could identify that the average speech intensity is highly correlated with the "clear intonation" assessments ($\rho = .805$; $p = .002$) and negatively correlated with the "speaks too quietly" assessment ($\rho = -.842$; $p < .001$). Further, the breathiness as observed with higher values of OQ_{NN} is negatively correlated with clear intonation ($\rho = -.615$; $p = .033$). Vocal variety is correlated with the standard deviation of f_0 ($\rho = .709$; $p = .010$) and negatively correlated with the monotonicity measure spectral stationarity ($\rho = -.586$; $p = .045$).

Based on the skeletal information, we can identify if the presenter is pacing too much on stage by using the leg movement descriptor ($\rho = .682$; $p = .021$). Additionally, arm movement is correlated with the experts' assessment if the presenter uses gestures to emphasize points of the presentation appropriately ($\rho = .710$; $p = .014$). Lastly, the "gazes at audience" assessment is correlated with the automatic behavior descriptor face gaze towards ($\rho = .621$; $p = .030$).

We are aware of the fact that a statistical correction for multiple testing would be required at this point. However, with the relatively small sample size this would require extremely high correlations $|\rho| \geq .800$ for each individual behavior. In order to address this issue from another direction, we chose to conduct a sanity-check regression analysis with a leave-one-presenter-out testing paradigm to show the relevance of the observed nonverbal behaviors in the following section.

5.4 Automatic Performance Assessment

Based on the above findings and automatic behavior descriptors, we conduct a presenter-independent approximation experiment. We use simple support vector regression with a polynomial kernel of degree three and in total eight features as input (i.e. five voice features, two skeleton features, and one gaze feature). With a leave one presenter out testing paradigm we achieve an overall absolute mean error of .660 with a standard deviation of .540. The automatically approximated performance assessment corresponds with the experts' mean overall assessment with $\rho = .617$ and $p = .025$.

6 Discussion

Based on expert assessments of a small number of presentations given to a virtual audience, we could identify several characteristic nonverbal behaviors that correlate positively or negatively with the overall perceived presenters' performance.

All three investigated modalities (i.e. voice, skeleton, and gaze) contribute to the assessment and Table 1 summarizes these findings. It is interesting to note, that some behaviors that anecdotally are associated with bad performances did not show any correlation with the overall assessment, such as interrupted speech or excessive gesturing. We believe that these behaviors might be outweighed by others and a more fine-grained overall performance estimation disseminating spoken, gestural, expressive, and structural quality might be required.

When approximating the expert-assessed nonverbal behaviors automatically, we could identify a number of basic behavior descriptors, such as the average intensity, overall leg movement, and gaze statistics, that are highly correlated with expert assessments (cf. Table 2). While, these basic descriptors achieved promising results using support vector regression in a speaker-independent experiment (cf. Section 5.4), they remain crude and on an abstract level. For example, the overall arm movement is correlated with appropriate gestural emphasis of arguments within a presentation, which would intuitively at least require knowledge about the arm gestures and the content of the spoken words. Hence, we plan to investigate multimodal information fusion to capture more meaningful and sophisticated measures of public speaking performances.

For future work, we additionally plan to investigate optimal ways of conveying the perceived information on the performance to the presenters. We will analyze ad-hoc visualizations, such as audience reactions or visual overlays, as well as post-hoc summaries and quantitative evaluations with typical statistical plots. We plan to base the audience's behavior on the presenter's automatically estimated performance to provide realtime feedback to the presenter. Here, we envision both subtle movements in the audience to create a more life-like and immersive experience for the presenter and more striking and interruptive behaviors to directly reflect the potential discontent or approval of the presenter's performance. The audience could for example show reduced interest in the presentation due to the lack of vocal variety in the presenter's voice. At present, we focused our analysis on nonverbal behaviors only and will expand the analysis to verbal contents in the future. Further, we will investigate usability and effectivity of different strategies, with respect to performance improvement and immersion.

7 Conclusions

This paper presents a proof-of-concept (and at present non-interactive) version of the research platform for public speaking training, called Cicero. Based on our research goals, stated in Section 1, we could identify the following main findings in this work: **R1** we reveal several expert estimates of nonverbal behaviors, such as flow of speech, vocal variety, or avoided eye contact with the audience, to be significantly correlated with an overall assessment of a presenter's performance; **R2** using multimodal information from three sensors we could identify automatic behavior descriptors that correlate strongly with expert estimates of nonverbal behaviors, comprising estimates for a clear intonation, vocal variety, pacing around, and eye contact with the audience. Lastly, **R3** we automatically approximate the experts' overall performance assessment with a mean error of .660 on a seven point scale. Further, the automatic approximation using support vector regression correlates significantly with the experts' opinion with Spearman's $\rho = .617$ (p = .025), which approaches the correlation between the experts' opinions (i.e. $\rho = .648$). Motivated by these promising results, we plan to expand the presented research platform Cicero in the near future to incorporate a more diverse and reactive virtual audience. Cicero will enable us to conduct a wide variety of experiments reaching from performance assessments to psychological experiments, which would not be possible with a real human audience.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 1118018 and the U.S. Army RDECOM. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We would like to thank our close collaborators Yuyu Xu, Teresa Dey, Margaux Lhommet, and Stacy Marsella for their help and advice while running this study. Additionally, we would like to thank the Toastmasters group Funny Bones of Culver City for their kind support in assessing the performance of our study participants.

References

1. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9), 341–345 (2001)
2. Drugman, T., Abeer, A.: Joint robust voicing detection and pitch estimation based on residual harmonics. In: Proceedings of Interspeech 2011, pp. 1973–1976. ISCA (2011)
3. Harris, S.R., Kemmerling, R.L., North, M.M.: Brief virtual reality therapy for public speaking anxiety. Cyberpsychology and Behavior 5, 543–550 (2002)
4. Hofmann, S.G., DiBartolo, P.M.: An instrument to assess self-statements during public speaking: Scale development and preliminary psychometric properties. Journal of Behavior Therapy, 499–515 (2000)
5. Itakura, F.: Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-23, 67–72 (1975)
6. Kane, J., Scherer, S., Morency, L.-P., Gobl, C.: A comparative study of glottal open quotient estimation techniques. To Appear in Proceedings of Interspeech 2013. ISCA (2013)
7. Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D.: Building interactive virtual humans for training environments. In: Proceedings of I/ITSEC (2007)
8. Koppensteiner, M., Grammer, K.: Motion patterns in political speech and their influence on personality ratings. Journal of Research in Personality 44, 374–379 (2010)
9. McCroskey, J.C.: Measures of communication-bound anxiety. Speech Monographs 37, 269–277 (1970)
10. North, M.M., North, S.M., Coble, J.R.: Virtual reality therapy: An effective treatment for the fear of public speaking. International Journal of Virtual Reality 3, 2–6 (1998)

11. Pertaub, D.P., Slater, M., Barker, C.: An experiment on public speaking anxiety in response to three different types of virtual audience. Presence: Teleoperators and Virtual Environments 11, 68–78 (2002)
12. Rammstedt, B., John, O.P.: Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. Journal of Research in Personality 41, 203–212 (2007)
13. Scherer, S., Layher, G., Kane, J., Neumann, H., Campbell, N.: An audiovisual political speech analysis incorporating eye-tracking and perception data. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 1114–1120. ELRA (2012)
14. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A(S.), Morency, L.-P.: Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS (LNAI), vol. 7502, pp. 455–463. Springer, Heidelberg (2012)
15. Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., Morency, L.-P.: Automatic behavior descriptors for psychological disorder analysis. In: Proceedings of IEEE Conference on Automatic Face and Gesture Recognition. IEEE (2013)
16. Shapiro, A.: Building a character animation system. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 98–109. Springer, Heidelberg (2011)
17. Strangert, E., Gustafson, J.: What makes a good speaker? Subject ratings, acoustic measurements and perceptual evaluations. In: Proceedings of Interspeech 2008, pp. 1688–1691. ISCA (2008)
18. Talkin, D.: A Robust Algorithm for Pitch Tracking. In: Kleijn, W.B., Paliwal, K.K. (eds.) Speech Coding and Synthesis, pp. 495–517. Elsevier (1995)
19. Thompson, E.R.: Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). Journal of Cross-Cultural Psychology 38(2), 227–242 (2007)
20. Wagner, J., Lingenfelser, F., Bee, N., André, E.: Social signal interpretation (ssi). In: KI - Kuenstliche Intelligenz, vol. 25, pp. 251–256 (2011), doi:10.1007/s13218-011-0115-x
21. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: A presence questionnaire. Presence 7(3), 225–240 (1998)

A Question-answering Agent Using Speech Driven Non-linear Machinima

Anuj Tewari, Timothy Brown, and John Canny

Electrical Engineering and Computer Science,
University of California,
Berkeley
{anuj,jfc}@cs.berkeley.edu, tbrown@berkeley.edu

Abstract. The preschool "literacy gap" is one of the most difficult challenges for education in the US. Children in the lowest SES (Socio-Economic Status) quartile have less than half the working vocabulary of those in the top quartile at age 3. On the other hand, preschool children are incessantly inquisitive, and will readily engage in question answering and asking activities if given the opportunity. We argue here that question asking/answering technologies can play a major role in early literacy. We discuss the system implementation of a virtual agent called Spot, that plays a 20-questions game with preschoolers. We describe how a platform for speech-driven non-linear machinima was used to develop this agent. We also present a preliminary computational evaluation of the performance of the system in view of the data collected through a wizard-of-study.

Keywords: Question-Answering Agent, Pedagogical Agent, Conversational Agent, Language Learning, Machinima.

1 Introduction

A large body of research has shown that the literacy gap between children is well-established before formal schooling begins, that it is enormous, and that it predicts academic performance throughout primary, middle and secondary school. Indeed rather than closing this gap, there is much evidence that formal schooling exacerbates it: once behind in reading and vocabulary, children read with lower comprehension, learn more slowly and have lower motivation than their more language-able peers. Many national organizations recognize the essential role of early literacy in a childs later educational and life opportunities [1],[2],[3]. Hart and Risley [4] report a factor of two difference in the working vocabularies of high vs. low-SES three-year-olds. The average low-SES child has heard 30 million fewer words than a high-SES child by this age. However, they also observed that SES alone is not a predictor of cognitive development at the preschool stage. "The richness of nouns, modifiers, and past-tense verbs in their parents' utterances, their parents' high propensity to ask yes/no questions, especially auxillary-fronted yes/no questions; and their parents' low propensity to initiate and use imperatives and prohibitions were more strongly predictive of the children's performance on the Stanford-Binet IQ test battery than was the family SES." Hart and Risley note that to close this gap is an enormous challenge and will require lengthy and regular language experiences for the child.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 129–138, 2013.

Since it has been conclusively established in recent research (cited above) that question-answering serves as a heavily-utilized tool in early child development, we want to test interventions in that area. Therefore, this paper talks about speech-enabled technologies aimed towards early childhood literacy that happens through conversations, and primarily question-answer sequences.

The paper makes the following contibutions:

- We propose a system architecture for dialogue controlled non-linear machinima. Machinima is the process of using in-game recording facilities to record segments of game action under high-level control of the player [23].
- We prototype a question-answering system for preschool children using the proposed system architecture.
- We do a wizard-of-study to collect "gold data" and evaluate our system's performance in view of that data.

2 Related Work

With the growing maturity of conversational technologies, the possibilities for integrating conversation and discourse in e- learning are receiving greater attention in both research and commercial settings. Conversational agents build on traditional education systems, providing a natural and practical interface for the learner. They are capable of offering bespoke support for each individual, and recognizing and building upon the strengths, interests and abilities of individuals in order to foster engaged and independent learners.

In terms of chatbots there are some known systems, like ELIZA [14], which was regarded as one of the first chatbots. ELIZA analysed input sentences and created its response based on reassembly rules associated with a decomposition of the input. This produced an impression of caring about its users, but it held no memory of the conversation and so could not enter into any form of targeted collaboration or negotiation. A.L.I.C.E. [15] is a chatbot built using Artificial Intelligence Markup Language (AIML), developed over the past 10 years. The chatbot is based on categories containing a stimulus, or pattern, and a template for the response. Category patterns are matched to find the most appropriate response to a user input. Further AIML tags provide for consideration of context, conditional branching and supervised learning to produce new responses. The Jabberwacky [16] chatbot has as its aim to simulate natural human chat in an interesting, entertaining and humorous manner. Jabberwacky learns from all its previous conversations with humans. Moreover, all these systems have been developed and tested with adults or adolescents in focus. We were not able to identify any system that was developed and deployed with pre-school children. Moreover, most of the systems are focused on a specific task, and try to engage the users in conversations around those tasks. All of these are also web-based and involve no speech-based interactions.

There is also some recent reflective work in media psychology [17] and Education [18] that critically analyses the results of experiments with pedagogical agents in general. It is also interesting to note that both these papers argue that most of the studies in this domain lack long-term evidence and credibility. Moreover, none of the works cited in these papers has question-answering as its focus. Inclusion of pre-school children in the target population is equally rare, for this kind of research.

3 System Architecture

3.1 Speech Recognition

This component of the system used Google Web Speech API. The reason for using Google's Speech API and not training our own recognizer was that it is most accurate open-ended speech recognition available, and also considerably reduces development time. Also advancing the state-of-the-art for recognition of children's speech is beyond the scope of our work. Moreover, this component was written in Javascript and communicated with the Machinima component of the system using websockets. The Javascript page was hosted by a local Apache server. This component ran the speech recognizer continuously, waiting for long pauses in speech. As soon as a pause was encountered, the recognition results were sent to the Machinima component for processing and response generation.

3.2 Non-linear Machinima

The speech recognition component exchanged information with Flex (Actionscript) code. The Flex code was built to run an AIR (Adobe Intergrated Runtime) socket server, and to do script-controlled Machinima. The AIR server was built to send information received from the speech recognition component to the language processing component. The script-controlled machinima code was built to use this information to choose the video to be played. The machinima code was built to run videos in a loop. Each time the video started, the Flex code would check the "emotional state" of the system. This "emotional state" of the system was determined by the language processing component. The "emotional state" is just a deterministic state that the language processing unit generates. For example, keywords like "happy", "sad", "idle", "talking" could be used to denote such a state. This mapping between the "emotional state" generated by the language processing unit, and the machinima video played by the system, can be supplied as an XML file. In terms of audio, the machinima component was built to maintain an array of pre-recorded audio, and read the mapping between video and audio as an XML file. However, for a more open-ended system this could be substituted with a speech synthesizer.

3.3 Language Processing

The language prcessing component was designed to take the incoming transcription of the user's speech, and generate an "emotional state" of the system (defined above). This particular component could be as complex as the developer wants it to be. For example, for a simple question-answering system (as discussed in next section), this component would just generate a "yes" or "no" response. For a fully open question-answering system, this component would simulate a chatbot. However, most visual chatbots only perform a few visual acts, therefore our machinima architecture could still be re-used. The entire language processing code was built in Java, and communicated with the machinima component, over a TCP socket connection. The language processing also allowed the developer to override the output generated by it, and send a command to the machinima component. This was done with the knowledge that all such automated systems have their restrictions, and sometimes a wizard-of-oz setup is required.

4 Question-Answering Agent

4.1 Interface

Using the system that we described so far as the foundation, a preliminary question-answering agent was built. A puppy character was employed to be the agent. He was named Spot, and will be referred to by that name henceforth.

Previous research also suggested that such agents should have lifelike characteristics [24]. To create engaging videos with minimal effort, we used Machinima from the SIMS Pets game. SIMS is a widely used god game that supports high-level control of characters, including non-human characters. Machinima is the process of using in-game recording facilities to record segments of game action under high-level control of the player [23]. A puppy character was created in the SIMS create-a-pet tool and machinima videos were recorded (with the inbuilt video capture in SIMS) with various different personality traits set. In the create-a-pet tool, depending on what personality is chosen the pet character responds through gestures.

In terms of the capabilities of the agent, we chose activities that were simple. The point was to demonstrate value to the proposed system architecture. Therefore, our system was built to conduct a session of twenty questions. In simpler words, Spot was programmed to show two objects, then shuffle them, and hide one of them inside a box. The user was then supposed to ask questions about the objects to determine which one was hidden under the box. The speech recognition component would transcribe whatever was being asked, the language processing routine would process the incoming transcript and determine a response, and the machinima component would play a video (and audio) corresponding to it. The dialogue design, interface design and evaluation of this system is outside the scope of this paper.

4.2 Question Analysis

Determining Whether or Not a Statement is a Question. In a typical case, the verb precedes the subject of the sentence. In the context of the activity at hand, most of the questions are simple in nature. Since the object in the box is unknown, the subject of the sentence is typically referring vaguely to the object as it or to an action that someone can do with or to the object, so the subject is along the lines of you or I. The types of verbs that we see at the beginning of these questions are a small subset of verbs and along the lines of can, does, or is. Looking for the index of the verb and subject of the sentence allows for a simple comparison of which one occurs first in order to see if it is a question in the general case, with the case where the index of the subject is greater than the index of the verb being a valid question.

However, there are some exceptions to the discussion above. Commands such as do this will be seen as questions with the above technique so the case of do this or do that is checked for and outputs that it is in fact not a question. The five Ws (who, what, when, where, and why) also cause complications for the technique for the general case. Although they are considered question words, if the subject follows them immediately then they are in fact sentences. (Example: what we found, was the following) Both of these exceptions are handled but probably unlikely to occur given the context of the game. A child is most likely not going to try to issue verbal commands to the game such as do this and using one of the five Ws in a sentence as opposed to a question requires a more complex sentence structure that preschoolers are not experienced in.

Extracting the Content of the Question. The main content is referring to the meat of the question. It is the property of the object that the question is referring to. To extract the main content, the prepositional phrase is removed if there is one by locating the preposition from a set of common prepositions in questions and removing the content after it. The main content is found by taking all of the words between a start and end index that are not filler words (a, an, the, have), one of the common question verbs, or a negation.

- If one of the five Ws occurs before the index of the subject then we start at the index of the W and end at the index of the verb
- In general we start at the index of the subject and end at the end of the sentence without the prepositional phrase

Moreover, when looking at the transcripts of the questions asked while the game was played, questions that contained prepositional phrases referred to a location. This location was typically in reference to where something belonged, like a hat on your head or an elephant in a zoo. Gaining the main content of these phrases was done by finding the preposition and taking the content after that if the word was not a filler, common subject, or a possessive.

4.3 Question Answering

Object Properties and Locations. An XML document was created that contained properties of the object, which in this case refers to physical attributes along with actions that can be done to or with an object. A locations category was also made which contains places where the object is commonly found. There is potential to automatically extract this information from online resources, but outside the scope of this paper.

Matching Content. Content is considered a match if it is within a Levenshtein distance of the length of the content string divided by three. This is done so that you do not need to enter every form of a word. For example, listing bouncy as a property will also answer the question does it bounce? correctly as well even though bounce is not listed. Once a question is asked the content of that question is extracted. If the main content of the question matches one of the properties in the xml document under that object then there are two scenarios that follow:

- There is no prepositional phrase so a positive response is generated
- There is a prepositional phrase.

1. If the content of the prepositional phrase matches one of the locations, a positive response is generated.
2. If the content does not match a location, then a negative response is generated.

If the main question content does not match any of the properties then a negative response is generated

Handling Negations. When doing the analysis on the question, the number of negations in the question is counted and taken modulo two. If the output of the above is a one then the positive outputs are switched to negative ones and negative outputs are changed to positive ones.

Reveal. If the question is whether the hidden object is one of the two possibilities and the player gets it right, a response of reveal is generated instead of the usual yes/no response.

5 Performance Evaluation

To evaluate the performance of the system we built, we needed to have a reasonable ”gold dataset” that could be used to compare with the responses that our system generated. This exercise is a necessary one before we directly use our system with children, as this would help come up with recovery mechanisms and dialogue repair required. A faulty piece in most dialogue system can be the speech recognition component. Therefore, we decided to do a wizard-of-oz data collection study in which we play the role of the speech recognition and language processing component. In simple words, instead of these components, a researcher sends direct commands to the machinima component based on what the child (user) says. Obviously, most of the errors in such a system would come from recognition errors or natural language understanding errors, so such a study was ideal for producing the ”gold data”. The efficiency of the automated question-answering system could then be seen as a comparison of the ”gold data” with the output of the built system. The following subsections explain how the data collection study was conducted.

5.1 Wizard-of-Oz Study for Data Collection

Participants. 20 children (10 boys, 10 girls) participated in our feasibility study. The participants in the study were 4 and 5 year old children at a preschool in California. Previous research suggested that 3-year-old children would be too young for such an experiment [8]. The preschool that was chosen as the location for the study was a research preschool.

Equipment and Setup. The study was conducted in a research room on the preschool premises, reserved for that purpose. The room was equipped with one-way mirror and audio equipment that allowed a visual supervisor to monitor the study at all times.

The presence of the visual supervisor was required by the preschools protocols. During the study two researchers were present for all sessions, in addition to the child. The children could see the researchers, but not the visual supervisor. A video camera recorded the childs and researchers activity at all times.

Method. During the study each child attended a session individually. Before each session a researcher from the team went to one of the classrooms and invited participants to attend a study sessions. Out of the consenting children one was escorted to the study room. As stated above, the study session could not last more than 20 minutes.

Each session comprised of multiple question-answering exchange trials. The child was told that the purpose of the game is to ask questions and figure out which object is being retained. The session started with a demo trial, where the two objects were cat and ball. In the demo trial, one of the researchers asked questions and Spot answered (based on the commands that another researcher sent to Spot). If the child did not understand the demonstration, it was repeated till the child was comfortable in contributing to the questions being asked. After the demo trial, 6 more trials followed. Spot was supposed to conduct the entire game session. As mentioned already Spot used a script to go through its dialogues. The part of setting up the game remained the same for all users. In each game session spot introduced itself, explained the rules of the game step-by-step and then went through the trials with different object pairs. In any given trial, Spot would first show the two items, identify both of them, then convert them into question marks. After this, through some animation, one question mark would leave the screen and the other one will go into a box. All of this is depicted in Figure. Once the object was hidden, the child was expected to ask questions to figure out the hidden object. After this, the game took different conversational routes for different participants.

Each child went through 7 pairs of objects, including the demo trial. For each trial, the stimulus pair of objects presented got more difficult. Increased difficulty meant increased similarity in the stimulus pair. For example, a cat and a ball are easily distinguishable, but a bicycle and a car are harder to differentiate. In formal terms, the more difficult stimulus pairs were closer to each other in terms of parts, functions and properties. A list of all the pairs is given in Table 1. It should be noted that no object was repeated across two trials. The researcher conducting the session tried to stick to the dialogue script designed for Spot, and deviated only if the child went off-topic or got confused. In simple words, the script was supposed to be overruled only if the conversation needed repair, despite the strategies used in the script. The deviations were allowed because parents in practice use a lot of strategies to engage children and technology cannot replicate all of those. The point of this activity was to collect audio data for the

Table 1. Object pairs used in the two phases

Target Item	Low Similarity	Moderate Similarity	High Similarity
Phase I	Book/Banana Elephant/Spoon	Table/Bed Shoe/Hat	Apple/Orange Bicycle/Car
Phase II	Chair/Rose Flower/Kite	Bear/Dog Chicken/Pig	Truck/Bus Clock/Watch

session, and then use it to evaluate the efficiency of the automated system. The collected data will help develop a system that mimics the experience of the wizard-of-oz study.

5.2 Computational Experiments

Once the data through the wizard-of-oz study had been collected, the next step was to run it by the automated system and measure the performance exhibited. Therefore, the recorded audio was divided into various audio snippets. After this they were converted into FLAC (Free Lossless Audio Compression) format and sent to Google's recognition endpoint that returned JSON objects as recognized hypothesis. Once the recognized transcripts were obtained, they were turned in as inputs to the language processing component and response generated by the system was recorded. If the generated response was the same as the one generated by the researchers during the wizard-of-study, that was considered a perfect response.

Speech Recognition Errors. A major component of errors in dialogue systems come from speech recognition errors. It is also hard to train a speech recognizer for children's voices. However, with advancement of cloud based speech recognition systems and Google releasing its web speech API, it has become easier for developers to include speech recognition into their applications. As explained before, we used the Google speech recognition system as a part of our system, instead of training our own recognizer. Therefore, it was critical to evaluate the performance of speech recognition alone, on children's voices. Out of a total number of 346 utterances, the speech recognition was unable to produce a correct transcription for only 49 utterances, which is just in 14% cases. In these 14% cases, the average Word Error Rate (WER) was 0.21. This means that even for the misrecognized transcriptions, the amount of misrecognition was low. Looking at the errors in more detail, the following were some recurring cases:

1. Some errors were because children used colloquial like ”does it meow” and ”does it make oink oink sounds?”, while asking questions.
2. Some errors were just language model based errors, like ”fur” getting recognized as ”for” and ”twirly tail” getting recognized as ”great detail”.

However, in spite of these errors, the performance of the recognition system is quite impressive by any standard, and the error rates beat any error rates reported by a system trained for children's speech.

Natural Language Understanding Errors. The most common reason why the algorithm output incorrect response in these questions was that the property was not listed in the xml file. If a property is not listed, then the algorithm thinks that the item being distinguished cannot do a certain action or that it has a certain characteristic. In addition to that, if a location where an action can be performed is not listed, the response will be incorrect. This can be seen in the case of ”Can you pedal it to the library?” where the output is a ”no” for a bicycle simply because it does not know that it can be pedaled specifically to a library. There are also cases where the property listed in the xml file results in the wrong answer, such as listing ”flow” as a property for kite caused a question asking if it was a flower to be answered as ”yes, it does” solely due to the fact that flower and flow are within an acceptable edit distance. Also the context of the question is not captured so asking ”Can you eat it?” and ”Do you eat at it?” to both be seen as questions about eat which will lead to errors such as saying yes to a question about being able to eat a table. Also the algorithm is only able to produce answers to well formed questions and fails on something like ”a watch?” which could be a question or a simple statement. There was also a case in the other extreme where the question was too complicated and did not fit the general form that we were looking for. Using ”and” and ”but” will in general throw off the algorithm at this point in time. Also adding adverbs and uncommon adjectives to properties will throw off the system because they are not listed in the properties xml file word for word.

In various cases, the difference from the actual transcript and what the software produced sounded alike but was out of the acceptable edit distance range so it was not recognized as a valid property of the object. Sounds were commonly misrepresented. By adding a lookup table for commonly mistaken phrases and what they actually should be, we were able to increase the accuracy of the algorithm.

However, in spite of the above restrictions, the question-answering algorithm had good matching accuracy. For a total of 346 utterances generated by the children in our study, the algorithm was able to generate the correct response in 297 cases. Which is a matching accuracy of 86%. In simple words, for 86% cases of the child talking to the system, the system was able to generate a response that the wizard generated during the data collection study. Also, when the lookup table of most common errors was added, the accuracy went up to 89%. If the errors of the speech recognition component were ignored and the question-answering component was given 100% accurate transcripts, the matching error rate was 94%. This means that there was only a drop of 5% accuracy because of speech recognition errors.

6 Conclusion

In this paper we propose a system architecture that could be used to do dialogue controlled non-linear machinima. The advantage to the proposed architecture is that the developer doesn’t need to rebuild speech recognition and script-controlled machinima components. Depending on the need of the system, just the language processing component can be tweaked. For the machinima component, the developer can just record new clips of their choice and assign them to responses generated by the language processing component. Given its simplicity, the system is restricted in what it offers.

In addition to the above contributions, we also built and evaluated a question-answering system for preschoolers. The wizard-of-oz study to collect data, helped in optimizing the system for better performance. The next step which is fairly obvious, is to evaluate the system in the field with realtime speech input. Even though we leave it as a future work, it will be interesting to explore the engagement and redirection strategies that the system could use to gain a child’s attention back after recovering from an error.

References

1. NELP: National Early Literacy Panel, Developing Early Literacy, National Institute for Literacy, NIFL (2008), `http://www.nifl.gov`

2. NFCL: National Family Literacy Organization, main site, http://www.famlit.org
3. NIH: Clear Communication: An NIH Health Literacy Initiative, http://www.nih.gov/clearcommunication/healthliteracy.htm
4. Hart, B., Risley, T.: Meaningful Differences in the Everyday Experience of Young American Children. Paul H. Brookes (1995)
5. Mol, S.E., Bus, A.G., de Jong, M.T., Smeets, D.J.: Added value of dialogic parent-child book readings: A meta-analysis. Early Education and Development 19, 7–26 (2008)
6. http://www.millee.org
7. Kumar, A., Tewari, A., Shroff, G., Chittamuru, D., Kam, M., Canny, J.: An Exploratory Study of Unsupervised Mobile Learning in Rural India. In: Proceedings of CHI 2010, Atlanta, Georgia (2010)
8. Chouinard, M.M.: Childrens questions: A mechanism for cognitive development. Monographs of the Society for Research in Child Development 72(1 Serial No. 286) (2007)
9. Wallace, R.: The elements of AIML style, Online text at http://www.alicebot.org/join.html
10. MacWhinney, B., Snow, C.: The Child Language Data Exchange System: An update. Journal of Child Language 17, 457–472 (1990)
11. Tewari, A., Goyal, N., Chan, M., Yau, T., Canny, J., Schroeder, U.: SPRING: Speech and Pronunciation Improvement through Games, for Hispanic children. In: To Appear in Proceedings of ICTD 2010, London (December 2010)
12. Hutchinson, H., Bederson, B.B., Druin, A.: Supporting Elementary-Age Children's Searching and Browsing: Design and Evaluation Using the International Children's Digital Library. Journal of the American Society for Information Science and Technology, John Wiley and Sons 58(11), 1618–1630 (2007)
13. Callanan, M.A., Oakes, L.M.: Preschoolers questions and parents explanations: Causal thinking in everyday activity. Cognitive Development 7, 213–233 (1992)
14. Weizenbaum, J.: ELIZA - A Computer Program For The Study of Natural Language Communications Between Man and Machine. Communications of the ACM 9(1), 36–45 (1966)
15. Wallace, R.S.: Chapter 00. The Anatomy of A.L.I.C.E., http://www.alicebot.org/documentation
16. Carpenter, R.: Jabberwacky.com (1997-2006), http://www.jabberwacky.com
17. Kramer, N.C.: Psychological Research on Embodied Conversational Agents: The Case of Pedagogical Agents. Journal of Media Psychology 2010 22(2), 47–51 (2010)
18. Clarebout, G., Elen, J., Johnson, W.L., Shaw, E.: Animated pedagogical agents. An Opportunity to be Grasped? Journal of Educational Multimedia and Hypermedia 11, 267–286 (2002)
19. Frazier, B.N., Gelman, S.A., Wellman, H.M.: Preschoolers Search for Explanatory Information Within AdultChild Conversation, Child Development, vol. 80(6), pp. 1592–1611 (November/December 2009)
20. Kam, M., Ramachandran, D., Devanathan, D., Tewari, A., Canny, J.: Localized Iterative Design for Language Learning in Underdeveloped Regions: The PACE Framework. In: Proceedings of the 2007 ACM Conference on Human Factors in Computing Systems (CHI 2007), pp. 1097–1106 (2007)
21. Hutchinson, H., Bederson, B.B., Druin, A.: Supporting Elementary-Age Children's Searching and Browsing: Design and Evaluation Using the International Children's Digital Library. Journal of the American Society for Information Science and Technology 58(11), 1618–1630 (2007)
22. MacWhinney, B., Snow, C.: The Child Language Data Exchange System: An update. Journal of Child Language 17, 457–472 (1990)
23. Lowood, H.: High-performance play: The making of machinima. In: Clarke, A., Mitchell, G. (eds.) Videogames and art: Intersections and interactions. Intellect, London
24. Perez-Marin, D., Pascual-Nieto, I.: Conversational Agents and Natural Language Interaction. Techniques and Effective Practices. IGI Global (2011)

A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance

Kirsten Bergmann[1] and Manuela Macedonia[2,3]

[1] SFB 673, CITEC, Faculty of Technology, Bielefeld University, Germany
[2] Department for Information Systems, University of Linz, Austria
[3] Research Group "Neural Mechanisms of Human Communication", Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
kirsten.bergmann@uni-bielefeld.de,
manuela.macedonia@jku.at

Abstract. An important and often laborious task in foreign language acquisition is vocabulary learning. Research has repeatedly demonstrated that performing iconic gestures together with novel words has a beneficial effect on learning performance. Can these findings be transferred onto virtual agents applied in gesture-supported vocabulary training? We present a study investigating whether iconic gestures performed by a virtual agent and imitated by learners have an impact on verbal memory for words in a foreign language. In a within-subject design we compared participants' memory performance achieved with the help of a virtual agent and those achieved with the help of a human trainer regarding both short-term learning effects and long-term decay effects. The overall results demonstrate improved memory scores when participants learned with a virtual agent. Especially high performers could profit from gesture-supported training with a virtual agent.

Keywords: Vocabulary acquisition, iconic gestures, memory performance.

1 Introduction

A major challenge in learning foreign languages is the acquisition of novel words. Generations of language learners have struggled with written or spoken materials, e.g., in form of word pair lists. Nowadays, multimodal learning has become more and more popular so that learning materials are enriched by pictures, videos or music to make the acquisition of new words easier. There is also evidence for the beneficial role of gestures in vocabulary acquisition: Performing an iconic gesture, i.e., a gesture whose physical form corresponds with object features [13], together with a novel word, enables faster learning and makes those words more resistant against forgetting [4,9,11,12,17,18, for details see section 2.1].

It stands to reason that these findings are of great importance for foreign language pedagogics, but so far they have only barely found their way into class rooms and teaching materials. Among the possibilities to train vocabulary items with gestures, a new generation of trainers has to be considered: embodied conversational agents. The embodied nature of virtual agents allows for multimodal communication. Nonverbal communication skills comprise gestural behavior, but also other behaviors like facial expressions.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 139–148, 2013.

Both factors play an important role in learning environments [1]. Moreover, virtual trainers are characterized by a high degree of flexibility. In terms of personalization an agent's communicative behaviors, specifically the use of gestures, can easily be adapted to particular needs and preferences of the learner. In addition, unlike a human teacher or tutor, a virtual character is always available and supportive.

Although there is an increasing number of studies dealing with the effects of virtual agents in foreign language learning, nonverbal communication plays only a minor role in agent-supported learning (cf. [6]). With particular regard to gestures, only pointing gestures have been considered so far [1,14]. And in these studies results with regard to memory performance diverge (for details see section 2.2). Our study presented here investigates whether iconic gestures performed by a virtual agent and imitated by learners have an impact on verbal memory for words in a foreign language. Furthermore, the aim of this study is to compare memory results achieved with the help of a virtual agent and those achieved with the help of a human trainer.

2 Background and Related Work

2.1 Gestures Have an Impact on Verbal Memory

A growing body of scientific evidence has demonstrated the so-called *enactment effect* [3] for more than 30 years: action words like 'to go' or phrases like 'play piano' are better recalled if they are performed than if learners listen to them or imagine to perform them. Going beyond verbal memory in native language, several studies have shown that gestures can also have an impact on memory for words and phrases in a foreign language [10]. The first systematic study on this topic was conducted by Quinn-Allen [17]. English-speaking students learned French expressions that simultaneously were accompanied by emblematic gestures. Better recall were achieved as compared to a control group. The study also demonstrated a beneficial long-term effect of gestures on information decay: Eleven weeks after encoding, gesture performers had forgotten less than control subjects. Other studies replicated and extended these findings. Macedonia [11] investigated single word retention for 36 items belonging to different word categories. In a within-subject design, participants learned words nouns, adjectives, verbs and prepositions of an artificial language corpus either by reading and listening to them or by additionally performing gestures. Results showed that retrieval (cued recall) was significantly better in the short- and long-term for enacted items.

Tellier [18] presented common words like house, swim, cry etc. to French children learning English. Half of the items were associated with a picture and the other half were illustrated by a gesture that the children saw in a video and thereafter performed. Enacted items were better memorized than items enriched visually by the pictures. Moreover, Macedonia and Knösche [12] demonstrated that enactment enhances memory performance not only for concrete, but for also for abstract words (nouns, verbs, and adverbs). In an additional transfer test, participants of this study were asked to produce new (non-canonical) sentences with the words they had learned during the training. Enacted items were recruited significantly more often than words learned audio-visually.

Other studies investigated whether the type of gestures performed while learning has an effect on word learning. Kelly et al. [4] trained young adults on Japanese verbs conveying common everyday meanings. The words were presented according to four modes: (i) speech, (ii) speech + congruent gesture, (iii) speech + incongruent gesture, and (iv) repeated speech. The results showed that participants memorized the largest number of words in the speech + congruent gesture mode, followed by the repeated speech mode.The least number of words was memorized when they were accompanied by an incongruent gesture. Another study controlling for the type of gestures was conducted by Macedonia et al. [9]. Subjects were trained on 92 concrete nouns. Half of the items were encoded with iconic gestures. They depicted some aspect of each words semantics and enriched the word with a plausible sensorimotor connotation. The other half of the items was learned with meaningless gestures. The results showed better memory performance for iconic gestures than for meaningless gestures in the short- and long-term (after 60 days).

2.2 Virtual Agents: Effects on Memory Performance

Due to the sparse number of experiments on this topic, the question whether agents have an impact on the memory performance of learners is still a matter of debate and we expand our view to virtual teachers, trainers or tutors in general learning contexts here. Beun and colleagues [2] compared two types of virtual agents (realistic and cartoon) with a no-agent condition in a text comprehension task. Subjects were provided with a short story they had to re-tell directly after the stimulus presentation. The presence of an embodied agent had a positive effect on the retainability of information. On the contrary, Moundridou and Virvou [15] did not find a beneficial effect of the presence of a virtual agent when students were solving math problems. Similarly, Mulken et al. [16] investigated whether the presence of a virtual agent was helpful for participants recall rate of technical material (information about machines) as well as non-technical material (introduction of fictitious people). Researchers compared an agent using deictic gestures with a pointer stick against a no-agent condition in which only the pointer sticks were visible. Recall tests directly after stimulus presentation showed that there was neither a positive nor a negative effect on recall.

In contrast to the aforementioned studies, Mitsako et al. [14] evaluated the effect of a gesticulating virtual agent as vocabulary trainer on learning performance in *repeated* interactions. German speaking participants learned English words either with or without a virtual agent in four sessions over eight days. In the no-agent condition the words were displayed on a screen and spoken by a text-to-speech system, whereas in the agent condition, a female agent was additionally present. She featured some idle movements and pointing gestures referring to the words to be learned. The presence of the agent had no effect on learning performance (‘persona-zero’ effect). Altogether, previous related studies revealed rather controversial effects of virtual agents on learners memory performance. This holds for the mere presence of embodied agents as well for use of nonverbal behavior. Nevertheless, given that existing implementations of nonverbal behavior are often by far not as natural as human behavior, researchers are still convinced that nonverbal communication is “a huge potential for agent systems” [6, p. 80]. Given that experiments have only tested the impact of deictic gestures on learning behaviors, iconic gestures still remain a field to be explored. In fact, it has been repeatedly demonstrated that they support verbal memory (see section 2.1). Another issue that has not been evaluated sufficiently are effects of virtual agents on long term memory performance. Mulken et al. [16] explicitly point to this and speculate that other effects might be found on a long term performance.

3 Study

We assess the effect of iconic gestures performed by a virtual agent compared to gestures performed by a human on memory performance for words in a foreign language. Given the embodied nature of virtual agents and their capabilities of expression, we expect that gesture-supported vocabulary training with a virtual agent will be as successful as training with equivalent gestures presented by a human. Second, we investigate both short-term effects of learning and long-term effects of information decay. From research on video stimuli of humans for the same purpose [4,9,11,12,17,18] we expect that gesture-based training with a virtual agent will also be beneficial for both.

The study employed a within-subject design manipulating the type of training: (1) Gesture-based training with video stimuli of a human (human gestures; HG), (2) Gesture-based training with video stimuli of a virtual agent (agent gestures; AG), and (3) a control condition without any gestures (Con). Participants trained vocabularies on three consecutive days. Their (short-term) learning performance was measured the next day prior to the next training session, respectively (day1, day2, day3). The long-term effect of information decay was measured additionally four weeks after training was finished (day30).

3.1 Participants

A total of 32 native speakers of German, aged from 20 to 40 (M=25.3, SD=3.5), participated in the study. 15 participants were female and 17 participants were male. All of them were either paid or received credits for their participation. Subjects were randomly assigned to two training groups to counterbalance training conditions and items.

3.2 Materials

The training material comprised 45 nouns in ‘Vimmi’, an artificial corpus created for experimental purposes [9,12]. It aims to avoid associations and to control for different factors that, in natural languages, can favor the memorization of particular vocabulary items. The Vimmi items were created according to Italian phonotactic rules. First they were randomly generated by Perl and thereafter adjusted to avoid tautological occurrence of syllables, high frequency of particular consonants or vowels, the appearance of strings sounding unusual to German-speaking subjects, association with words from European languages taught at school (English, French, Italian, and Spanish), and with proper nouns comprising names of products available on the German market. The artificial words were assigned common meanings like bridge and suitcase. Word familiarity was controlled for using the word frequency counter of German[1]. The 45 words were recorded in 45 single audio files, with each file having a length of approximately 0.8s.

[1] http://Wortschatz.Uni-Leipzig.de

(a) Gesture for the word 'staircase'

(b) Gesture for the word 'mirror'

Fig. 1. Example stimuli performed by the virtual agent and human actress

The gestures mirrored some feature of the semantics of the word that was arbitrarily chosen from a range of possibilities discussed in the team. For instance, for the word 'staircase', the actress modeled several steps of different height on after another (figure 1a). For the human gesture stimuli we video-taped a human actress. The agent's gestures were modeled to match the human gestures as much as possible. They were specified in the Multimodal Utterance Representation Markup Language (MURML; [7]) and realized with the Articulated Communicator Engine (ACE; [5]), a toolkit for building animated embodied agents that are able to generate human-like multimodal utterances. The agent's gestures were also rendered into video data. For examples of both human and agent gesture stimuli see figure 1.

3.3 Procedure

Participants were informed that they took part in an experiment on foreign language learning with the goal to memorize as many words as possible. Participants also were informed that their performance would be assessed at different time points through different kinds of written tests. The training lasted approximately 45min per day on three consecutive days. Training was performed in groups. The total of 45 words were subdivided into three blocks of 15 words each, in which the three training conditions daily alternated and counterbalanced the experimental conditions. In each block the items were randomly subdivided into three smaller blocks of five items each. A block was first shown and participants were instructed to watch it. Thereafter, the block was played again six times and participants were cued to imitate the gesture and/or to repeat the word in Vimmi after seeing and hearing it. All the words were randomized within the blocks. In total, every vocabulary item was presented seven times every day. After each training block, a break of five minutes followed. The software used for the training was Presentation[2].

[2] v16.3 by Neurobehavioral Systems, Albany, CA.

3.4 Tests

Memory performance was assessed daily starting from the second experiment day measuring the learning outcome of the first training day etc. (day 1-3) and on day 30. Participants were administered a free and thereafter a cued recall test. In the free recall test participants were provided with an empty sheet. They were instructed to write as many items as possible in both languages. Items could be loose (i.e., only German or only Vimmi) or matched (i.e., Vimmi and German). In the cued recall test participants were given a randomized list of the 45 trained items to be translated from German into Vimmi and then a further randomized list of the same words to be translated from Vimmi into German (duration 5 min), with the instructions to translate the items from one language into the other. The order of the translation from one into the other language alternated daily. Items were considered correct if their spelling corresponded 100% to the word spelling provided during training (score of 1). Partial correctness was only considered in case of minor mistakes, e.g., interchanged letters like 'asemo-aseno' for the Vimmi words or nominalization effects like 'to pipe-pipe' in German. Here a score of 0.5 was given. All other items were considered wrong and given a score of 0.

4 Results

Data from 29 participants was analyzed for both short-term and long-term effects. Three participants had to be removed from the data set. This was due to data storage problems in two cases. In one case it turned out that the participant did not fit into the group of young adults we aimed to investigate with our study.

4.1 Short-Term Effects

In order to assess the influence of training on memory performance in free and cued recall respectively, a repeated-measures ANOVA with the within-subject factors TIME (day1, day2, day3) and TRAINING (HG, AG, Con) was conducted. Means and standard deviations are summarized in table 1.

Free Recall. For the aggregated measure of free recall Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of TIME (χ^2=20.45, p<.001). Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ϵ=.65). There was a significant main effect of TIME one free recall rate (F(2,56)=187.26, p<.001). Contrasts revealed that memory performance increased significantly from day2 to day3 (F(1,28)=160.23, p<.001). More interestingly, there was also a main effect of TRAINING one free recall rate at a significant level (F(2,56)=4.24, p=.019). Memory performance was significantly better for words trained with the agent gestures (AG) as compared to words that had been trained in the control condition (F(1,28)=6.25, p=.019).

Table 1. Short term training results for the aggregated measure of free and cued recall: Mean scores of proportional recall and standard deviations in parentheses

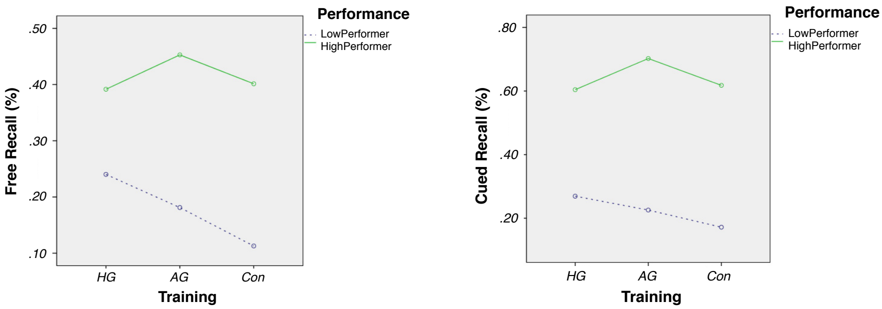
		HG	AG	Con
Free Recall	day1	15.56 (6.64)	18.05 (9.23)	11.92 (7.65)
	day2	33.10 (12.02)	34.79 (16.02)	28.24 (16.21)
	day3	46.09 (18.11)	46.59 (17.38)	43.98 (19.04)
Cued Recall	day1	9.66 (11.13)	11.78 (13.77)	13.16 (11.24)
	day2	41.44 (25.05)	43.10 (27.80)	41.03 (28.66)
	day3	60.98 (25.35)	63.28 (25.69)	60.11 (29.67)

Cued Recall. For the aggregated measure of free recall Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of TIME (χ^2=19.55, p<.001). Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ϵ=.66). There was a significant main effect of TIME on free recall rate ($F(2,56)=139.69, p<.001$). Contrasts revealed that memory performance increased significantly from day2 to day3 ($F(1,28)=146.79, p<.001$).

4.2 Long-term Effects

For the analysis of long-term effects on memory, we considered another variable, namely learning performance across learners in vocabulary. We subdivided participants in high and low performers. This we determined by a median-split of the group. The average retrieval performance for the 29 participants over the four time points showed a mean value of 34.57% (SD 15.24). The median value of 30.6% split the group into two subgroups: 13 low performers (4 females and 9 males) with a mean performance of 21.20% and 16 high performers (10 females and 6 males) with a mean performance of 45.43%. In order to assess the influence of training on memory performance in free and cued recall respectively, a mixed design ANOVA with the within-subject factor TRAINING (HG, AG, Con) and the between-subject factor PERFORMANCE (high performers, low performers) was conducted.

Free Recall. For the aggregated measure of free recall the main effect of TRAINING was significant ($F(2,54)=3.68, p=.032$). Compared to the control condition the amount of correctly recalled items was significantly higher in both the training condition with human gestures ($F(1,27)=4.72$, $p=.039$) as well as the training condition with agent gestures ($F(1,27)=9.17, p=.005$). Not surprisingly, there was also a significant effect of PERFORMANCE, indicating that high performers had a significantly enhanced memory performance as compared to low performers ($F(1,27)=19.39, p<.001$). We also found a significant interaction effect between TRAINING and PERFORMANCE ($F(2,54)=4.37$, $p=.017$) indicating that the different training conditions affected low and high performers differently. In order to break down this interaction, we compared the different training groups across high and low performers. This analysis revealed significant interactions when between the gesture-based training with human stimuli and the control condition across high and low performing participants ($F(1,27)=6.41$, $p=.017$).



(a) Interaction of TRAINING and PERFORMANCE in free recall.

(b) Interaction of TRAINING and PERFORMANCE in cued recall.

Fig. 2. Interaction effects of TRAINING and PERFORMANCE after 30 days

The graph in figure 2 plotting the above results suggests that human gestures were helpful to low performers while the supporting effect was lower for high performers. Interestingly, the gestures performed by the agent had an impact on the high performing group.

Cued Recall. In this measure we found a significant main effect of TRAINING (F(2,54)=3.49, p=.038). Gesture-based training with the agent led to significantly better retrieval than the audiovisual training (F(1,27)=5.90, p=.022). There was also a significant main effect of PERFORMANCE indicating that high performers had a significantly enhanced memory performance as compared to low performers (F(1,27)=58.06, p<.001). In addition, there was a significant interaction effect between TRAINING and PERFORMANCE (F(2,54)=3.99, p=.024) indicating that memory scores achieved in the different training conditions were not the same for low and high performers. Again, contrasts were used to break down this interaction. They revealed significant interactions between high and low performers' training with human gestures and the control condition (F(1,27)=4.19, p=.05). The graph in figure 2 shows that the human gestures were more helpful to low performing participants than for high performers. By contrast, the agent's gestures had an impact and were more helpful for the high performing group.

5 Discussion and Conclusions

The present study had two goals: first to test the impact of gestural training provided by a virtual agent on memory for words in a foreign language. Second, this study compared memory performance after a training with an agent and memory performance after a training with a human. We predicted that gestural training would enhance memory for words and that training provided by an human and by an agent would not differently affect memory performance. From a virtual agent perspective this hypothesis could actually be exceeded: For most of the tests we employed, training with a virtual agent led to better memory performance than training with a human. For both types of long-term measures (free and cued recall) memory performance for items learned in the AG condition outperformed memory performance for items learned in the control condition. The same effect was present for short-term measures of free recall. However, training with human gestures did not reach significance in all tests which contradicts empirical evidence of multiple studies (see Sect. 2.1). In terms of an explanation, we speculate that participants of our study only trained for three days. In other studies with the same material training spanned over a longer period of time [9,12]. Nevertheless, in absence of significant benefits of human gestures, the beneficial effects of a virtual agents gestures is even more striking.

To further elucidate the effect of agent-based training being more successful than human-based training, we took participants overall performance into account for the long term effects. Interestingly, we found that agent-based training was particularly successful for high-performers. For low-performers the human-based training resulted in higher memory scores. How can we explain this difference? Here we assume that the virtual agents gestures were less natural than the video recordings of the human actress. This difference in naturalness might bring a hindered processing about. Observers of these gestures might need a few more moments of thinking to clarify what exactly the gesture means, how it relates to the word to be learned and how to imitate it. It may be that this additional thinking further consolidates the information in high performers. For low performers, on the contrary, this further processing might be too much so that they do not benefit from the agents gestures as much as from natural human gestures. Leutner et al. [8, p.83] raise cognitive load theory in this context: "If an agent behaves in a way that is perceived as somewhat strange for the learners, this thinking about strangeness requires cognitive resources, and consequently imposes extraneous cognitive load in working memory". Interpreted in the light of this context our results suggest that high performers, who have stronger language acquisition skills [9], are in a better position to handle this additional load, and they even seem to profit from it.

In future work we aim to further pursue this explanation by collecting ratings of naturalness and the degree of iconicity for the different gesture stimuli employed in our study. Moreover, as we could show in the present study that virtual agents using iconic gestures are actually supportive for vocabulary training, we can think of more comprehensive language training tools based on embodied characters. A virtual trainer permanently at the users disposal on mobile devices can offer individualized training, not only in second language education but also in other domains like language rehabilitation.

Acknowledgements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence 277 "Cognitive Interaction Technology" (CITEC).

References

1. Baylor, A.L., Kim, S.: Designing nonverbal communication for pedagogical agents: When less is more. Computers in Human Behavior 25, 450–457 (2009)
2. Beun, R.-J., de Vos, E., Witteman, C.L.M.: Embodied conversational agents: Effects on memory performance and anthropomorphisation. In: Rist, T., Aylett, R.S., Ballin, D., Rickel, J. (eds.) IVA 2003. LNCS (LNAI), vol. 2792, pp. 315–319. Springer, Heidelberg (2003)

3. Engelkamp, J., Krumnacker, H.: Imaginale und motorische Prozesse beim Behalten verbalen Materials. Zeitschrift für experimentelle und angewandte Psychologie pp. 511–533 (1980)
4. Kelly, S.D., McDevitt, T., Esch, M.: Brief training with co-speech gesture lends a hand to word learning in a foreign language. Language and Cognitive Processes 24, 313–334 (2009)
5. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15, 39–52 (2004)
6. Krämer, N., Bente, G.: Personalizing e-learning. The social effects of pedagogical agents. Educational Psycholoy Review 22, 71–87 (2010)
7. Kranstedt, A., Kopp, S., Wachsmuth, I.: MURML: A multimodal utterance representation markup language for conversational agents. In: AAMAS 2002 Workshop Embodied Conversational Agents-Let's Specify and Evaluate Them! (2002)
8. Leutner, D., Wirth, J.: Commentary on: Personalizing e-learning. The Social Effects of Pedagogical Agents. Educational Psycholoy Review 22, 71–87 (2010)
9. Macedonia, M., Müller, K., Friederici, A.: Neural correlates of high performance in foreign language vocabulary learning. Mind, Brain and Education 4(3), 125–134 (2010)
10. Macedonia, M., von Kriegstein, K.: Gestures enhance foreign language learning. Biolinguistics 6(3-4), 393–416 (2012)
11. Macedonia, M.: Voice Movement Icons' during encoding of foreign language. Ph.D. thesis, Salzburg University (2003)
12. Macedonia, M., Knösche, T.R.: Body in mind: How gestures empower foreign language learning. Mind, Brain, and Education 5, 196–211 (2011)
13. McNeill, D.: Hand and Mind—What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)
14. Miksatko, J., Kipp, K.H., Kipp, M.: The persona zero-effect: Evaluating virtual character benefits on a learning task with repeated interactions. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 475–481. Springer, Heidelberg (2010)
15. Moundridou, M., Virvou, M.: Evaluating the persona effect of an interface agent in a tutoring system. Journal of Computer Assisted Learning 18, 253–261 (2002)
16. Mulken, S., André, E., Müller, J.: The persona effect: How substantial is it? People and Computers pp. 53—66 (1998)
17. Quinn-Allen, L.: The effects of emblematic gestures on the development and access of mental representations of french expressions. The Modern Language Journal 79, 521–529 (1995)
18. Tellier, M.: The effect of gestures on second language memorisation by young children. Gesture 8, 219–235 (2008)

Using Linguistic Alignment to Enhance Learning Experience with Pedagogical Agents: The Special Case of Dialect

Vivien Kühne, Astrid Marieke Rosenthal-von der Pütten, and Nicole C. Krämer

University of Duisburg-Essen, Department of Social Psychology:
Media and Communication, Forsthausweg 2, 47048 Duisburg, Germany
vivien.kuehne@stud.uni-due.de,
{a.rosenthalvdpuetten,nicole.kraemer}@uni-due.de

Abstract. Empirical research showed that verbal and nonverbal alignment occurs in HCI in the same way as in HHI [1-3]. Against the background of similarity attraction [4], a “we-feeling” within dialect-origin [5] and different investigations regarding speaking variations [6,7], the present study analyses the effect of the dialectical language usage of a virtual pedagogical agent within a tutoring setting and the ramifications for the learning situation. An experimental study with a between subject design (N=47) was conducted in which the virtual interlocutor explained and subsequently questioned the subjects about medical topics in either dialect or High German (via Wizard-of-Oz-scenario). The results show that linguistic alignment occurs in both conditions, but even more in interaction with the High German-speaking agent. Furthermore the dialect-using agent was rated as more likable while there were no effects with regard to social presence. Implications for theory and development are discussed.

Keywords: ECA, virtual agent, experimental study, linguistic alignment, communication adaptation theory, High German and dialect, social effects.

1 Introduction

In the 21th century we face virtual agents in online shops, airports, in museums and increasingly often in learning systems. Especially in the case of pedagogical agents, it seems to be important to use virtual characters which are as close to the learner as possible to increase learning motivation [8]. A possible characteristic which has not been considered so far is the use of dialect. Since dialect usage can establish a “we-feeling” [5] between people of the same dialect-origin and because people tend to like people similar to themselves (principle of similarity attraction; [4]), the usage of dialect in human-agent interactions could have the potential to bolster up motivational effects of pedagogical agents. According to Krämer and Bente [9], social factors have important influence on the interaction with a pedagogical agent in a tutoring situation.

Recent studies have shown that people align their nonverbal and verbal behavior in interactions with virtual agents just as they do in human-human interaction (HHI), e.g. regarding the amount of words used during conversation [3] or participants’ smiling

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 149–158, 2013.

behavior [2]. With regard to pedagogical agents, Rosenthal-von der Pütten et al. [10] investigated the linguistic alignment towards a tutoring agent by varying the lexical word-choice and found that participants align both to the agent using everyday language and the agent using technical terms. However, dialect has not been addressed so far. On that account, the present study includes the social factor of dialect and investigates linguistic alignment towards a dialect-speaking agent (Rhine-Ruhr region) and the differences between the ramifications of a dialect- or High German-agent. The possible alignment, the evaluation as well as a possible change of test-anxiety before and after the interaction may give further suggestions for the design and presentation of virtual agents and in this context, for designing better tutoring-systems.

1.1 Dialect in the Rhine-Ruhr Region

The standard language of any country serves as the main medium for communication of a culture, provides rules for syntax and lexis, and serves as benchmark for this language [11]. However, every culture has its own everyday modifications of their language, called colloquial language and dialect, which are more common in daily life. They are not only important because of the communicative function; they also indicate a membership to some kind of group (geographically and mentally) and therefore have a social function like creating a "we-feeling" [5]. This similarity attraction between people manifests not only in language, but also in attitude and values [4].

1.2 Linguistic Alignment in Human-Human Interaction

People use the appearance, nonverbal and verbal behavior of others to derive conclusions for an appropriate interaction and to find a shared communication concept [12]. According to Brennan [13], people have a "vocabulary problem", because they have thousands of possibilities to express themselves. An easy way to find a solution is alignment. People generate "conceptual pacts" [13] and use the strategy of convergence [14] or linguistic alignment [1], a dynamic and adaptive exchange between communication partners for a successful interaction, to manage the comprehension of an interaction. Branigan et al. [1] distinguish between conscious and unconscious alignment which both can occur in one interaction. For instance, people tend to repeat the things they have just said or heard and to align to single words (lexical alignment), sentences and structures (syntactical alignment), and to nonverbal features of speech like speech rate, pauses and pronunciation (for a review cf. [1]). Also of great importance is the linguistic alignment in the context of geographical and cultural conditions. With regard to certain dialects, people can align because of sympathy or integration, or they can refuse to align because of antipathy [6]. Till now it is underresearched how we can exploit this social aspect of language to enhance learning situations. In sum, linguistic alignment in HHI has different facets and includes variables which are not easily controllable. Human-computer interaction (HCI) and human-agent interaction (HAI) enable researchers to control these variables and facilitate systematic investigations which might lead to a better understanding of the processes and effects of linguistic alignment [15].

1.3 Linguistic Alignment in Human-Computer Interaction

That humans align with their interaction partner is also observable in HCI and HAI. In HCI research it was found that people aligned with regard to speaking pauses, sentence structures, speech rate, the usage of personal pronouns, choice of words and an emphasized articulation towards a computer-system (cf. [1] for a review). Branigan et al. [16] showed in a series of experiments that people tend to align even more to computers than to humans. Moreover, the expertise of the computer systems plays an important role, since it was found that participants adapted more strongly to computers that were presented as less capable than to computers that were presented as more capable. With regard to virtual agents, it has also been shown that humans align with their virtual interlocutor. People mimicked the agent's smiling behavior [2], aligned their amount of words during an interview [3], and showed lexical alignment [10]. The aspect of dialect has not been addressed in HAI so far, although some researchers examined speech inconsistencies and variations like interruptions, repetitions [17] or the usage of fillers like "uh" [18] and the alignment regarding these variations.

1.4 Research Questions

In the present study we explore whether participants linguistically align to a High German or dialect-speaking agent and how alignment affects the evaluation of the interaction. Based on previous findings on linguistic alignment in HAI we expect that in our study participants will also align to the virtual agent in their use of dialect or High German, respectively. We thus hypothesize that H1) participants, who are interacting with a dialect-speaking agent, use more dialectical words than participants who are interacting with a High German-speaking agent. Taking participants' natural tendency to use dialectal words into account measured by a previously assessed baseline, we hypothesize that H2a) participants in the dialect condition will use more dialectal words during than before the interaction and H2b) participants in the High German condition will use fewer dialectal words during than before the interaction.

Against the background of similarity attraction [4] and the "we-feeling" of "dialectal in-groups" described by Bichel [5] and the assumption that people are motivated to talk to others, who are similar to themselves, by appearance, origin or language, the following hypotheses are posited: H3a) The dialect-speaking agent will be rated more likable than the High German-speaking agent and H3b) participants will report higher social presence when interacting with the dialect-speaking agent compared to the High German agent. Moreover, we expect that H4) the dialect-speaking agent will be rated more positively by those participants who more frequently use dialect themselves compared to those who use dialect less frequently and that H5) the conversation with the dialect-speaking agent is rated more positively than the conversation with the High German-speaking agent.

Furthermore, we expect that our system might lead to a better tutoring and, regarding the coverstory mentioned below, is successful in reducing test anxiety. Thus we expect that H6) participants will report less test-anxiety after the interaction with the system than before.

2 Method

2.1 Experimental Design and Independent Variables

We used a one-factorial between subjects design with either a High German- or dialect-speaking agent. The utilized Rhine-Ruhr regional dialect is one of the most common in Northern Germany and is strongly influenced by the industrialization and the polish migration. Possible modifications are short-cuts of pronouns (e.g. „sie“ = „se“) or articles (e.g. “ein” = “en”) as well as different pronunciations (e.g. “Alltag” = “Alltach”). In the present study, we varied the usage of those pronouns, articles and prepositions. For every sentence we altered two or three words from standard High German to dialect. We chose a female avatar (cf. Fig 1) by the Charamel Company (www.charamel.com), which was controlled by the investigator (Wizard-of-Oz-scenario). The nonverbal behavior of the agent (e.g. idle behavior like blinking and posture shifts, gestures) has been kept constant between conditions.

Fig. 1. The virtual avatar Gloria

2.2 Dependent Variables

The dependent variables used in this experiment were the *linguistic alignment of the participant, the perception of the agent, the social presence, the evaluation of the interaction in general* and *the situational test-anxiety of the participant.* These variables were measured by quantitative analysis (online questionnaires before and after the interaction, verbal behavior during the interaction).

Linguistic Alignment. Participant’s verbal behavior before and during the interaction was recorded and transcribed. The used dialectical words were counted and dialect ratios were calculated (dialectical words/total of used words) for each participant for the interaction with the agent and for a previously surveyed baseline.

Perception of the Agent. The likability of the agent was measured using a semantic differential with 8 bi-polar pairs of adjectives (friendly-unfriendly, likable-unlikable, pleasant-unpleasant, honest-dishonest, nice-mean, warmhearted-cold, compassionate-unconcerned, committed-uncommitted) which were rated on a 5-point Likert-scale (Cronbach's $\alpha = .844$). In addition, we assessed as how pleasant the interaction with the tutor was evaluated. Participants rated on a 5-point Likert-scale from "displeasing" to "pleasing".

Social Presence. Social presence was assessed with two subscales from Nowak and Biocca's [19] social presence questionnaire: the subscale *perceived other's copresence* with 12 items (Cronbachs's $\alpha = .713$) and the subscale *self-reported copresence* with six items (Cronbach's $\alpha = .578$) were measured on a 5-point Likert-scale (from "agree" to "disagree"). The scale *self-reported copresence* was excluded from further analysis due to low internal consistency.

Evaluation of the Interaction. For the evaluation of the interaction with the virtual agent we used the *positive-evaluation*-scale with six items (e.g. "It was interesting to communicate with the tutor") rated on a 5-point Likert-scale (from "disagree" to "agree", Cronbach's $\alpha = .882$).

Situational Test-anxiety. According to the cover story (see below), the participants were asked about their current feeling in this test-situation [20] before and after the interaction. To determine test-anxiety, a 7-point Likert-scale was used with eight adjectives (afraid, excited, uncertain, worried, tense, scary, fearful, and nervous; Cronbach's $\alpha = .960$).

2.3 Participants and Procedure

A total of 47 persons (24 female) participated in the study with a mean age of 22.9 years (SD=1.78) ranging from 20 to 28 years. They were recruited via general advertising on campus and online advertising. The study was announced as an evaluation of new software that should help students to reduce test-anxiety in oral exams. To be included in the study participants had to fulfill the precondition of growing up in the Rhine-Ruhr-region. Upon arrival participants signed informed consent and received instructions. First, they filled in the pre-questionnaires (demographics, personality and situational test-anxiety) while the investigator ostensibly left the room in order not to disturb the participant during interaction. However, she retreated to another room from which she could control the interaction using pre-built scripts in a Wizard-of-Oz-scenario guaranteeing the same interactions in both conditions. After finishing the questionnaire participants put on a head-set and started the interaction on a second computer screen by saying "start". To generate a baseline for their natural usage of dialectal words, participants were first requested to talk three minutes about a random topic (e.g. the first week in university) in order to "adapt the computer to their speaking-characteristics". Subsequently, the interaction with the agent started which was divided in two sessions. In the first session, participants were told to play a word-understanding game with the agent in order "to train the agent to their voice". In this game they should move different objects (e.g. box, paper) according to the 15 instructions of the agent (e.g. "Open the box. Put the stuffed animal into the box…"). After every move, participants should describe what exactly they were doing. During the second part of the interaction, the agent reported about the diseases Diabetes and Alzheimer (short version of [10]) and explained the clinical picture, the causes and effects of each disease. Directly after all of the 14 explanations, it asked questions about the recently specified content simulating a test situation. We used only open questions to avoid yes/no-answers. After the interaction, the participants had to fill in the second part of the questionnaire (situational test-anxiety, evaluation of the agent and interaction) after which they were fully debriefed and thanked for their participation.

3 Results

Linguistic Alignment. To examine whether participants who were interacting with a dialect-speaking agent used more dialectical words than participants interacting with a High German-speaking agent (H1), we conducted a one-way ANOVA with language usage of the agent as independent and dialect usage of the participant during the interaction as dependent variable. As expected, we found a main effect of the agent's language use (High German vs. dialect). Participants who interacted with the dialect-speaking agent used more dialectical words than participants who talked with the High German-speaking agent (cf. Table 1). To take into account the natural dialectical word usage of the participants, we conducted an analysis of covariance with the dialectical usage of the baseline as the covariate. There was also a significant effect of the agent's language usage on participants' usage of dialectal words after controlling for participants' usual dialectal usage ($F(2; 45) = 18.969$; $p = <.001$; partial $eta^2 = .463$). Furthermore, the differences within the groups before and during the interaction were investigated and main effects were found for both groups (H2a & H2b). The one-way ANOVAs revealed that participants who talked with the High German-speaking agent used fewer dialectical words during the interaction than before (cf. Table 2). Participants interacting with the dialect-speaking agent used more dialectical words during the interaction than before.

Perception of the Agent. To test whether the dialect-speaking agent was perceived as being more likable (H3), we conducted a one-way ANOVA with the language use of the agent as independent variable and the likability of the agent as dependent variable. As expected, the dialect-speaking agent was evaluated significantly more likeable than the High German-speaking agent (cf. Table 1). A one-way ANOVA was calculated with the dependent variable pleasantness of the interaction which revealed no significant effects. There was no correlation of the likability of the dialect-speaking agent and the own natural dialect usage (H4).

Social Presence. In order to examine whether the dialect-speaking agent elicits more social presence, a one-factorial ANOVA was calculated which showed no significant effect.

Perception of the Interaction. A one-way ANOVA with the *positive-evaluation*-scale as dependent variable and agent's language usage as independent variable revealed a main effect contrary to H5: The interaction was evaluated more positive when talking to the High German- than to the dialect-speaking agent (cf. Table 1).

Situational Test-anxiety. The test-anxiety mean values indicate a lower situational test-anxiety after the interaction with the system (Before: M=3.16, SD=1.634; After: M=2.88, SD=1.577). However, the one-way ANOVA with repeated measurement for the factor situational test-anxiety showed no significant effect.

Table 1. ANOVA - Means and standard deviations for the used dialectical words during the interaction, the likability of the agent, and the positive evaluation of the interaction

	High German *M (SD)*	Dialect *M (SD)*	F	p	ηp²	df
Used dialectical words in game	4.97 (3.97)	9.91 (6.26)	10.570	.002	.190	1
Used dialectical words in test	7.58 (2.11)	10.49 (2.05)	22.895	<.001	.337	1
Used dialectical words total	6.39 (1.93)	10.31 (3.32)	24.841	<.001	.356	1
Likability	2.35 (.72)	2.73 (.53)	4.058	.050	.083	1
Positive evaluation of interaction	3.37 (.84)	2.78 (.86)	5.220	.027	.104	1

Table 2. ANOVA - Means and standard deviations for the used dialectical words before and during the interaction

Condition	Baseline *M (SD)*	Interaction *M (SD)*	F	p	ηp²	df
High German agent	11.23 (3.082)	6.39 (1.931)	59.369	.000	.721	1
Dialect agent	8.43 (3.237)	10.32 (3.318)	7.334	.013	.250	1

4 Discussion

We presented an experimental study examining the effects of dialect in HAI regarding linguistic alignment, the likeability of the agent and the perception of the interaction. Linguistic alignment was found in both conditions. People who talked with the High German-speaking agent used fewer dialectical words, those who talked to the dialect-speaking agent used more dialectical words than usually. Our results support existing findings which have shown linguistic alignment in HCI for instance by Branigan et al. [1] and Brennan [13]. Moreover, the results are in line with the findings of Ferguson [21] and Giles [6] indicating that alignment in pronunciation and accent is used to guarantee a better understanding. Comparing both conditions, people who talked with the High German-speaking agent linguistically aligned to a greater extent than people who talked to the dialect-speaking agent. This is a quite intuitive result, because in general dialect-speaking Germans are used to adapt to High German when speaking with Germans not from their dialectal region. An explanation for the distinct alignment in the High German condition might be the possible weak manipulation of the dialect-speaking agent, who spoke a rather light version of the Rhine-Ruhr dialect. A stronger dialect might have revealed stronger alignment in the dialect condition, but this is uncommon in an academic setting and rather used in rather socially disadvantaged classes. However, all effects were persistent when controlling for participants' usual dialect usage (assessed by a baseline). Besides the actual alignment in participants' behavior, we assumed that dialect adds a social component to the interaction resulting in positive social effects. Indeed, the dialect-speaking agent was rated more likable than the High German-speaking agent. This supports the assumption that using dialect creates a "we-feeling" [5] between the agent and the participants. In accordance with the similarity-attraction theory [4] this shows a higher positive attitude towards the in-group agent. Although the interaction with both agents was rated as rather pleasant, dialectical usage seems to have no significant effect concerning the perception of the pleasantness of the interaction. Furthermore it seems that the mentioned similarity attraction and the "we-feeling" also have no significant effect towards the perception of social presence. In this context, another explanation could be a perceived discrepancy between behavior and appearance. Perhaps, the serious appearance of the agent didn't fit to the rather chummy verbal behavior as which the Rhine-Ruhr dialect might have been perceived. Consistency across modalities and other factors (like behavior and appearance), however, is an important factor for successful communication [22,23]. Surprisingly, the interaction itself was evaluated more positively when interacting with the High German-speaking agent. This might be due to the test situation. In academia, people are used to be in test situations with High German conditions. Thus, it might be confusing to be questioned by a dialect-speaking interlocutor, who, moreover, is dressed in a business suit. With regard to the assessed test-anxiety, no significant decrease in test-anxiety after the interaction was found, although the descriptive data indicate a reduction of test-anxiety. A considerable reduce of test-anxiety is presumably only possible when participants use the system more frequently [10]. Moreover, the test situation was interweaved with tutoring sessions where the agent presented information (in contrast to real oral exams) and participants are aware that the situation is fictitious and has no real consequences.

Limitations and Perspectives. One limitation of this study is that participants showed a significantly different dialect usage between the two groups before the interaction. This limitation could be avoided, for instance, by additional containment of the examined dialectal region or questions about the intensity of dialect usage. Another possibility is to divide the experiment into two sessions and assign participants in the second experimental session based on results of the previously assessed baseline. Additionally, it has to be mentioned that the structure of the conversation between participant and agent was highly structured to some extent. That is why it could be interesting to analyze free speech of the participants. Regarding the assumption of Branigan et al. [1] that alignment occurs even more in HCI than in HHI, it might further be interesting to compare the dialectical alignment with a HHI condition to draw conclusions on the different degrees of alignment in HCI and HHI. Moreover, different dialects or lighter and stronger versions of the same dialect could be used to determine the effects of this social language component and examine how virtual tutoring systems can benefit from using dialect. In this context the durability and sustainability of alignment and the effects on evaluation over a longer period of system usage are further interesting fields of examination. This long-term evaluation would also be interesting for the effects on reducing test-anxiety. With respect to the tutoring system itself, some suggestions for improvement can be proposed. The described context of the learning situation and the agent's appearance should be taken into account. While it seems to be appropriate for a virtual examiner to speak High-German and be dressed rather formally, a tutoring agent which solely supports learning might benefit from a more informal appearance and the use of dialect. Another limitation was the rather restricted mimic feedback of the agent, which is meaningful for social interaction with an agent [24]. Moreover the effect of linguistic alignment (dialectical and standard) on knowledge transfer should be tested in future work in order to design an effective tutoring system.

References

1. Branigan, H.P., Pickering, M.J., Pearson, J., et al.: Linguistic alignment between people and computers. Journal of Pragmatics 42(9), 2355–2368 (2010), doi:10.1016/j.pragma.2009.12.012
2. Krämer, N.C., Kopp, S., Becker-Asano, C., et al.: Smile and the world will smile with you - The effects of a virtual agent's smile on users' evaluation and behavior. International Journal of Human-Computer Studies 71(3), 335–349 (2013), doi:10.1016/j.ijhcs.2012.09.006
3. von der Pütten, A.M., Hoffmann, L., Klatt, J., Krämer, N.C.: Quid Pro Quo? Reciprocal Self-disclosure and Communicative Accomodation Towards a Virtual Interviewer. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 183–194. Springer, Heidelberg (2011)
4. Aronson, E., Wilson, T.D., Akert, R.M.: Social psychology, 7th edn. Prentice Hall, Upper Saddle River (2010)
5. Bichel, U.: Problem und Begriff der Umgangssprache in der germanistischen Forschung. Max Niemeyer Verlag, Tübingen (1973)
6. Giles, H.: Accent mobility: A model and some data. Anthropological Linguistics 15, 87–105 (1973)
7. Iacobelli, F., Cassell, J.: Ethnic Identity and Engagement in Embodied Conversational Agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D., et al. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 57–63. Springer, Heidelberg (2007)
8. Moreno, R., Flowerday, T.: Students' choice of animated pedagogical agents in science learning: A test of the similarity-attraction hypothesis on gender and ethnicity. Contemporary Educational Psychology 31(2), 186–207 (2006), doi:10.1016/j.cedpsych.2005.05.002
9. Krämer, N.C., Bente, G.: Personalizing e-Learning. The Social Effects of Pedagogical Agents. Educ. Psychol. Rev. 22(1), 71–87 (2010), doi:10.1007/s10648-010-9123-x
10. Rosenthal-von der Pütten, A.M., Wiering, L., Krämer, N.C.: Great minds think alike. Experimental study on lexical alignment in human-agent interaction. i-com (in press)
11. Fekeler-Lepszy, E.: Gesprochene Sprache im Ruhrgebiet. K. Farin & H.-J. Zwingmann (1983)
12. Garrod, S., Anderson, A.: Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. Cognition 27(2), 181–218 (1987), doi:10.1016/0010-0277(87)90018-7

13. Brennan, S.E.: Lexical entrainment in spontaneous dialog. In: Proceedings of the 1996 International Symposium on Spoken Dialogue, ISSD 1996, pp. 41–44 (1996)
14. Giles, H., Coupland, N., Coupland, J.: Accommodation theory: communication, context, and consequence. In: Giles, H., Coupland, J., Coupland, N. (eds.) Contexts of Accommodation: Developments in Applied Sociolinguistics, pp. 1–68. Cambridge University Press, Cambridge (1991)
15. Kuhlen, A.K., Brennan, S.E.: Language in dialogue: when confederates might be hazardous to your data. Psychonomic Bulletin & Review 20(1), 54–72 (2013)
16. Branigan, H.P., Pickering, M.J., Pearson, J., et al.: The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. Cognition 121(1), 41–57 (2011), doi:10.1016/j.cognition.2011.05.011
17. Brennan, S.E., Schober, M.F.: How Listeners Compensate for Disfluencies in Spontaneous Speech. Journal of Memory and Language 44(2), 274–296 (2001), doi:10.1006/jmla.2000.2753
18. Pfeifer, L.M., Bickmore, T.: Should Agents Speak Like, um, Humans? The Use of Conversational Fillers by Virtual Agents. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H., et al. (eds.) IVA 2009. LNCS, vol. 5773, pp. 460–466. Springer, Heidelberg (2009)
19. Nowak, K.L., Biocca, F.: The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. PRESENCE: Teleoperators and Virtual Environments 12(5), 481–494 (2003), doi:10.1162/105474603322761289
20. Jacobs, B.: Fragebogen zur aktuellen Prüfungsangst: 288 (2010)
21. Ferguson, C.: Toward a characterization of English foreigner talk. Anthropological Linguistics 17(1), 1–14 (1975)
22. Nass, C., Gong, L.: Maximized modality or constrained consistency. In: Proceedings of the AVSP 1999 Conference, pp. 1–5 (1999)
23. Goetz, J., Kiesler, S.B., Powers, A.: Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation. In: Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2003), pp. 55–60 (2003), doi:10.1109/ROMAN.2003.1251796
24. Nass, C., Isbister, K., Lee, E., et al.: Truth Is Beauty: Researching Embodied Conversational Agents. In: Cassell, J. (ed.) Embodied Conversational Agents, pp. 374–402. MIT Press, Cambridge (2000)

A Virtual Patient to Assess Pediatric Intensive Care Unit (PICU) Nurses' Pain Assessment and Intervention Practices

Sangyoon Lee[1], Cynthia M. LaFond[2], Andrew E. Johnson[1], Catherine Vincent[2], Jason Leigh[1], and Luc Renambot[1]

[1] Dept. of Computer Science, University of Illinois at Chicago, IL, USA
{slee14,ajohnson,spiff,renambot}@uic.edu
[2] College of Nursing, University of Illinois at Chicago, IL, USA
{clafon2,vincentc}@uic.edu

Abstract. Pediatric intensive care (PICU) nurses play a crucial role in managing children's pain. While virtual patient (VP) technology has shown significant benefits in assisting with the practice of healthcare, there has been little research effort in this particular setting. This paper presents a pilot evaluation study to determine the validity of VP vignettes - including PICU nurses' recognition of the facial expressions (smiling and grimacing) of VPs and nurses' descriptions as to whether the VP vignettes are consistent with their professional experiences. The results of our initial study (n=20) confirm that nurses identified given expressions correctly (98.5%) and validated the similarity of the vignettes to their experiences with patients in the PICU (87%). A better understanding of nurses' pain management practices will aid the development of future VP interventions.

Keywords: Virtual Patient Vignette, Pain Assessment, Intervention Practice, Pediatric.

1 Introduction

The use of a virtual patient (VP) in healthcare and clinical settings has been actively studied. VPs provide robust tools for training medical professionals [1,2] and for examining how patient profiles influence pain assessment [3,4]. This is especially true [5,6] when the VP is deployed in conjunction with a traditional written vignette that provides a medical context in which to study disparities in nurses' pain-related decision making. VPs enhance written vignettes by providing a consistent stimulus from which nurses respond, likely improving the uniformity of data [7]. Furthermore, variables (such as facial actions) among VPs can be more carefully controlled than video among actual patients or actors, especially in the case of school-age children, as fraudulent pain expressions are easily detected in this age group [8]. Additional criticism of the written vignette includes its simplicity and decreased demand for interpretation of information [7]. To better simulate nurses' workflows and usual visual methods of

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 159–167, 2013.

obtaining patient information, our work incorporated VPs as well as video and simulated patient electronic medical records to replace all written vignette information (e.g., vital signs, medication orders). These VP vignettes were piloted with pediatric intensive care units (PICU) nurses in a mixed-methods study to evaluate the effect of child behavior and pain type upon their pain assessment and intervention choices. Uncontrolled pain is the second most common adverse event in PICUs in the United States [9,10]. Nurses have a crucial role in managing children's pain; a better understanding of their pain management practices will aid the development of future VP interventions. However, to date VP efforts have been developed for nurses working with adults. Our study aims to validate the VP vignette with respect to children's pain assessment - including PICU nurses' recognition of specific facial expressions and descriptions of the vignettes consistency with their professional experiences. We will also discuss the feasibility of implementing a VP vignette within a hospital setting. The rest of this paper is organized as follows: Section 2 reviews studies related to our work; Section 3 describes the PICU vignette scenarios and the implementation details; Section 4 presents the user study, preliminary results, and a discussion; and, Section 5 presents our conclusions.

2 Related Work

Among the disciplines that can benefit from virtual human (VH) technology, healthcare and clinical settings have been widely studied; in fact, the use of VP models to train healthcare professionals is a popular topic in the literature. Mehra and Lee analyzed early examples of VHs in healthcare and illustrated the fundamental principles and the architecture behind VHs [11]. A more specific example of VP use is found in [1], a VP for clinical therapists. The study showed how VPs could help novice clinicians practice interviewing and diagnosing skills. Conversely, Pontier and Siddiqui presented a virtual therapist preferred by users responding to questionnaires about depression [12].

In the literature on emotional expression research, Ekman's classic study [13] has been extensively adopted to model human emotions in VHs. His Facial Action Coding System (FACS) and Action Units (AU) provide foundations to analyze facial muscle movement in the categorical emotions; the resulting empirical data can be used to develop more realistic emotions in VH interfaces. To this end, Hirsh et. al. used FACS to realize pain expressions in their VP models, used to evaluate disparities in pain assessments amongst patients with variations in sex (male or female), race (Caucasian or African-American), and age (young or old) [14]. They modeled the VH pain expressions with the following AUs: brow lowerer (AU4), cheek raiser (AU6), eyelid tightener (AU7), nose wrinkler (AU9), upper lip raiser (AU10), and eye closure (AU43). This systematic composition of facial expressions offers a more controlled uniform realization method for VPs. Our study relies on the same principles to implement the grimace expression; however, we also iteratively validated our model with a FACS certified expert and our study is specifically designed for children's pain assessment (10-years-old) within the PICU environment.

3 Virtual Patient Vignette

Our VP scenario is derived from the Pain Beliefs and Practices Questionnaire (PBPQ) case studies [15,16]. The written PBPQ cases present 10-year-old boys who had abdominal surgery a day before, have a self-assessed pain rating of 8 on a 0-10 numeric pain scale, and have stable vital signs. One boy smiles and jokes whereas the other grimaces. Nurses are asked to rate each child's pain from 0 to 10 and choose the analgesia dose they would provide (if any). Because of the subjective nature of pain and the child's verbalization of pain poorly relieved from a prior dose of morphine, the response considered most correct for these vignettes included an agreement with the child's pain rating of 8 and increasing the morphine dose (regardless of the child's facial expression or diagnosis). We adapted these scenarios for a VP vignette and included two additional patients experiencing pain from a sickle-cell vaso-occlusive crisis (one smiling, one grimacing). The adapted vignettes were reviewed by content experts, advanced practice nurses with expertise working with critically ill children.

Fig. 1. Virtual Patient Head Models

The LifeLike Responsive Avatar Framework (LRAF) [17] is used to realize a VP. First, we made 4 patient models from photos of 9-11 year-old boys to simulate the PBPQ scenario. Second, we designed facial expressions (smile and grimace) with a FACS certified expert. Last, the VP vignette application was implemented. We created a child model for the first time. All our previous VH designs were based on people with ages ranging from 20-60 years old. We were able to generate models that looked close enough to the target children via our design pipeline and a head generation method (Figure 1).

The grimace expression is based on FACS and observed facial movements, or FACS AU [18]. The initial set of AUs selected for the grimacing patient was brow lowerer (AU4), cheek raiser (AU6), nose wrinkler (AU9), and mouth stretcher (AU27). In addition, we applied dynamic facial wrinkles to the facial expressions. During our iterative review session with a FACS coder, she confirmed that the wrinkle feature enhanced the quality of the perceived expression. The smile expression is selected from one of LRAF's basic expression templates (open-smile) together with slight lip stretcher (AU20) and lips part (AU25). In the final configuration, we chose AU 4, 6, 7, 9, 20, and 25 with a wrinkle intensity 0.6 at

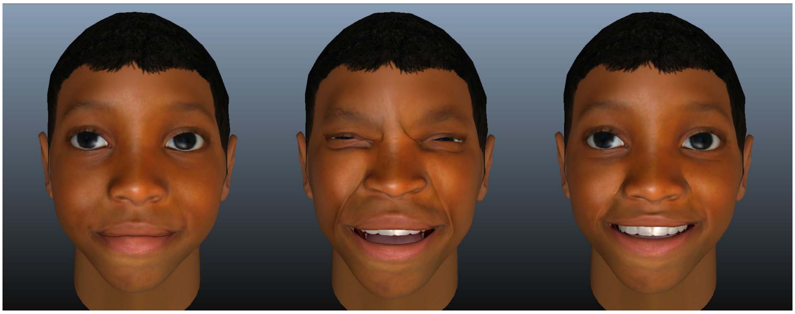


Fig. 2. Virtual Patient Facial Expression (left: neutral face; center: grimace; right: smile)

the peak of grimace animation (Figure 2. center) to achieve the desired facial expression. A wrinkle intensity is 0.7 is used for the smile expression (Figure 2. right).

Due to limited hardware accessibility on-site (hospitals), we implemented a web application with recorded videos (15 seconds each). Figure 3 illustrates the deployed application. In the main GUI, a subject can navigate menu buttons to see patient record information, VP animation, patient monitor (current vital signs), the medication administration record, and the patient's self-assessed pain level.

Fig. 3. Pain Vignette Application (left: main GUI; center: virtual patient; right: patient monitor)

4 Pilot Study

The goal of the study is to validate our VP vignettes with the PICU population. It includes PICU nurses' recognition of the intended facial expressions as well as their descriptions of the vignettes consistency with their professional experiences. We present preliminary results from a mixed-methods study in which PICU nurses' beliefs regarding children's pain and their simulated pain assessment and intervention practices were evaluated; of specific interest for the larger study was the effect of child behavior (smiling or grimacing), pain type (sickle cell vaso-occlusive crisis or abdominal surgery), and vignette type (written or virtual) upon the nurses' pain assessment and intervention choices.

4.1 Design and Procedure

For our VP vignettes, we replaced all written PBPQ vignette information with digital media, such as animated VPs, simulated vital signs, simulated electronic patient flow-sheets, and medication orders. Patient information and medication administration records were created to mimic computer screens that nurses use in their daily work. Vital signs were captured from a simulated patient monitor. It is a 30-second video clip showing stable vital signs with some mild variability in heart rate and oxygen saturation (BP: 102/60, HR: 80, R: 20, SpO2: 98%). The VP animation shows the bust of a child and zooms in to a close-up of his face. For the smiling vignette, the VP looks left, smiles, turns back to front, and smiles again. Smiling while looking left simulates joking with visitors (no audio or visitors' visual representations given). In the grimacing vignette, the VP illustrates the same head motion as the smiling VP whereas the grimacing VP only grimaces when turning his head forward. These movements were chosen to maintain consistency with the written PBPQ vignette descriptions of patients. Both VP videos have the same length, 15 seconds.

Procedure: Approval was obtained from the Institutional Review Boards. Once consent was obtained and a demographic questionnaire was completed, each subject was shown five vignettes with a different model (one practice model and four test models). At the end of each session, the subjects were asked to rate each child's pain (0-10) and to determine corresponding interventions. To avoid an order effect, the sequence of the test vignettes was randomly assigned. Semi-structured interviews followed the completion of all vignettes to elicit details regarding the subjects' choices. For each of the vignettes, nurses were asked what they were thinking about as they rated the child's pain and made a decision regarding pain intervention. After discussing all four vignettes, the PICU nurses were encouraged to share any additional information regarding assessing and intervening for children's pain. Finally, they were asked how consistent the vignettes were with their experience as a PICU nurse. Transcriptions of the interviews were analyzed using qualitative content analysis [19].

Participants: A convenience sample of nurses was recruited from PICUs at two hospitals in the Midwest, a large children's hospital and a university-based hospital. Of the first 20 PICU nurses to participate in the study, 85% were female, 70% identified their race as white, 20% Asian, 5% Black or African American, and 5% Hawaiian or Pacific Islander, and 5% identified their ethnic category as Hispanic or Latino. The nurses' years of PICU nursing experience ranged from 1-29 years, with a mean of 12.1 years (SD 9.2). All of the nurses described caring for children experiencing pain in the past 3 months, with the majority (60%) reporting to care for 1-5 children in pain each week. The remaining 40% reported caring for children in pain more frequently (30% 6-10 per week, 10% more than 15 per week).

4.2 Results and Discussion

We sought two measures: (1) did subjects correctly identify facial expressions presented; and, (2) did subjects find the vignette consistent with their professional practice? Additionally, we sought to determine (3) how feasible is the use of VP vignettes as an instrument to elicit responses from nurses within the hospital setting?

Identification of Facial Expressions: The semi-structured interviews were analyzed to determine nurses' recognition of the intended facial expressions of the VPs. Though the nurses were not asked to identify the facial expressions of the VPs, many of them discussed the VPs' facial expressions while describing pain assessment and intervention choices. When facial expressions were described, the nurses identified a VP's intended facial expression 98.5% (64/65) of the time. When discussing the smiling patients, nurses used the word smile or smiling for 77.5% of the VP viewings (31/40); 10% (4/40) of the expressions were described with terms such as happy. For the remaining 12.5% (5/40), the patients' facial expression was not discussed some of these nurses reported that facial expressions were not an important aspect of their pain assessment. When discussing the two grimacing patients, 69% of the time (27/39 VP viewings) nurses used the words grimace or wince to describe the patients' facial expressions. Two times, nurses referred to the expression as showing pain and distressed. One nurse used the word smile while describing a patient with a grimace (1/39). However, it is unclear if this response was a misidentification of the facial expression or an inaccurate recollection of the patient being discussed.

Consistency of Vignette with Professional Practice: At the end of each interview, nurses were asked if the VP vignettes were consistent with their professional experience as a PICU nurse. Eighty-seven percent of the nurses (15/18, two audio recordings were interrupted) validated the similarity of the vignettes to their experiences with patients in the PICU. Comments provided included *"I've seen patients like this before," "I didn't see anything here I haven't seen a million times already," and "It's good. It exemplifies real life as far as what you see with our patients."* The remaining three nurses (17%) did not specifically confirm consistency of the vignettes, but instead elaborated on their personal experiences with patients in pain. Though the majority of nurses confirmed the vignette's accurate depiction of patients, some discrepancies were identified with the medication order (30%), and the lack of vital sign changes with pain (25%).

Feasibility

(a) Time and expense: Differences in the facial structures of the boys in the photos led to inconsistencies in the intensity and presence of some of the AUs, requiring additional manipulation of each VP to achieve equivalence. Additionally, some of the predefined AUs did not incorporate the response of other areas of the face to the movement of the action unit. This was especially problematic for the lower face leading to the need to engage additional action units to accomplish the originally intended AU. The use of a FACS coder is essential if realistic expressions are desired, especially in the case of a research study in which findings rely upon the participants' response to the expressions. These additional steps and the involvement of a consultant add to the expense and the time until completion of a VP vignette.

(b) Practicality: As previously mentioned, web-based vignettes allowed for flexibility in the locations in which subjects could meet the principal investigator. All 20 of the nurses in this study preferred to meet the principal investigator before or after their shifts at the hospital where they worked. The only required equipment was a laptop; however, we chose to use a 24-inch LCD monitor so that the nurses could better view the vignettes and a wireless mouse to ease their navigation.

(c) Usability: As previously mentioned, the nurses were introduced to a practice vignette prior to viewing and responding to the study vignettes. All of the nurses became proficient with navigating the practice vignette in a short amount of time (roughly 3-5 minutes) and completed viewing all four test vignettes, rating the VP pain level and documenting interventions, in generally 20 minutes or less.

5 Conclusion and Future Wrok

In this research, we designed a VP vignette for PICU nurses' pain assessment based on PBPQ case studies. The written PBPQ scenario was automated using VP animations, videos of vital signs, and images of patient information and medication orders. We created a VP that targeted school-age children and looked close enough to convey PBPQ scenarios. The VP vignettes were easily ported to multiple study sites and the PICU nurses were able to quickly navigate and assimilate the information within the vignettes. The main barriers to developing the VP vignettes were the need for a certified FACS coder and the multiple revisions required to achieve the desired expressions.

Pilot study results suggested that nurses identified given expressions correctly and validated the similarity of the vignettes to their experiences with patients in the PICU. The nurses' request for more information regarding the patients, including the ability to further assess the patients' pain, supports their engagement in the vignettes, and speaks to the potential usefulness of interactive VP vignettes for future research and training. We are currently completing data collection and plan to further analyze data to compare the use of a VP to elicit responses from PICU nurses' regarding their pain assessment and intervention practices to the classical written method.

Acknowledgments. This publication is based in part on work supported by the National Science Foundation (NSF), awards CNS-0420477 and CNS-0703916. This work was also partially supported by Seth and Denise Rosen Memorial Research award, The International Nursing Association for Clinical Simulation and Learning Debra Spunt award, Sigma Theta Tau Alpha Lamba Research award, and University of Illinois at Chicago College of Nursing Student Research award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution.

References

1. Kenny, P., Parsons, T.D., Gratch, J., Leuski, A., Rizzo, A.A.: Virtual Patients for Clinical Therapist Skills Training. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 197–210. Springer, Heidelberg (2007)
2. Rossen, B., Cendan, J., Lok, B.: Using Virtual Humans to Bootstrap the Creation of Other Virtual Humans. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 392–398. Springer, Heidelberg (2010)
3. Stutts, L.A., Hirsh, A.T., George, S.Z., Robinson, M.E.: Investigating patient characteristics on pain assessment using virtual human technology. European Journal of Pain 14(10), 1040–1045 (2010)
4. Wandner, L.D., Stutts, L.A., Alqudah, A.F., Craggs, J.G., Scipio, C.D., Hirsh, A.T., Robinson, M.E.: Virtual human technology: patient demographics and healthcare training fac-tors in pain observation and treatment recommendations. Journal of Pain Research 3, 241–247 (2010)
5. Hirsh, A.T., George, S.Z., Robinson, M.E.: Pain assessment and treatment disparities: a virtual human technology investigation. Pain 143(1), 106–113 (2009)
6. Hirsh, A.T., Dillworth, T.M., Ehde, D.M., Jensen, M.P.: Sex differences in pain and psychological functioning in persons with limb loss. Journal of Pain 11(1), 79–86 (2010)
7. Hughes, R., Huby, M.: The application of vignettes in social and nursing research. Journal of Advanced Nursing 37(4), 382–386 (2002)
8. Larochette, A., Chambers, C.T., Craig, K.D.: Genuine, suppressed and faked facial ex-pressions of pain in children. Pain 126(1), 64–71 (2006)
9. Agarwal, S., Classen, D., Larsen, G., Tofil, N.M., Hayes, L.W., Sullivan, J.E., Storgion, S.A., Coopes, B.J., Craig, V., Jaderlund, C., Bisarya, H., Parast, L., Sharek, P.: Prevalence of ad-verse events in pediatric intensive care units in the United States. Pediatric Critical Care Medicine 11(5), 568–578 (2010)
10. Larsen, G.Y., Donaldson, A.E., Parker, H.B., Grant, M.J.C.: Preventable harm occur-ring to critically ill children. Pediatric Critical Care Medicine 8(4), 331–336 (2007)
11. Mehra, V., Lee, W.S.: Virtual humans in healthcare. In: Proceedings of the 3rd IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications (2004)
12. Pontier, M., Siddiqui, G.F.: A Virtual Therapist That Responds Empathically to Your Answers. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 417–425. Springer, Heidelberg (2008)
13. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measure-ment of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
14. Hirsh, A., Alqudah, A.F., Stutts, L.A., Robinson, M.E.: Virtual human technology: Capturing sex, race, and age influences in individual pain decision policies. Pain (2008)
15. Manworren, R.C.B.: Development and Testing of the Pediatric Nurses Knowledge and Attitudes Survey Regarding Pain. Pediatric Nursing 27(2), 151–158 (2001)
16. Vincent, C., Wilkie, D.J., Wang, E.: Pediatric Nurses Beliefs and Pain Management Practices: An Intervention Pilot. Western Journal of Nursing Research 33(6), 825–845 (2010)

17. Lee, S., Carlson, G., Jones, S., Johnson, A., Leigh, J., Renambot, L.: Designing an Expressive Avatar of a Real Person. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 64–76. Springer, Heidelberg (2010)
18. Wilkie, D.J.: Facial Expressions of Pain in Lung Cancer, Analgesia, vol. 1, pp. 91–99 (1995)
19. Hsieh, H.F., Shannon, S.E.: Three Approaches to Qualitative Content Analysis. Qualitative Health Research 15(9), 1277–1288 (2005)

Controlling the Listener Response Rate of Virtual Agents

Iwan de Kok and Dirk Heylen

Human Media Interaction Group, University of Twente
P.O. Box 217, 7500AE Enschede, The Netherlands
{i.a.dekok,heylen}@utwente.nl

Abstract. This paper presents a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models in predicting generic listener responses for embodied conversational agents. Based on the time since the last generated listener response, the proposed dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. The proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to adapt the behavior to the situation, targeted role or personality of the virtual agent. We show that the generated behavior is more stable under changing conditions than the behavior of the traditional fixed threshold.

1 Introduction

In conversation, both speakers and listeners are active participants. Having a conversation requires complex coordination between verbal and nonverbal behavior to shape the information which is passed on from one interlocutor to the other. This is true for both the interlocutor who is speaking as well as the interlocutor currently listening. The speaker provides the information, while the listener is constantly providing feedback to the speaker by signallying his/her attendance, understanding and/or appraisal. This behavior is an essential part of a succesfull interaction. It has been proven to increase the quality of the speaker's speech [11,1], understanding of the speaker's speech by the listener [11,1] and rapport between the interlocutors [4].

The responses that interlocutors give while listening can be divided into generic and specific listener responses [1] (or alternatively continuers and assessments [3]). Generic listener responses are not specifically connected to what the speaker is saying. One could theoretically interchange two generic listener responses, typically head nods or minimal vocalizations such as "mhm", without having too much of an impact on the flow and meaning of the conversation. They merely function to signal attendence and a general notion of understanding to let the speaker know he/she can continue. Specific listener responses do have a tie to the content of the speaker's speech. They usually give an assessment

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 168–179, 2013.

of what has been said, such as "oh wow!" or gasping in horror, and/or give a specific signal of understanding, such as repeating key words of the speech.

Computational models of listening behavior for embodied conversational agents have also been developed and have shown success in replicating the function they fullfill in everyday conversation [12,10,19,6,18]. For the development of computational models for these and other embodied conversational agents different strategies for both types of listener responses have been considered. For specific listener responses the models have focused on the incremental understanding of the content of the speech [23,22], since these listener responses have strong ties to what is being said. Models for generic listener responses have focussed on more shallow features of the speaker's speech, such as acoustic features [2,24] and eye gaze [13]. The approach that has been utilized was initially handcrafted rules [24], but nowadays the corpus based machine learning approach has proven to outperform these handcrafted rules [13].

These machine learning based models are optimized to match the ground truth labels found in the corpus as closely as possible, usually measured by a F_1-measure [8]. The goal is generally not to copy the behavior that was found in a corpus, but to build an agent that responds to new input. Also, one might want to vary the behaviour of agent. So, when using such models in a conversational embodied agent, other aspects of the generated listening behavior may be more important. When designing a conversational embodied agent the designers usually have a personality or role in mind they want their agent to fulfill. A lot of factors are important when managing the impression a user has about the personality of an agent, one of which is their listening behavior. The timing, amount and form of the listener responses that are produced by an embodied conversational agent have been proven to influence the impression the user has about their personality [20].

Therefore, it is important that the produced listening behavior is consistent with the targeted personality of your embodied conversational agent and is so under every circumstance and for every user. In other words, it is important that the listening behavior that an embodied conversational agent performs is stable, recognizable and conform the expectations the user and designer have of the behavior.

This is typically not provided by the current state-of-the-art models for generating generic listener responses. Changes in conditions, such as different interlocutors with different voice characteristics and speaking styles, can have a big impact on the features used as input by these models, which in turn can have a big impact on the predictions made by the models.

In this paper we will present a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models for predicting generic listener responses for embodied conversational agents. Based on the time since the last listener response the proposed dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. The proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to create the behavior that is desired for the targeted role, personality or situation.

More details about the dynamic thresholding approach are given in Section 2. We evaluate the approach on a corpus in Section 3. We conclude this paper with our final thoughts in Section 4.

2 Interpreting the Prediction Value Curve

The best known models to determine the timing of generic listener responses are corpus-based machine learning models [15,21,14,13,5,16]. These models are learned from an annotated corpus of human-human interactions. From this corpus features are extracted, such as the eye gaze of the speaker and speech features. Based on these features a model is learned that infers the relation between these features and the occurrence of a listener response in the corpus.

The output of such a listener response prediction model is a prediction value indicating the likelihood of a listener response occuring at each time frame. After sequencing and smoothing these prediction values one gets a prediction value curve. In Figure 1 two example prediction value curves are presented, plotted in black (disregard the red line for now). These examples were taken from the first minute of two interactions from the MultiLis corpus and produced by the model we use in our evaluation. More details on the corpus and model follow later.

From these prediction value curves the timing predictions for listener responses can be extracted. This is usually done by detecting peaks in the prediction value curve and comparing these peaks to a threshold. If this peak exceeds the threshold, it is considered appropriate to give a listener response at the time of the peak. Typically the threshold is a fixed value that is determined in the validation phase of the development of the model, which can be decreased or increased to generate more or fewer responses respectively to express more attention or a different personality type. In Figure 1 the threshold that was found to give the highest F_1 score for the model is indicated by the horizontal line at 0.2122.

2.1 Limitations of the Fixed Threshold

However, this way of selecting the threshold has a problem. The amount of listener responses predicted by the model using a certain threshold is inconsistent. The same model using the same threshold applied to two different speakers can result in a significant variation in listener response rate which might be unwanted. This is illustrated by the two prediction value curves in Figure 1. By looking at the number of peaks that exceed the threshold we can see that applying the optimal threshold according to the validation step has resulted in big difference in response rate. For the first interaction nineteen listener responses are predicted in the first minute, while only one is predicted for the second interaction. The explanation for the lower prediction values in the bottom curve

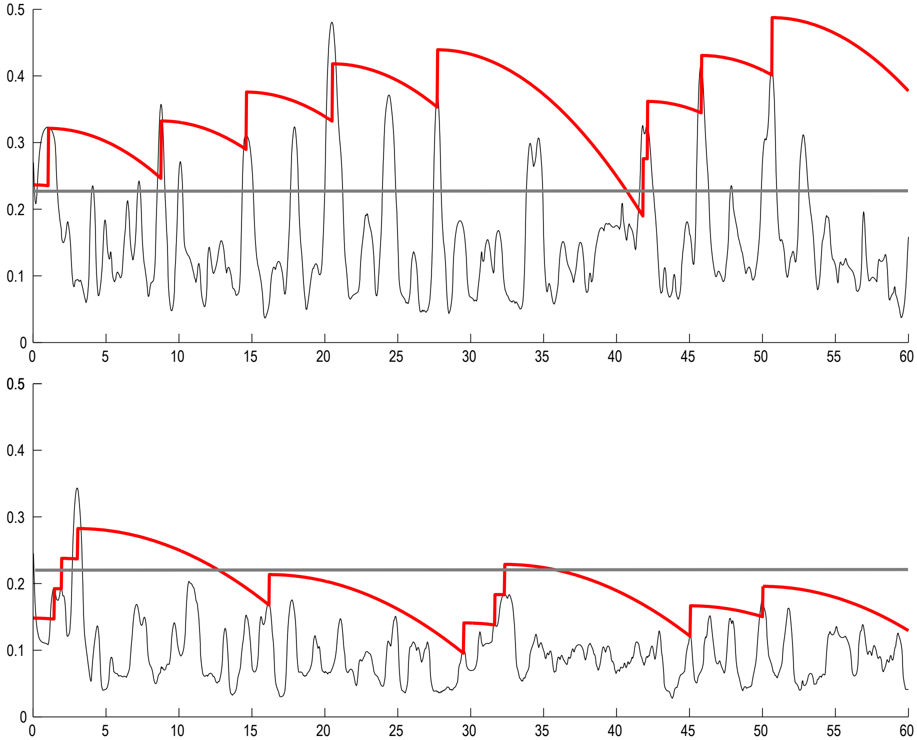


Fig. 1. The prediction value curves of the parallel listener consensus model applied to the first minute of two interactions. On the horizontal axis time is presented in seconds and on the vertical axis the likelihood of a listener response according to the model is presented. The gray horizontal line is the validated threshold obtained during the validation step. In red the proposed dynamic threshold is shown. The threshold start initially high and decreases over time at an increasing rate until a peak exceeds the threshold. Then the threshold is increased by a fixed amount after which the threshold start decreasing slowly again.

in Figure 1 lies in the fact that in this case the speaker hardly ever looked at the listener where gaze is one of the most important cues that the prediction model is using. Even though the speaker does not look often at the listener, opportunities to give a listener response are available, since the listeners in the corpus did respond during this minute. So, it might be in the interest of the virtual listener to give them. Either to comfort and encourage the speaker to continue speaking and/or to built a better rapport with the speaker.

With a fixed threshold this is hard to do, since their is no reliable way of knowing how much one needs to lower the threshold to get the desired response rate. This is because selecting peaks based on a fixed validated threshold is subject to changing conditions. These changing conditions do not limit themselves to eye gaze behavior as in the previous example, but also other aspects of a speaking style can change. Examples include, but are not limited to speaking in a louder or softer voice, with higher or lower pitch, different speech rates or varying degrees in which intonation is used. These aspects can even change during an interaction with the same speaker, not necessarily because the speakers changes these characteristics of their speaking style, but they may also change their position relative to the microphone or video camera. All these things can influence the features used by the model, which in turn influence the prediction values returned by the model. For some conditions the general prediction value from a model will always be lower than for other conditions and the peaks in this curve may remain below the fixed validated threshold.

Essentially, these differing conditions makes it very difficult for a designer of an embodied conversational agent to give the agent the personality and behavior the designer has in mind. For one user the agent may be responsive and attentive, while for another almost no listener responses are generated at all.

So, we need another way of extracting the appropriate timings of listener responses from the prediction value curves. To give the designer of the nonverbal listening behavior of a virtual agent the tools to control the generated behavior, the solution needs to ensure the following characteristics:

– **Stable Response Rate -** The solution should be able to generate s similar response rate under changing conditions, such as different speakers, audio/video quality or feature extraction accuracy.
– **Evenly Distributed Responses -** Not only the overall rate should be stable, but the distribution needs to be even as well. Periods with many or few responses are perceived as unnatural behavior [17].
– **Adjustable Response Rate -** The response rate needs to be easily adjustable, so the designer of the agent can generate the desired behavior and even change this behavior during the interaction.

Next we will present a formula that is designed to adjust the thresholds dynamically according to these principles and we will show that it works.

2.2 Dynamic Thresholding

The solution we propose is to have a changing threshold during the interaction. So, instead of a fixed threshold determined at the development stage of the prediction model, we propose a dynamic threshold that changes over time depending on the time since the last predicted listener response. At the start of the listening period the threshold is relatively high and it starts decreasing at an increasing rate until a peak in the curve exceeds the threshold. When a listener response is predicted the threshold jumps up and starts decreasing again at an initially slow rate. This will ensure both the stable response rate and the even distribution of the responses.

In Figure 1 the dynamic threshold for the two interactions is shown in red. Here we can see that for both the example interactions the number of predicted response opportunities is nine. So, the resulting response rate is much more stable than the response rate the fixed threshold would have given, meaning we have met our first required characteristic. The response are also more evenly distributed. There are no long periods without responses, the longest gap being 14 seconds. There are still issues with two or three consecutive predicted response opportunities, but an extra rule stating no two listener responses to be generated within a certain amount of time would solve this. For the final required characteristic we will take a closer look at the formula that created this dynamic threshold.

$$T_t = T_{t_{last}} + j - (\frac{t - t_{last}}{g.r})^2.\frac{j}{d} \tag{1}$$

The formula that created the dynamic thresholds presented in red in Figure 1 is presented in Equation 1. It calculates the dynamic threshold (T) at time t. Time t is measured in frames. Gap parameter g is the mean time in seconds between two predicted listener responses. Parameter r is the sampling rate of the system, needed to convert timing in seconds into timing in frames. Parameter t_{last} is the time of the listener response that was last generated and $T_{t_{last}}$ is the dynamic threshold at that time. Parameter j is the jump parameter, which represents the amount that the dynamic threshold increases after predicting a listener responses. The final parameter d is the dropoff parameter, which controls the amount the dynamic thresholds decreases over time.

Gap parameter g is the parameter that will give the designer of the agent control over the response rate of the agent by defining the mean gap between two predicted listener responses. However, before this gap parameter will give the expected behavior, the jump parameter j and dropoff parameter d need to be determined. For the jump parameter j we recommend using the standard deviation of the prediction value curve as value. This will make the jumps after predicting a listener response appropriate to the variation found in the prediction value curve and thus make the dynamic threshold even more adaptable to differing conditions. That leaves the dropoff parameter and this needs to be calibrated on a development set of example interactions. The procedure for this is to try different combinations of parameter d and g and minimize the absolute difference in expected number of listener responses and predicted number of listener responses for the values of g you expect the virtual listener to use.

3 Evaluation

In the previous section we have presented our dynamic threshold formula and highlighted its merits on two segments of one minute. In this section we will evaluate the performance of the dynamic threshold over a larger sample size. We will do this on a subset of the interactions of the MultiLis corpus, presented in Section 3.1, using a prediction model, presented in Section 3.2, that is trained on the other subset of interaction of the MultiLis corpus. The procedure of this evaluation is explained in Section 3.3.

To support the dynamic thresholding formula the evaluation will aim to give answers to the following questions: Does the dynamic threshold succeed in stabilizing the response rate? Does it avoid periods with few or many responses?

3.1 MultiLis Corpus

The MultiLis corpus [7] is a Dutch spoken multimodal corpus of 32 mediated face-to-face interactions totaling 131 minutes. Participants (29 male, 3 female, mean age 25) were assigned the role of either speaker or listener during an interaction. In each session four participants were invited to record four interactions. Each participant was once speaker and three times listener.

The interactions of the corpus were between one speaker and three listeners. The three listener were tricked into believing to be the sole listeners, but were recorded in parallel listening to the same speaker. The speakers saw only one of the listeners, believing that they had a one-on-one conversation. To create this illusion all listeners were placed in a cubicle and saw the speaker on the screen in front of them. The camera was placed behind an interrogation mirror, positioned directly behind the position on which the interlocutor was projected. This made it possible to create the feeling of eye contact.

To ensure that the illusion of a one-on-one conversation was not broken, interaction between participants was limited. Speakers and listeners were instructed not to ask for clarifications or to elicit explicit feedback from each other, so no turn-switching would take place. The speaker received a task of either watching a short video clip before the interaction and summarizing it to the listener, or learning a recipe in the 10 minutes before the interaction and reciting it to the listener. The listener needed to remember as many details of what the speaker told as possible, since questions about the content were asked afterwards.

In our evaluations we use ten interactions from the corpus totalling little over 40 minutes. These ten interactions were not used in the development of the prediction model, which will be presented in the following section.

3.2 Prediction Model

As our listener response prediction model we use the best performing model from the paper by De Kok et al. [9]. This model is the "Consensus 2" model. It is a Conditional Random Fields model trained on the other 22 interactions from the MultiLis corpus. As input features it uses the eye gaze of the speaker and several acoustic features of the speech signal, such as pitch, intensity and silence. As ground truth labels the model is trained using only the moments where at least two of the three listeners in the MultiLis corpus have given a listener response.

3.3 Procedure

We applied the listener response prediction model to the ten interactions to obtain the prediction value curve for each interaction. We then applied eleven fixed and eleven dynamic thresholds to these prediction value cur-ves. We varied the fixed thresholds between 0.15 and 0.35. To have dynamic thresholds we can directly compare to the fixed thresholds, we look at the resulting overall response rate of the thresholds. We then set gap parameter g from the dynamic threshold

Table 1. The table illustrates the effect of different fixed thresholds on the response rate in responses per minute of a listener response prediction model. The cells are gray shaded for easier interpretation, with higher response rate being darker. It illustrates that, although increasing the threshold decreases the overall response rate at a predictable pace, the effects on individual interactions varies wildly.

	Response Rate for Fixed Threshold (responses/minute)										
		Interaction									
Threshold	**Overall**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.15	21.6	19.0	23.8	26.1	21.9	12.9	21.8	24.9	24.2	22.7	21.5
0.17	16.8	15.7	20.2	21.6	17.2	6.4	16.7	22.3	19.2	16.4	16.1
0.19	13.2	13.7	16.2	18.6	12.1	3.0	12.0	18.4	15.5	13.4	13.1
0.21	11.5	12.8	14.8	16.7	11.1	2.0	8.8	15.3	13.0	12.5	11.5
0.23	10.1	11.3	13.3	15.7	10.0	1.5	7.2	14.4	11.3	10.8	10.9
0.25	8.7	9.3	11.5	14.1	8.7	1.0	5.5	13.6	9.8	9.6	8.4
0.27	7.1	7.0	10.4	13.4	8.0	1.0	3.5	11.8	8.2	7.2	6.6
0.29	5.5	4.0	8.6	12.4	6.7	1.0	2.9	10.9	6.3	5.3	4.1
0.31	4.4	3.1	6.1	10.1	5.4	0.7	2.1	10.1	5.3	3.8	3.2
0.33	3.1	2.2	4.3	5.9	4.1	0.7	1.9	8.7	2.8	2.8	2.7
0.35	2.4	1.8	2.5	4.9	3.1	0.5	1.5	6.6	2.3	1.3	2.5

formula such that the response rate of that threshold was (almost) the same. For both the fixed and the dynamic threshold we do not allow any listener response within one second of the previous listener response. These predicted listener responses are discarded.

For the dynamic threshold other parameter beside the gap parameter g need to be set as well. For this evaluation we initialized the dynamic threshold with the mean of the prediction value curve as initial $T_{t_{last}}$ and the standard deviation of the prediction value curve as jump parameter j. To select the value for the dropoff parameter d, we tried several combinations of gap parameter g and dropoff paramater d on the ten interactions. We selected the value for dropoff parameter d such that the difference between the expected number of listener responses based on the value for gap parameter g and the resulting number of predicted listener responses was minimized. This was true for value 1.4.

3.4 Results

The first question we will answer is, does the dynamic threshold succeed in stabilizing the response rate? For this we look at the response rates that were the result of applying the eleven fixed and dynamic thresholds on each interaction. These response rates in responses per minute are presented in Tables 1 (fixed thresholds) and Table 2 (dynamic thresholds). In the first column the height of the fixed threshold and the gap parameter g that were varied are presented. A comparison of the second columns of both Table 1 and Table 2 shows that a similar overall response rate is predicted by the model by the both thresholds.

In the remaining columns the response rates for each interactions are presented. The cells are gray shaded for easier interpretation, with higher response

Table 2. The table illustrates the effect of different gap parameters in the dynamic threshold on the response rate in responses per minute of a listener response prediction model. The cells are gray shaded for easier interpretation, with higher response rate being darker. It illustrates that, for each interaction a similar response rate is generated.

	Response Rate for Dynamic Threshold (responses/minute)										
		Interaction									
Gap g	**Overall**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2.8	18.1	17.7	19.1	18.6	18.0	17.8	18.4	19.2	17.0	18.5	18.1
3.6	14.8	14.1	14.8	15.4	14.7	14.4	14.6	15.7	14.5	15.1	15.2
4.5	12.1	11.9	12.2	12.4	12.1	11.9	11.9	13.1	12.0	12.1	12.2
5.2	10.6	10.4	10.4	10.8	10.5	10.2	11.0	11.4	10.5	10.8	10.6
5.9	9.5	9.8	9.4	10.1	9.3	8.9	9.3	10.1	9.2	9.8	9.9
6.9	8.4	8.8	8.3	8.5	8.0	7.9	8.3	9.2	8.2	8.7	8.6
8.4	7.0	6.8	7.6	7.5	6.9	6.7	6.7	7.9	6.7	7.4	7.2
10.9	5.7	5.6	5.8	6.2	5.9	5.2	5.5	6.6	5.3	5.7	5.7
13.7	4.7	4.3	4.3	5.2	4.9	4.7	4.5	5.2	4.5	4.9	4.7
19.2	3.6	3.3	4.0	3.6	3.6	3.5	3.4	3.9	3.3	3.8	3.6
25.2	2.9	2.6	3.2	3.1	3.1	3.0	2.8	3.5	2.7	3.2	2.9

rate being darker. These rates show that for the fixed threshold the response rate of each interaction can vary wildly. Especially the predicted response rate for interaction 5 is a lot lower than for the other interactions, while the predicted response rate for interaction 7 is generally higher. For the dynamic threshold the response rates for each individual interaction are a lot more stable. Each interaction has more or less the same response rate. So indeed, the dynamic threshold has succeeded in stabilizing the response rate.

The next question we will answer is, does the dynamic threshold avoid periods with many or few responses? First we look at whether the dynamic threshold avoids periods with many responses. For this, we will see how many responses were discarded by the extra rule that no predicted responses should be within 1 second of each other. The number of discarded predictions is presented in Table 3 and the left graph in Figure 2. They show that this number is about twice as high in case of the fixed threshold compared to the dynamic threshold. So, the dynamic threshold helps to avoid periods with many responses.

To see whether the dynamic threshold helps to avoid periods with few predicted responses, an analysis of the gaps between two consecutive predicted responses is performed. We looked at the mean, standard deviation and maximum of these gaps. The results of this analysis are presented in Table 3. By looking and comparing the columns μ and max of both thresholds, we can see that the difference between the mean and maximum gap is a lot more stable for the dynamic threshold (see also right graph in Figure 2). This is also reflected in the lower standard deviation for the dynamic threshold (see middle graph in Figure 2). So, we can conclude that the dynamic threshold also resulted in a more even distribution of the predicted responses.

Table 3. Presentation of the results of the analysis of the gaps between predicted responses using a fixed threshold and our dynamic threshold

Fixed Threshold					Dynamic Threshold				
	Discarded	Gap (s)				Discarded	Gap (s)		
Threshold	Predictions	μ	σ	max	Gap g	Predictions	μ	σ	max
0.15	384	2.75	1.85	24.17	**2.8**	129	3.31	1.41	7.33
0.17	202	3.56	2.93	47.10	**3.6**	75	4.06	1.89	9.37
0.19	114	4.51	5.12	85.63	**4.5**	52	4.97	2.53	10.93
0.21	79	5.21	6.06	85.63	**5.2**	43	5.65	2.98	13.67
0.23	62	5.91	6.68	85.63	**5.9**	40	6.35	3.42	15.17
0.25	55	6.94	8.25	97.60	**6.9**	27	7.17	4.09	17.13
0.27	41	8.51	9.60	97.60	**8.4**	18	8.61	4.95	20.97
0.29	23	11.01	12.48	97.60	**10.9**	16	10.74	6.79	27.03
0.31	16	13.73	15.56	97.60	**13.7**	11	13.13	8.30	33.47
0.33	11	19.15	19.77	101.93	**19.2**	6	17.53	12.13	43.63
0.35	9	22.80	22.61	101.93	**25.2**	5	21.49	16.11	55.87

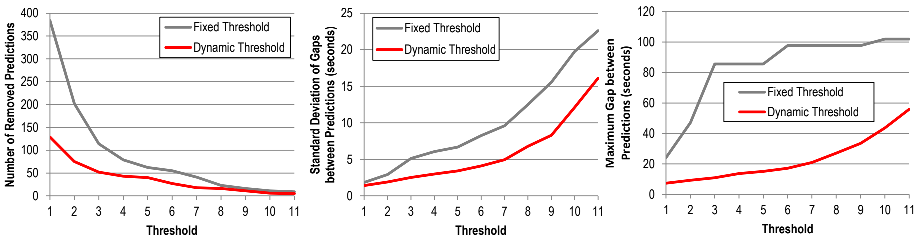


Fig. 2. Three graphs to illustrate that for all thresholds the dynamic thresholding needs less predictions to be removed (left graph), has a lower standard deviation for the gaps between two consecutive predictions (middle graph) and has a smaller maximum gap between consecutive predictions (right graph)

4 Conclusion

In this paper we have presented a novel way of interpreting the prediction value curves that are the output of the current state-of-the-art models in predicting generic listener responses for embodied conversational agents. Based on the time since the last generated listener response the dynamic thresholding approach varies the threshold that peaks in the prediction value curve need to exceed in order to be selected as a suitable place for a listener response. We have shown through an objective evaluation that this approach generates behavior that is more stable under changing conditions.

Furthermore, the proposed formula for this dynamic threshold includes a parameter which controls the response rate of the generated behavior. This gives the designer of the listening behavior of a virtual listener the tools to create the behavior that is desired for the targeted role, personality or situation. When the role, personality or situation requires a low listener response rate, the designer now has a reliable way of ensuring this response rate for any interlocutor.

A subjective evaluation of the generated behavior has been performed, but full coverage of this evaluation was considered outside the scope of this paper. In this subjective evaluation the behavior generated with the dynamic thresholding was preferred over the behavior generated with the fixed threshold.

The current state-of-the-art models for predicting generic listener responses are general models aimed to work for every situation, for every interaction partner and for every other context one can think of. The dynamic threshold proposed in this paper is a way of achieving this with the current state-of-the-art prediciton models, but ultimately we need more advanced models that can adapt to different speaking styles, conversational settings and/or changing conditions by themselves.

References

1. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology 79(6), 941–952 (2000)
2. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. European ACL pp. 51–58 (2003)
3. Goodwin, C.: Between and within: Alternative sequential treatments of continuers and assessments. Human Studies 9(2-3), 205–217 (1986)
4. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)
5. Huang, L., Morency, L.P., Gratch, J.: Learning Backchannel Prediction Model from Parasocial Consensus Sampling: A Subjective Evaluation. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 159–172 (2010)
6. Huang, L., Morency, L.P., Gratch, J.: Parasocial Consensus Sampling: Combining Multiple Perspectives to Learn Virtual Human Behavior. In: Proceedings of Autonomous Agents and Multi-Agent Systems, Toronto, Canada, pp. 1265–1272 (2010)
7. de Kok, I., Heylen, D.: The MultiLis Corpus – Dealing with Individual Differences in Nonverbal Listening Behavior. In: Esposito, A., Esposito, A.M., Martone, R., Müller, V.C., Scarpetta, G. (eds.) COST 2102 Int. Training School 2010. LNCS, vol. 6456, pp. 362–375. Springer, Heidelberg (2011)
8. de Kok, I., Heylen, D.: A survey on evaluation metrics for backchannel prediction models. In: Interdisciplinary Workshop on Feedback Behaviors in Dialog, pp. 15–18 (2012)
9. de Kok, I., Ozkan, D., Heylen, D., Morency, L.-P.: Learning and Evaluating Response Prediction Models using Parallel Listener Consensus. In: Proceeding of International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (2010)
10. Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling Embodied Feedback with Virtual Humans. In: Wachsmuth, I., Knoblich, G. (eds.) Modeling Communication. LNCS (LNAI), vol. 4930, pp. 18–37. Springer, Heidelberg (2008)

11. Kraut, R.E., Lewis, S.H., Swezey, L.W.: Listener responsiveness and the coordination of conversation. Journal of Personality and Social Psychology 43(4), 718–731 (1982)
12. Maatman, R.M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 25–36. Springer, Heidelberg (2005)
13. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems 20(1), 70–84 (2011)
14. Nishimura, R., Kitaoka, N., Nakagawa, S.: A spoken dialog system for chat-like conversations considering response timing. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 599–606. Springer, Heidelberg (2007)
15. Noguchi, H., Den, Y.: Prosody-based detection of the context of backchannel responses. In: Fifth International Conference on Spoken Language Processing (1998)
16. Ozkan, D., Morency, L.P.: Latent Mixture of Discriminative Experts. IEEE Transaction on Multimedia 15(2), 326–338 (2013)
17. Poppe, R., Truong, K.P., Heylen, D.: Perceptual evaluation of backchannel strategies for artificial listeners. Autonomous Agents and Multi-Agent Systems (January 2013)
18. Sakai, Y., Nonaka, Y., Yasuda, K., Nakano, Y.I.: Listener agent for elderly people with dementia. In: Proceedings of HRI 2012, pp. 199–200 (2012)
19. Schröder, M., Bevacqua, E., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., Pammi, S., Pantic, M., Schuller, B., Pelachaud, C., de Sevin, E., Wollmer, M., Valstar, M.: A demonstration of audiovisual sensitive artificial listeners. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–2. IEEE, Amsterdam (September 2009)
20. de Sevin, E., Hyniewska, S.J., Pelachaud, C.: Influence of personality traits on backchannel selection. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 187–193. Springer, Heidelberg (2010)
21. Takeuchi, M., Kitaoka, N., Nakagawa, S.: Timing detection for realtime dialog systems using prosodic and linguistic information. In: International Conference on Speech Prosody, pp. 529–532 (2004)
22. Traum, D., DeVault, D., Lee, J., Wang, Z., Marsella, S.: Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 275–288. Springer, Heidelberg (2012)
23. Wang, Z., Lee, J., Marsella, S.: Towards More Comprehensive Listening Behavior: Beyond the Bobble Head. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 216–227. Springer, Heidelberg (2011)
24. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207 (2000)

The Influence of Prosody on the Requirements for Gesture-Text Alignment

Yingying Wang and Michael Neff

University of California, Davis
{yiwang,mpneff}@ucdavis.edu

Abstract. Designing an agent capable of multimodal communication requires synchronization of the agent's performance across its communication channels: text, prosody, gesture, body movement and facial expressions. The synchronization of gesture and spoken text has significant repercussions for agent design. To explore this issue, we examined people's sensitivity to misalignments between gesture and spoken text, varying both the gesture type and the prosodic emphasis. This study included ratings of individual clips and ratings of paired clips with different alignments. Subjects were unable to notice alignment errors of up to ± 0.6s when shown a single clip. However, when shown paired clips, gestures occurring after the lexical affiliate are rated less positively. There is also evidence that stronger prosody cues make people more sensitive to misalignment. This suggests that agent designers may be able to "cheat" when it comes to maintaining tight synchronization between audio and gesture without a decrease in agent naturalness, but this cheating may not be optimal.

1 Introduction

Agents capable of multimodal communication simultaneously express content across multiple channels. These channels include body movement, facial expressions, gesture, spoken text and prosody changes. The need to coordinate these channels puts high demands on the planning and animation subsystems of an agent, yet it is not clear how tight the synchronization must be in order to generate a believable agent. This paper looks at the need for alignment between gesture and spoken text, and how this requirement may vary across the two main factors likely to influence it: the type of gesture performed and variation in the prosody of the accompanying text.

To better understand the need for gesture-text alignment, and how prosody may influence it, we ran two sets of experiments. The first asked people to rate single clips for naturalness that always had strong prosodic emphasis on the lexical affiliate, but varied gesture type and alignment. The second asked people to select a preferable clip from two side-by-side clips and examined variation in the gesture type, prosody and alignment. Results indicate that people have quite low sensitivity to alignment when viewing a single clip. However, when given a second, side-by-side clip, they in general show a lower preference for gestures occurring after the lexical affiliate.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 180–188, 2013.

2 Background

According to McNeil in [15], gestures are classified into 4 categories, "beat" - a rhythmic flick of finger, hand or arm to highlight what is being said, "deictic" - a pointing gesture with direction, "iconic" - a representation of a concrete object, or painting with hand and "metaphoric" - explanation of an abstract concept. Besides this, [10,20,7] also provide their own taxonomies of gestures.

In virtual character gesture research, solutions for coordinating gestures with other modalities have received a lot of focus. Kopp in [12] presents an incremental production model that can help generate multimodal behaviors from utterance planning. Stone et al.[23] worked on a framework for creating talking characters with both sound and motion data from the real human performance. Cassell et al.[5] takes in text input and generate nonverbal behaviors including both facial expressions and gesture. The system presented in [19] uses plain text as input and their goal is to find a solution for automatically adding gestures.

From previous research, among the multiple communication modalities, prosody is shown to have a close correlation with gesture. According to [24], prosody correlates to emphasis, which makes it coincide with emphasizing gestures like "beats". Adolphs in [1] and Schroder in [22] mentioned that prosody contains emotive information, which according to [25] and [17] is frequently reflected in non-verbal behavior. Prosody features were previously used to generate facial expression [2,9], head motion [6], head orientation [21], and gesture [18,14].

According to McNeil [15], gesture strokes end at or before, but never after the prosodic stress peak of the accompanying syllable. Results from [26] showed that 90% gestures cooccur with lexeme syllable in the speech, and 65% to 75% cases contain a prosody accent. Experiments in [11] indicate that gestures which are performed 0.2 second or 0.6 second earlier w.r.t. the accompanying text get higher ratings for their naturalness.

3 Experiment 1: Perception of Gesture Misalignment with Single Clip

Our main hypotheses are that the perceived gesture misalignment varies based on gesture type and also prosodic emphasis. To verify the effect of gesture type and prosodic emphasis, our experiment design include four different types of gestures: "deictic" (D), "metaphoric" (M), "iconic" (I) and "beat" (B) – being placed on utterances with weak (W) prosody or strong (S) prosody.

3.1 Gesture Form

We designed the following four utterances to include the "deictic", "metaphoric", "iconic" and "beat" gesture types. The designed gestures are associated with the highlighted lexeme. A "pointing" gesture was used as the "deictic" gesture, indicating "you" in the text; a "progressive" gesture symbolizes the progress of the conversation; the "iconic" gesture shapes the size of the box; and a dismissal flip of hand motion was used as one-peak “beat” gesture to express the negative content. The gesture motions were generated by editing motion captured data.

T1(D): I know that **you** took it.
T2(M): The conversation **ran** for a long time.
T3(I): I bought a **box** at the store.
T4(B): This is **not** the case.

3.2 Prosody Emphasis

We generated two utterances for each text, with weak prosody and strong prosody. Based on previous research [6,4,9,26,14,13], we describe prosody using pitch and intensity. The utterances were recorded from a male adult with different variations of flat and emphasized prosody. We used the Praat speech analysis tool [3] to analyze the recorded utterance, and guarantee no significant intensity or pitch variations in the weak prosody utterances, while for strong prosody cases, there was at least a 10dB change of intensity and a 100Hz change of pitch for the highlighted lexeme. To differentiate the scenario, we use x-y notation, where x indicates the gesture, and y indicates the prosody. Thus we have 8 utterances, D-W, D-S, M-W, M-S, I-W, I-S, B-W and B-S.

3.3 Alignment Timing

To include the misalignment of gesture w.r.t. the utterance, gestures were placed on the lexeme, with 7 different offsets: -0.6s, -0.4s, -0.2s, 0s, +0.2s, +0.4s, +0.6s. Thus 7 motion clips were used in total for each utterance to verify the perceived naturalness. Each clip lasts about 10 seconds and contains one gesture in the utterance.

3.4 Experiment Execution

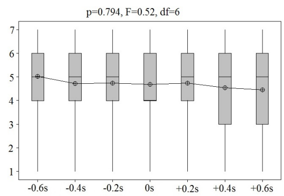
For our experiment, a male virtual character rendered in Maya[8] is used to match the voice for the gesture performance. The character’s face is blocked, and thus neither asynchronous lip movement nor facial expression will affect subjects’ judgment. A front viewpoint was chosen which displayed the character from the knee to head. The clips are generated at size 640 x 480 with high quality, see Figure (1).

We conduct our experiments by putting all the movie clips on mechanical turk. Subjects watched each clip and were then asked to rate the naturalness of the character’s behavior. A 7-point Likert rating scale was used, with 1 indicating “least natural”, and 7 indicating “most natural”. We encouraged subjects to try to use the entire rating scale. Subjects were allowed to replay the clips as many times as they wanted.

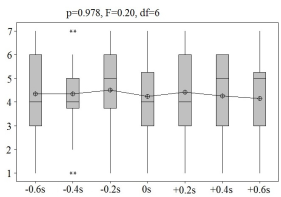
Fig. 1. "Deictic" gesture in the utterance T1

3.5 Result

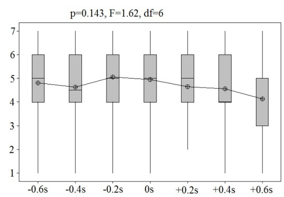
40 subjects participated in our perception study. For each utterance, we ran the ANOVA to check the perceived difference of the gesture timing factor. No significant rating differences were found over the alignment timing factor for all the 8 utterances. Results for strong prosody cases are illustrated in Figure 2. Thus we conclude, given a single clip, subjects do not have adequate sensitivity to differentiate misaligned gestures.



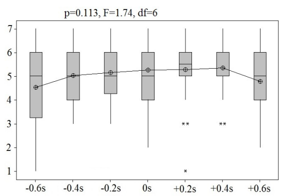
(a) D-S Ratings



(b) M-S Ratings



(c) I-S Ratings



(d) B-S Ratings

Fig. 2. Naturalness ratings for single clips with strong prosody

4 Experiment 2: Perception of Gesture Misalignment with Paired Clips

4.1 Experiment Design

We maintained the gesture type, prosodic emphasis and alignment timing, and re-organize the single clips from Experiment 1 into comparison pairs. According to previous research [11,15,16], we assume -0.2s is the most natural gesture timing (gesture occurring slightly before the word), and pair clips of "-0.2s" with clips of "-0.6s", "-0.4s", "0s", "+0.2s", "+0.4s" and "+0.6s", given the same utterance. Paired clips were shown in different left and right orders to eliminate an order effect. For each utterance, 12 pairs of movie clips were generated, see Figure (3).

We conducted our experiments by putting the paired movie clips of all utterances on mechanical turk. To evaluate the perceived naturalness, subjects were asked to choose the more natural one by selecting a relative naturalness rating between the two clips. Five options were offered, "Strongly prefer the left clip", "Slightly prefer the left clip", "Almost the same", "Slightly prefer the right clip" and "Strongly prefer the right clip", with numeric scores -2, -1, 0, 1 and 2. A naturalness rating was calculated from subject's choice, 1 point will be added if the clip is slightly preferred, 2 point will be added if strongly preferred, and 0 point if not preferred or subjects cannot tell the difference.

4.2 Result

40 subjects were recruited for the perception task. For paired clips in the movie, we run the 2-sample T test on their ratings to check subjects' preference. Due to the use of multiple T-tests, Bonferroni correction was used, which sets the significance at level 0.001. In general, people were more likely to detect late gestures as being unnatural, and tend to give higher natural ratings to early gestures. However, the detailed situation differs based on gesture types and prosodic emphasis, which we will discuss separately. The ratings from different left/right orders for the same clip are combined, as our ANOVA does not detect significant effect on left/right ordering.

Fig. 3. Side-by-side comparison of "deictic" misalignment

D-W and D-S: For "deictic" gesture, results for weak prosody and strong prosody are listed in Table 1 and Table 2. With weak prosody, subjects can only detect the difference when "deictic" gesture is more than +0.2s later than the lexeme, otherwise, they don't differentiate the gesture alignment with the -0.2s alignment. Given strong prosody, the rating difference is significant between -0.2s alignment and 0s alignment, which indicates subjects strongly prefer the "deictic" gesture being placed 0.2s earlier than the utterance rather than exactly on the utterance. In both prosody settings, subjects prefer earlier "deictic" gesture, but do not really care too much how much earlier it is within the time window -0.6s to -0.2s.

M-W and M-S: For "metaphoric" gesture, results for weak prosody and strong prosody are listed in Table 3 and Table 4. Subjects can detect the all the late alignment of "metaphoric" gestures given the weak prosody. However, when using the strong prosody utterance, subjects no longer detect the difference between the -0.2s alignment and the 0s alignment.

I-W and I-S: For "iconic" gesture, results for weak prosody and strong prosody are listed in Table 5 and Table 6. Given weak prosody, the rating difference between -0.2s alignment and early -0.4s alignment and all late placements are significant. However, when using strong prosody utterance, subjects do not differentiate -0.2s alignment with 0s alignment, nor do they notice the difference between -0.2s, -0.4s and -0.6s alignment. In weak prosody settings, subjects indicate preference for earlier "iconic" gesture placement, and the -0.4s alignment is more preferable than -0.2s alignment.

B-W and B-S: For "beat" gesture, results for weak prosody and strong prosody are listed in Table 7 and Table 8. Subjects' sensitivity to gesture alignment does not vary too much given 2 different prosody settings. Late "beat" gesture can be detected, while earlier placement does not seem to differ too much according to subjects' ratings.

Between Gesture Types: Our findings in the experiments verify that gestures performed earlier than its spoken text are perceived as more natural. Between different gesture types, the situations are not exactly the same. The influence of prosody on the perceived naturalness varies for different gesture types. Strong prosody can sharpen subjects' sensitivity to gesture misalignment. For the "deictic" and "beat" gestures, subjects were more able to detect misaligned clips when the prosody was strong. For the "metaphoric" and "iconic" gestures, in the weak prosody case, there was an example of an earlier -0.4s alignment being preferred over the -0.2s alignment. This did not happen in the strong prosody case. So the prosody signal may have weakened the preference for an early gesture. It can be explained as "metaphoric" and "iconic" gestures are more likely to contain information not tightly coupled with the prosodic emphasis, while "deictic" and "beat" have sharper forms which are more related to the voice.

Table 1. Average ratings and T test results from paired clips for utterance D-W

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.431) -0.6s (0.444) T=-0.14, P=0.892	-0.2s (0.284) -0.4s (0.432) T=-1.59, P=0.114	-0.2s (0.465) 0s (0.31) T=1.56, P=0.12	-0.2s (0.696) +0.2s (0.261) T=3.91, **P<0.0001**	-0.2s (0.922) +0.4s (0.195) T=6.87, **P<0.0001**	-0.2s (1.139) +0.6s (0.153) T=9.11, **P<0.0001**

Table 2. Average ratings and T test results from paired clips for utterance D-S

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.608) -0.6s (0.405) T=1.81, P=0.072	-0.2s (0.239) -0.4s (0.451) T=-2.4, P=0.018	-0.2s (0.639) 0s (0.449) T=4.61 **P<0.0001**	-0.2s (0.973) +0.2s (0.151) T=8.01, **P<0.0001**	-0.2s (1.141) +0.4s (0.141) T=9.08, **P<0.0001**	-0.2s (1.040) +0.6s (0.133) T=8.34, **P<0.0001**

Table 3. Average ratings and T test results from paired clips for utterance M-W

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.284) -0.6s (0.608) T=-3.22, P=0.002	-0.2s (0.227) -0.4s (0.507) T=-3.01, P=0.003	-0.2s (0.581) 0s (0.230) T=3.69, **P<0.0001**	-0.2s (0.693) +0.2s (0.187) T=5.27, **P<0.0001**	-0.2s (0.760) +0.4s (0.263) T=4.49, **P<0.0001**	-0.2s (0.895) +0.6s (0.250) T=5.53, **P<0.0001**

Table 4. Average ratings and T test results from paired clips for utterance M-S

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.5) -0.6s (0.474) T=0.23, P=0.818	-0.2s (0.321) -0.4s (0.385) T=-0.68, P=0.498	-0.2s (0.52) 0s (0.267) T=2.58, P=0.011	-0.2s (0.72) +0.2s (0.267) T=4.23 **P<0.0001**	-0.2s (0.933) +0.4s (0.160) T=7.05, **P<0.0001**	-0.2s (0.960) +0.6s (0.213) T=6.77 **P<0.0001**

Table 5. Average ratings and T test results from paired clips for utterance I-W

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.347) -0.6s (0.533) T=-2.06, P=0.041	-0.2s (0.186) -0.4s (0.5) T=-3.45, **P=0.0001**	-0.2s (0.568) 0s (0.149) T=5.17, **P<0.0001**	-0.2s (0.84) +0.2s (0.16) T=7.3, **P<0.0001**	-0.2s (1.054) +0.4s (0.108) T=10.37, **P<0.0001**	-0.2s (1.068) +0.6s (0.137) T=8.77, **P<0.0001**

Table 6. Average ratings and T test results from paired clips for utterance I-S

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.349) -0.6s (0.538) T=-2.22, P=0.028	-0.2s (0.418) -0.4s (0.439) T=-0.2, P=0.839	-0.2s (0.422) 0s (0.328) T=0.94 P=0.347	-0.2s (0.721) +0.2s (0.246) T=4.0 **P<0.0001**	-0.2s (0.723) +0.4s (0.292) T=3.63, **P<0.0001**	-0.2s (0.885) +0.6s (0.262) T=4.87, **P<0.001**

Table 7. Average ratings and T test results from paired clips for utterance B-W

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.346) -0.6s (0.538) T=-1.94, P=0.054	-0.2s (0.25) -0.4s (0.438) T=-2.05, P=0.042	-0.2s (0.539) 0s (0.276) T=2.75, P=0.007	-0.2s (0.797) +0.2s (0.114) T=8.21, **P<0.0001**	-0.2s (1) +0.4s (0.132) T=8.28, **P<0.0001**	-0.2s (1.208) +0.6s (0.091) T=11.31, **P<0.0001**

Table 8. Average ratings and T test results from paired clips for utterance B-S

Alignment	-0.6s	-0.4s	0s	+0.2s	+0.4s	+0.6s
-0.2s	-0.2s (0.449) -0.6s (0.487) T=-0.37, P=0.712	-0.2s (0.244) -0.4s (0.423) T=-2.22, P=0.028	-0.2s (0.519) 0s (0.247) T=2.91 P=0.004	-0.2s (0.857) +0.2s (0.065) T=9.57 **P<0.0001**	-0.2s (1.039) +0.4s (0.143) T=8.68, **P<0.0001**	-0.2s (1.138) +0.6s (0.125) T=9.47, **P<0.0001**

5 Discussion and Conclusion

This paper summarizes a series of studies exploring people's sensitivity to gesture alignment with text and how this varies as prosody changes. When shown a single clip with a strong prosody signal, our subjects on Mechanical Turk were not able to reliably detect misalignment, even given relatively large misalignments of 0.6s. However, when shown side-by-side clips, subjects generally viewed gestures occurring later, greater than 0.2s after the lexical affiliate, less favorably. Results were more mixed for gestures occurring early, and in some cases, gestures occurring 0.4s before the lexical affiliate were preferred to those occurring 0.2s before the lexical affiliate, our presumed best alignment.

With regards to prosody, the picture is complex. We had hoped to find a clear indication that when prosody was strong, people had higher demands for alignment. This does appear to be true for "deictic" and "beat" gestures, but the opposite picture emerged for the "metaphoric" and "iconic" examples. This is perhaps not surprising as "deictic" and "beat" gestures may both have sharper forms, more tightly coupled with the emphasis in the voice, whereas "metaphoric" and "iconic" gestures are more likely to contain information not copied in the speech. This may be an area worth further study.

In terms of agent design, it would appear that it is not necessary to maintain tight alignment, as people seem to have low sensitivity to this, especially if seeing only one clip. Where variation from the speech is allowed, it seems clear that it is preferable to move the gestures earlier in time, not later.

Acknowledgments. Financial support for this work was provided in part through NSF grants IIS 1115872 and IIS 0845529.

References

1. Adolphs, R.: Neural systems for recognizing emotion. Current Opinion in Neurobiology 12(2), 169–177 (2002)
2. Albrecht, I., Haber, J., Seidel, H.P.: Automatic generation of non-verbal facial expressions from speech. In: Advances in Modelling, Animation and Rendering, pp. 283–293. Springer, Heidelberg (2002)
3. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2002)
4. Bregler, C., Covell, M., Slaney, M.: Video rewrite: Driving visual speech with audio. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, pp. 353–360. ACM Press/Addison-Wesley Publishing Co (1997)

5. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 477–486. ACM (2001)
6. Chuang, E., Bregler, C.: Mood swings: expressive speech animation. ACM Transactions on Graphics (TOG) 24(2), 331–347 (2005)
7. Efron, D.: Gesture and Environments. King's Crown Press (1941)
8. Inc., A.: Maya, 3d computer graphics software (2008)
9. Ju, E., Lee, J.: Expressive facial gestures from motion capture data, vol. 27(2), pp. 381–388 (2008)
10. Kendon, A.: Current issues in the study of gesture. In: The Biological Foundations of Gestures: Motor and Semiotic Aspects, pp. 23–47 (1986)
11. Kirchhof, C.: On the audiovisual integration of speech and gesture. In: The 5th Conference of the International Society for Gesture Studies, ISGS (2012)
12. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1), 39–52 (2004)
13. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. ACM Transactions on Graphics (TOG) 29(4), 124 (2010)
14. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. ACM Transactions on Graphics (TOG) 28, 172 (2009)
15. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
16. McNeill, D.: Gesture and thought. University of Chicago Press (2008)
17. Montepare, J., Koff, E., Zaitchik, D., Albert, M.: The use of body movements and gestures as cues to emotions in younger and older adults. Journal of Nonverbal Behavior 23(2), 133–152 (1999)
18. Morency, L.P., Sidner, C., Lee, C., Darrell, T.: Head gestures for perceptual interfaces: The role of context in improving recognition. Artificial Intelligence 171(8), 568–585 (2007)
19. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) 27(1), 5 (2008)
20. Rimé, B., Schiaratura, L.: Gesture and speech (1991)
21. Sargin, M.E., Erzin, E., Yemez, Y., Tekalp, A., Erdem, A., Erdem, C., Ozkan, M.: Prosody-driven head-gesture animation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 2, pp. II–677. IEEE (2007)
22. Schröder, M.: Speech and emotion research
23. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with hands: Creating animated conversational characters from recordings of human performance. ACM Transactions on Graphics (TOG) 23(3), 506–513 (2004)
24. Terken, J.: Fundamental frequency and perceived prominence of accented syllables. The Journal of the Acoustical Society of America 89, 1768 (1991)
25. Wallbott, H.G.: Bodily expression of emotion. European Journal of Social Psychology 28(6), 879–896 (1998)
26. Yasinnik, Y., Renwick, M., Shattuck-Hufnagel, S.: The timing of speech-accompanying gestures with respect to prosody. In: Proceedings of the International Conference: From Sound to Sense, vol. 50, pp. 10–13 (2004)

Prediction of Visual Backchannels in the Absence of Visual Context Using Mutual Influence

Derya Ozkan and Louis-Philippe Morency

University of Southern Caligornia
Institute for Creative Technologies
10215 Waterfront Drive, Playa Vista, CA
{ozkan,morency}@ict.usc.edu

Abstract. Based on the phenomena of mutual influence between participants of a face-to-face conversation, we propose a context-based prediction approach for modeling visual backchannels. Our goal is to create intelligent virtual listeners with the ability of providing backchannel feedbacks, enabling natural and fluid interactions. In our proposed approach, we first anticipate the speaker behaviors, and then use this anticipated visual context to obtain more accurate listener backchannel moments. We model the mutual influence between speaker and listener gestures using a latent variable sequential model. We compared our approach with state-of-the-art prediction models on a publicly available dataset and showed importance of modeling the mutual influence between the speaker and the listener.

Keywords: nonverbal behavior, embodied conversational agent.

1 Introduction

During face-to-face communication, participants often mutually influence each other through their verbal and nonverbal behaviors. For instance, a speaker will decide to give more explanations or simply continue with the story based on the feedbacks from the listener. Similarly, participants often mimic each others gestures to convey empathy and rapport [1–3]. This phenomena, which we refer as mutual influence in this paper, is essential for fluid human interactions; but research is still needed to replicate this process with virtual humans.

A good example of human behaviors that involves mutual influence is backchannel feedbacks (i.e. the nods and paraverbal signals such as "uh-hu" and "mm-hmm" that listeners produce as someone is speaking). Backchannel feedbacks have received considerable attention due to their pervasiveness across languages and conversational contexts. They play a significant role in determining the nature of a social exchange by showing rapport and engagement [4]. When these signals are positive, coordinated and reciprocated, they can lead to feelings of rapport and promote beneficial outcomes in diverse areas such as negotiations and conflict resolution [5], psychotherapeutic effectiveness [6], improved test performance

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 189–202, 2013.

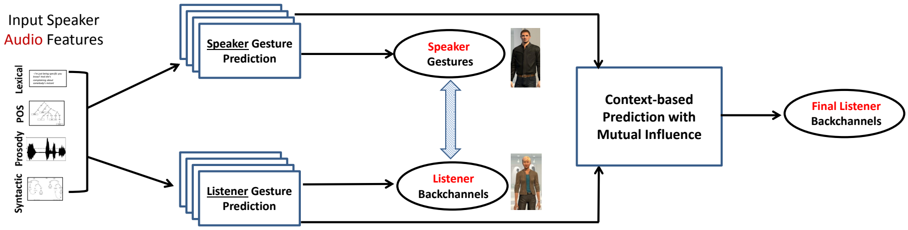


Fig. 1. An overview of our approach for predicting listener backchannels in the absence of visual information. Our approach takes into account the context from the speaker by first predicting the nonverbal behaviors of the speaker and uses these predictions to improve the final listener backchannels.

in classrooms [7] and improved quality of child care [8]. By correctly predicting backchannel feedback, we can improve the way that the machines communicate with human. For instance, a virtual human that provides head nods at reasonable points in the conversation can have a stronger sense of rapport.

One of the challenges in building intelligent virtual agents with such abilities is absence of the visual information. In many real-world applications, we often have only the speech and/or text to be spoken by the virtual human, without any visual context. Another scenario where no visual context is available is phone-to-phone conversations. If we want to create a virtual (i.e. customer service) representative that is capable of providing backchannel feedbacks, the only source of information is the interlocutor's (customer's) voice. As discussed above, a good prediction model of backchannels should be able to take into account the mutual influence between participants even in the absence of visual context. This can be achieved by anticipating the nonverbal behaviors of the speaker, and using the anticipated visual context to model the mutual influence between the speaker and the listener.

In this paper, we present a context-based prediction model to predict backchannels of a listener during dyadic conversations. An overview of our approach is given in Figure 1. We assume an environment where the visual gestures of the speaker are not available. Based on this assumption, we first predict the visual context (i.e. nonverbal behaviors) of the speaker and the backchannels of the listener using only the auditory observations (features) from the speaker. We model the mutual influence between the speaker and the listener by using a latent variable model based on Latent Mixture of Discriminative Experts (LMDE) [9]. We evaluate our approach using 45 storytelling dyadic interactions from the RAPPORT dataset [10]. In our experiments, we compare our approach with previous approaches based on Conditional Random Fields (CRF) [11], Latent-Dynamic CRFs [12], and CRF Mixture of Experts (a.k.a Logarithmic Opinion Pools [13]), and a rule based random predictor [14].

The paper is organized as follows. We first present the related works in Section 2. Then we present our context-based prediction approach in Section 3.

Experimental setup, and results are given in Section 4 and Section 5, respectively. Finally, we conclude in Section 6.

2 Related Works

Although human can naturally display and interpret nonverbal signals in social context, computers are not equipped with such abilities. Therefore, supporting such fluid interactions has become an important topic in computer science research [15]. Many different models have been proposed to recognize [16, 17], or predict [14, 18, 10] certain nonverbal behaviors.

The application described in this paper uses audio cues from the speaker to predict the social behavior of the participant. This type of predictive models has been mostly studied in the context of embodied conversational agents [19, 20]. Several researchers have developed models to predict when backchannel should happen. In general, these results are difficult to compare as they utilize different corpora and present varying evaluation metrics. Ward and Tsukahara [14] propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Models were produced manually through an analysis of English and Japanese conversational data. Later in 2003, Ward [21] studied both the forms and functions of sounds like h-nmm, hh-aaaah, hn-hn, unkay, nyeah, ummum, uuh and um-hmuh -hm in American English conversation.

Fujie et al. [22] use Hidden Markov Models to perform head nod recognition. In their proposal, they combined head gesture detection with prosodic low-level features from the same person to determine strongly positive, weak positive and negative responses to yes/no type utterances. Maatman et al. [18] present a multimodal approach where Ward and Tsukhara's prosodic algorithm is combined with a simple method of mimicking head nods. No formal evaluation of the predictive accuracy of the approach was provided but subsequent evaluations have demonstrated that generated behaviors do improve subjective feelings of rapport [23] and speech fluency [4]. Morency et al. [10] showed that Conditional Random Field models can be used to learn predictive features of backchannel feedback. In their approach, multimodal features are simply concatenated in one large feature vector for the CRF model. They show statistical improvement when compared to the rule-based approach of Ward and Tsukahara [14].

The Semaine Project of EU-FP7 [24] focuses on building *Sensitive Artificial Listeners*. Towards this effort, Gravano [25] focuses on backchannel-inviting cues as part of as part of their study of turn-taking phenomena. They first analyze individual acoustic, prosodic and textual backchannel-inviting cues; then, they investigate how such cues combine together to form complex signals. In [26], Neiber focuses on the communicative functions of vocal feedback like "mhm", "okay" and "yeah, thats right". They categorize feedback as non-lexical, lexical and phrase based feedback.

In this paper, we present an approach to predict the backchannels of a listener using the anticipated visual context of the speaker. More specifically, we focus on the visual feedbacks of the listener: head nods. We assume an environment, in which the visual information for both the listener and the speaker is absent.

3 Context-based Backchannel Prediction with Mutual Influence

The goal of our approach is to predict listener backchannels in dyadic conversations by using the mutual influence between the speaker and the listener. We assume a situation where no visual context from neither the speaker nor the listener is available. In other words, we have no access to speaker's visual information, but only the speech/text information from the speaker. In our approach, we explicitly model multiple dimensions of the speech information such as prosody, lexicons, syntactic structure and part-of-speech tags. These different dimensions contain complementary information, and our approach will model the hidden dynamic between them.

In our context-based prediction approach, we first infer the speaker gestures, and then exploit this visual context to improve the final listener backchannel predictions (see Figure 1). In order to model the mutual influence between the speaker and the listener, we use a variant of the Latent Mixture of Discriminative Experts

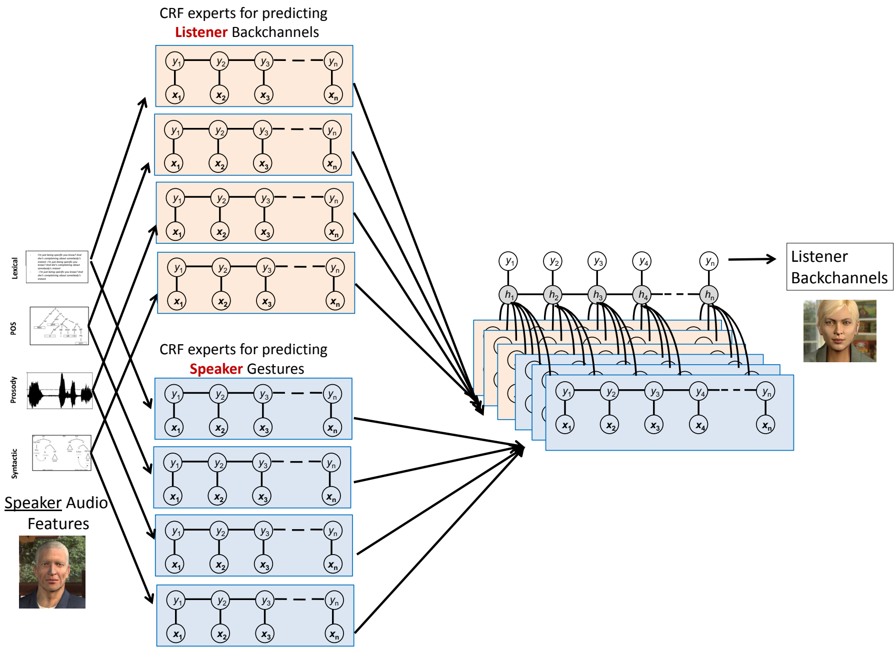


Fig. 2. Our approach for predicting speaker gestures in dyadic conversations. Using the speaker audio features as input, we first learn a CRF model (expert) per each audio channel and for both speaker gestures and listener backchannels. Then, we merge these CRF experts using a latent variable model that is capable of learning the hidden dynamic among the experts. This second step allows us to incorporate the mutual influence between the speaker and the listener.

(LMDE) [9] called mutual-LMDE. LMDE was originally proposed to integrate data from multiple modalities. One of the main advantages of this model is that it can automatically discover the hidden structure among modalities and learn the dynamic between them. We extend the LMDE model to also take into consideration the mutual influence between speaker and listener.

Our mutual-LMDE model is based on a two step process (an overview is shown in Figure 2): in the first step, we learn discriminative experts for speaker gestures and listener backchannels. Speaker expert models are trained using a Conditional Random Field (CRF) [11] on one of the four speech dimensions (prosody, lexicons, syntactic structure and part-of-speech tags). These individual experts make up for the visual context from the speaker. We learn experts for listener backchannels similar to speaker gestures, but using the actual listener backchannel feedback as our labels. In the second step, we merge the speaker experts (visual context) with listener experts by using a latent variable model. This process involves using the outputs of these expert models as an input to a Latent Dynamic Conditional Random Field (LDCRF) [12] that is capable of modeling the mutual influence between listener and speaker gestures.

The task of our LMDE model is to learn a mapping between a sequence of multimodal observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$ and a sequence of labels $\mathbf{y} = \{y_1, y_2, ..., y_m\}$. Each y_j is a class label for the jth frame of a video sequence and is a member of a set $\mathcal{Y}$ of possible class labels, for example, $\mathcal{Y} = \{$`backchannel`, `no feedback`$\}$. Each frame observation x_j is represented by a feature vector $\in \mathbf{R}^d$, for example, the prosodic features at each sample. For each sequence, we also assume a vector of "sub-structure" variables $\mathbf{h} = \{h_1, h_2, ..., h_m\}$. These variables are not observed in the training examples and will therefore form a set of hidden variables in the model.

Following Morency et al. [12], we define mutual-LMDE model as follows:

$$P(\mathbf{y} \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}:\forall h_j \in \mathcal{H}_{y_j}} P(\mathbf{h} \mid \mathbf{x}, \theta). \tag{1}$$

where θ are model parameters learned during training and $P(\mathbf{h}| \mathbf{x}, \theta)$ is defined as follows:

$$P(\mathbf{h}| \mathbf{x}, \theta) = \frac{\exp\begin{pmatrix} \sum_p \theta_p \cdot \mathbf{T}_p(\mathbf{h}) + \\ \sum_l \theta_l \cdot \mathbf{S}_l(\mathbf{h}, \mathbf{x}) + \sum_s \theta_s \cdot \mathbf{S}_s(\mathbf{h}, \mathbf{x}) \end{pmatrix}}{\mathcal{Z}(\mathbf{x'}, \theta)}, \tag{2}$$

Different from Ozkan et al. [9] and Morency et al. [12], we learn three sets of θ parameters: (1) θ_p related to the transition between hidden states, (2) θ_l related to *listener* expert outputs, and (3) θ_s related to *speaker* expert outputs. θ_s and θ_l model the relationships between expert outputs and the hidden states h_j. $\mathcal{Z}$ is the partition function. $\mathbf{T}_p(\mathbf{h}, \mathbf{x'})$ is the transition function between the hidden states. $\mathbf{S}_l(\mathbf{h}, \mathbf{x})$ is the listener state function and is defined as follows:

$$\mathbf{S}_l(\mathbf{h}, \mathbf{x}) = \sum_j s_s(h_j, [q_{j_1} q_{j_2} .. q_{j_\alpha} .. q_{j_{|e|}}]) \tag{3}$$

Each q_{j_α} is the marginal probability of expert α at frame j, and equals to $P_\alpha(y_j = a|\mathbf{x}, \lambda_\alpha)$. Each expert conditional distribution is defined by $P_\alpha(\mathbf{y}|\mathbf{x}, \lambda_\alpha)$ using the usual conditional random field formulation:

$$P_\alpha(\mathbf{y}|\ \mathbf{x}, \lambda_\alpha) = \frac{\exp\left(\sum_k \lambda_{\alpha,k} \cdot \mathbf{F}_{\alpha,k}(\mathbf{y}, \mathbf{x})\right)}{\mathcal{Z}_\alpha(\mathbf{x}, \lambda_\alpha)}, \tag{4}$$

where λ_α represent the model parameters of each expert α. $\mathbf{F}_{\alpha,k}$ is either a state function $s_k(y_j, \mathbf{x}, j)$ or a transition function $t_k(y_{j-1}, y_j, \mathbf{x}, j)$. Each expert α contains a different subset of state functions $s_k(y, \mathbf{x}, j)$, defined in Section 4.3.

Speaker state function $\mathbf{S}_s(\mathbf{h}, \mathbf{x})$ is defined similar to $\mathbf{S}_l(\mathbf{h}, \mathbf{x})$. The main difference is that, we use listener backchannels as sequence labels, $\mathbf{y}$, when learning $P_\alpha(\mathbf{y}|\ \mathbf{x}, \lambda_\alpha)$ for listener experts $\mathbf{S}_l(\mathbf{h}, \mathbf{x})$, and use speaker gestures as sequence labels $\mathbf{y}$ for speaker experts $\mathbf{S}_s(\mathbf{h}, \mathbf{x})$.

In our framework, each speaker expert learns a different aspect of speech for speaker gestures. Similarly, the listener experts allows us to obtain discriminative characters of speech for listener backchannel feedbacks. By using a latent variable model to combine these individual experts, our mutual-LMDE model is able to learn both the mutual influence between the speaker and the listener, and the hidden structure among the experts. More details about training and inference of LMDE can be found in Ozkan et al. [9].

4 Experimental Setup

As mentioned in the previous section, we evaluate our mutual-LMDE on the multimodal task of predicting listener nonverbal backchannel. In this section, we first describe the dataset, the gesture and backchannel annotation technique and multimodal speaker features. Then, we explain the baseline models used for comparison in our tests, and the experimental setup.

4.1 Dataset

We are using the RAPPORT dataset from [4], which contains 45 dyadic interactions between a speaker and a listener. Data is drawn from a study of face-to-face narrative discourse ("quasi-monologic" storytelling). In this dataset, participants in groups of two were told they were participating in a study to evaluate a communicative technology. Subjects were randomly assigned the role of speaker and listener. The speaker viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. After the speaker finished viewing the video, the listener was led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. The listener was asked to not talk during the story retelling. Elicited stories were approximately two minutes in length on average. Participants sat approximately 8 feet apart. All video sequences were manually transcribed and manually annotated to determine the ground truth backchannels. The next section describes our annotation procedure.

4.2 Gesture and Backchannel Annotations

In our experiments, we focus on visual backchannels of a listener: head nods. Similarly, we use speaker head nods as speaker nonverbal behaviors. A head nod gesture starts when the person starts moving his/her head vertically. The head nod gesture ends when the person stops moving or when a new head nod is started. A new head nod starts if the amplitude of the current head cycle is higher than the previous head cycle. Some listeners' responses may be longer than others although they all correspond to one single respond. In our data, annotators found a total of 666 head nods. The duration of these nods varied from 0.16 seconds to 7.73 seconds. Mean and standard deviation of backchannel durations are 1.6 and 1.2 respectively. The minimum number of head nods given by one listener during one interaction is 1, the maximum is 47, mean and standard deviations are 14.8 and 10.9 respectively.

Following Ward and Tsukahara's [14] original work on backchannel prediction, we train our models to predict only the start time of the backchannel start cue (i.e. head nod). Following again Ward and Tsukahara [14], we define the backchannel duration as a window of 1.0 seconds centered around the start time of the backchannel. A backchannel cue will be correctly predicted if at least one prediction of our LMDE model happens during this 1.0 seconds duration. All models tested in this paper use this same testing backchannel duration of 1.0 seconds.

4.3 Multimodal Features and Experts

This section describes the different multimodal audio features used to create our four experts.

Prosody. Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody [27, 14, 28]. For example, Ward and Tsukahara [14] show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. Listener feedback often follows speaker pauses or filled pauses such as "um" (see [28]). Using openSMILE [29] toolbox, we extract the following prosodic features, including standard linguistic annotations and the prosodic features suggested by Ward and Tsukhara:

- downslopes in pitch continuing for at least 40ms
- regions of pitch lower than the 26th percentile continuing for at least 110ms (i.e., lowness)
- drop or rise in energy of speech (i.e., energy edge)
- fast drop or rise in energy of speech (i.e., energy fast edge)
- vowel volume (i.e., vowels are usually spoken softer)
- pause in speech (i.e., no speech)

Lexical. Some studies have suggested an association between lexical features and listener feedback [28]. Using the transcriptions, we included all individual words (i.e., unigrams) spoken by the speaker during the interactions.

Part-of-Speech Tags. In [28], combination of pause duration and a statistical part-of-speech language model is shown to achieve the best performance for placing backchannels. Following this work, we use a CRF part-of-speech (POS) tagger to automatically assign a part of speech label to each word. We also include these part-of-speech tags (e.g. noun, verb, etc.) in our experiments.

Syntactic Structure. Finally, we attempt to capture syntactic information that may provide relevant cues by extracting three types of features from a syntactic dependency structure corresponding to the utterance. The syntactic structure is produced automatically using a data-driven left-to-right shift-reduce dependency parser [30], trained POS on dependency trees extracted from the Switchboard section of the Penn Treebank [31], converted to dependency trees using the Penn2Malt tool[1]. The three syntactic features are:

– Grammatical function for each word (e.g. subject, object, etc.), taken directly from the dependency labels produced by the parser
– Part-of-speech of the syntactic head of each word, taken from the dependency links produced by the parser
– Distance and direction from each word to its syntactic head, computed from the dependency links produced by the parser

Although our current method for extracting these features requires that the entire utterance be available for processing, this provides us with a first step towards integrating information about syntactic structure in multimodal prediction models. Many of these features could in principle be computed incrementally with only a slight degradation in accuracy, with the exception of features that require dependency links where a word's syntactic head is to the right of the word itself. We leave an investigation that examines only syntactic features that can be produced incrementally in real time as future work.

4.4 Baseline Models

Individual Experts. Our first baseline model consists of a set of CRF chain models, each trained with different set of multimodal features (as described in the previous section). In other words, only visual, prosodic, lexical or syntactic features are used to train a single CRF expert. (See Figure 3a).

Multimodal Classifiers. Our second baseline consists of two models: CRF and LDCRF [12]. To train these models, we concatenate all multimodal features (lexical, syntactic and prosodic) in one input vector. Graphical representation of these baseline models are given in Figure 3-(a) and Figure 3-(b).

[1] `http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html`

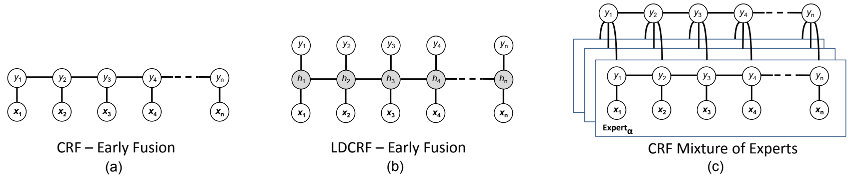


Fig. 3. Baseline Models: **a)** Conditional Random Fields (CRF), **b)** Latent Dynamic Conditional Random Fields(LDCRF), **c)** CRF Mixture of Experts (no latent variable)

LMDE. To show the importance of visual context from the speaker, we train an LMDE model without using any of the speaker experts. In other words, our baseline LMDE model is trained to directly predict listener backchannels from the speaker audio features.

Pause-random Classifier. Random backchannel generator randomly generates backchannels whenever some pre-defined conditions in the speech is purveyed. These conditions include pauses that come after at least 700 milliseconds of speech and absence of backchannel feedback within the preceeding 800 milliseconds. This random classifier has also been used by Ward and Tsukahara [14] for comparison.

CRF Mixture of Experts. To show the importance of latent variable in our context-based prediction model, we trained a CRF-based mixture of discriminative experts. A graphical representation of a CRF Mixture of experts is given in Figure 3. This model is similar to the Logarithmic Opinion Pool (LOP) CRF suggested by Smith et al. [13], in the sense that they both factor the CRF distribution into a weighted product of individual expert CRF distributions. The main difference between LOP and CRF Mixture of Experts model is in the definition of optimization functions. Training of CRF Mixture of Experts is performed in two steps: Expert models are learned in the first step, and the second level CRF model parameters are learned in the second step.

LMDE with Speaker Nods. Our final set of baseline models include an LMDE model that directly uses the visual context from the speaker (speaker nods). In this baseline model, we first train only the listener expert models as in the first step of our proposed approach. Then, in the second step, we use the annotated (actual) speaker gestures together with the listener experts as input to the latent variable model. So, the main difference of this baseline model with our approach is that our approach first anticipates the speaker nonverbal behaviors through CRF experts instead of directly using them.

4.5 Methodology

We performed held-out testing by randomly selecting a subset of 11 interactions (out of 45) for the test set. The training set contains the remaining 34 dyadic interactions. All models in this paper were evaluated with the same training and test sets. Validation of all model parameters (regularization term and number of hidden states) was performed using a 3-fold cross-validation strategy on the training set. The regularization term was validated with values $10^k, k = -1..3$. Two different number of hidden states were tested for the LDCRF models: 3, and 4 (note that LDCRF with 1 hidden state is equivalent to Mixture of CRF Experts model).

The performance is measured by using the conventional metrics: precision, recall, and F-measure. Precision is the probability that predicted backchannels correspond to actual listener behavior. Recall is the probability that a backchannel produced by a listener in our test set was predicted by the model. We use the same weight for both precision and recall, so-called F_1, which is the weighted harmonic mean of precision and recall. F_1 scores for each sequence is calculated first, then the final F_1 result is computed by averaging these sequence scores.

Before reviewing the prediction results, is it important to remember that backchannel feedback is an optional phenomena, where the actual listener may or may not decide on giving feedback [14]. Therefore, results from prediction tasks are expected to have lower accuracies as opposed to recognition tasks where labels are directly observed (e.g., part-of-speech tagging).

During testing, we find all the "peaks" (i.e., local maxima) from marginal probabilities $P(y_j = a|\ \mathbf{x}, \theta)$. For the f1-score, the prediction model needs to decide on a specific threshold (i.e., amount of backchannel) for the marginal probabilities for all users. The value of this threshold is automatically set during validation. Since we are predicting the start time of a backchannel, an actual listener backchannel is correctly predicted if at least one model prediction happen within the 1 second interval window around the start time of the listener backchannel.

The training of all CRFs and LDCRFs were done using the hCRF library[2]. The LMDE model was implemented in Matlab based on the hCRF library. The input observations were computed at 30 frames per second. Given the continuous labeling nature of our model, prediction outputs were also computed at 30Hz.

5 Results

In this section we present the results of our empirical evaluation. We designed our experiments to test different characteristics of our mutual-LMDE approach: (1) integration of multiple sources of information, and (2) mutual influence.

Performances of individual CRF experts for predicting listener backchannels and speaker gestures are presented in Table 1. Our approach combines all these experts to model the mutual influence between the speaker and the listener. This integration of multiple resources improve the prediction accuracy for listener backchannels. Therefore, we get an f-1 score of 0.32 with our mutual-LMDE model.

[2] `http://sourceforge.net/projects/hrcf/`

Table 1. Test performances of the individual expert models for listener backchannel and speaker gesture (head nod) predictions

	Listener			**Speaker**		
Expert	f1	Precision	Recall	f1	Precision	Recall
Prosodic	0.1913	0.1060	0.9803	0.2789	0.1669	0.8478
Lexical	0.2073	0.1377	0.4198	0.2959	0.2068	0.5203
POS	0.2346	0.1446	0.6220	0.3274	0.2182	0.6556
Syntactic	0.2045	0.1287	0.4956	0.3175	0.2330	0.4983
mutual-LMDE	**0.3212**	0.2633	0.4117	**0.3313**	0.2456	0.5087

Table 2. Comparison of different models with our approach

Model	f1	Precision	Recall
Early CRF	0.2173	0.1423	0.4591
Early LDCRF	0.2115	0.1231	0.7495
LMDE	0.2764	0.2055	0.4219
Pause-Random	0.1456	0.1322	0.2031
CRF Mixture	0.1963	0.1718	0.2288
LMDE+Speaker Nods	0.2614	0.2071	0.3541
mutual-LMDE	**0.3212**	0.2633	0.4117

In our second set of experiments, we evaluate the importance of modeling mutual influence. Table 2 summarizes our results. The prediction models in the top three rows of the table do not take into account the mutual influence between the speaker and the listener. These models are trained on the speaker audio features to directly infer the listener backchannels. Among these models, LMDE gives the best f-1 score, which proves the importance of late fusion of multiple sources of information (different speech channels). However, our mutual-LMDE model outperform all these three models, which indicates the importance of using mutual influence between the interlocutors.

The models listed in the last three rows of Table 2, model the mutual influence. CRF Mixture model does not perform as good as other LMDE models. The main reason for this decrease in performance is that the LMDE model uses a latent variable to capture the dynamic among different sources of information, whereas the CRF Mixture approach directly models these information. Although the last LMDE approach use the speaker nonverbal behavior information directly in the second step of LMDE, it does not perform as good as our mutual-LMDE model, in which we first infer these speaker behaviors instead of directly using them. We hypothese that, by inferring the speaker backchannels, we are able to model a better average speaker feedback behavior and remove the variations in the actual speaker backchannels.

Our framework addresses the problem of listener backchannel prediction by modeling the mutual influence. A related issue is modeling the recursive influence between the listener and the speaker. For instance, backchannels of a listener might trigger more visual gestures from the speaker. Although we do not explicitly model this recursive influence in our current study, the proposed framework can be extended to address this issue as well. For instance, we can use the listener observations (features) in the learning process for speaker experts to model how listener behaviors affect speaker behaviors. The study of these recursive models is part of our future work.

6 Conclusions

In this paper, we proposed a context-based approach for predicting the backchannels of a listener in a dyadic conversation. To model the mutual influence between the speaker and the listener, we used a variant of Latent Mixture of Discriminative Experts model. Our mutual-LMDE approach consists of two steps: we first learn expert models to predict speaker gestures (head nods), and the listener backchannel feedbacks. Then, we use visual context (predicted speaker gestures) from the speaker to improve the final listener backchannels.

We evaluated our approach on 45 dyadic interactions from the RAPPORT dataset. Our experiments have shown improvement over all previous approaches. The results suggest two main conclusion: (1) By modeling the mutual influence between the participants of a dyadic interaction, we can better model the backchannel feedbacks of the listener. (2) In case of no available visual speaker information, predicted speaker visual context helps us to learn an average speaker behavior that is more effectual and less noisy than actual speaker behaviors.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Ross, M.D., Menzler, S., Zimmermann, E.: Rapid facial mimicry in orangutan play. Biol. Lett. 4, 27–30 (2008)
2. Hatfield, E., Cacioppo, J., Rapson, R.: Emotional contagion. In: Clark, M.S. (ed.) Review of Personality and Social Psychology: Emotion and Social Behavior, pp. 151–171 (1992)
3. Riek, L.D., Paul, P.C., Robinson, P.: When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. Journal on Multimodal User Interfaces 3, 99–108 (2010)
4. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)
5. Drolet, A.L., Morris, M.W.: Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. Journal of Experimental Social Psychology 36(1), 26–50 (2000)

6. Tsui, P., Schultz, G.: Failure of rapport: Why psychotheraputic engagement fails in the treatment of asian clients. American Journal of Orthopsychiatry 55, 561–569 (1985)
7. Fuchs, D.: Examiner familiarity effects on test performance: implications for training and practice. Topics in Early Childhood Special Education 7, 90–104 (1987)
8. Burns, M.: Rapport and relationships: The basis of child care. Journal of Child Care 2, 47–57 (1984)
9. Ozkan, D., Morency, L.P.: Latent mixture of discriminative experts. IEEE Transactions on Multimedia 15(2), 326–338 (2013)
10. Morency, L.P., de Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Conference on Intelligent Virutal Agents, IVA (2008)
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labelling sequence data. In: International Conference on Machine Learning, ICML (2001)
12. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEE Conference on Computer Vision and Pattern Recognition, CVPR (2007)
13. Smith, A., Cohn, T., Osborne, M.: Logarithmic opinion pools for conditional random fields. In: Association for Computational Linguistics (ACL), pp. 18–25 (2005)
14. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in english and japanese. Journal of Pragmatics 23, 1177–1207 (2000)
15. Pantic, M., Pentland, A., Nijholt, A., Huang, T.: Human computing and machine understanding of human behavior: A survey. In: ACM International Conferance on Multimodal Interfaces, pp. 239–248 (2006)
16. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(3), 311–324 (2007)
17. Sebea, N., Cohenb, I., Netherl, T.: Multimodal approaches for emotion recognition: A survey (2005)
18. Maatman, R.M., Gratch, J., Marsella, S.: Natural behavior of a listening agent. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 25–36. Springer, Heidelberg (2005)
19. Nakano, Y., Reinstein, G., Stocky, T., Cassell, J.: Towards a model of face-to-face grounding. In: Association for Computational Linguistics, ACL (2003)
20. Nakano, Y., Murata, K., Enomoto, M., Arimoto, Y., Asa, Y., Sagawa, H.: Predicting evidence of understanding by monitoring user's task manipulation in multimodal conversations. In: Association for Computational Linguistics (ACL), pp. 121–124 (2007)
21. Ward, N.: Non-lexical conversational sounds in American English (2003)
22. Fujie, S., Ejiri, Y., Nakajima, K., Matsusaka, Y., Kobayashi, T.: A conversation robot using head gesture recognition as para-linguistic information. In: IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 159–164 (2004)
23. Kang, S.H., Gratch, J., Wang, N., Watt, J.: Does the contingency of agents' nonverbal feedback affect users' social anxiety? In: International Conference on Autonomous Agents and Multiagent Systems, AAMAS (2008)
24. Semaine the sensitive agent project
25. Gravano, A.: Turn-taking and affirmative cue words in taskoriented dialogue. Technical report (2009)

26. Neiberg, D.: Modelling Paralinguistic Conversational Interaction: Towards social awareness in spoken human-machine dialogue. PhD thesis, KTH, Speech Communication and Technology, QC 20120914 (2012)
27. Nishimura, R., Kitaoka, N., Nakagawa, S.: A spoken dialog system for chat-like conversations considering response timing. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 599–606. Springer, Heidelberg (2007)
28. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: European Chapter of the Association for Computational Linguistics (EACL), pp. 51–58 (2003)
29. Eyben, F., Wöllmer, M., Schuller, B.: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Affective Computing and Intelligent Interaction (ACII), pp. 576–581 (2009)
30. Sagae, K., Tsujii, J.: Dependency parsing and domain adaptation with LR models and parser ensembles. In: Association for Computational Linguistics (ACL), pp. 1044–1050 (2007)
31. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The penn treebank: annotating predicate argument structure. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 114–119 (1994)

Modeling the Semantic Coordination of Speech and Gesture under Cognitive and Linguistic Constraints

Kirsten Bergmann, Sebastian Kahl, and Stefan Kopp

SFB 673, CITEC, Faculty of Technology, Bielefeld University, Germany
{kbergman,skahl,skopp}@techfak.uni-bielefeld.de

Abstract. This paper addresses the semantic coordination of speech and gesture, a major prerequisite when endowing virtual agents with convincing multimodal behavior. Previous research has focused on building rule- or data-based models specific for a particular language, culture or individual speaker, but without considering the underlying cognitive processes. We present a flexible cognitive model in which both linguistic as well as cognitive constraints are considered in order to simulate natural semantic coordination across speech and gesture. An implementation of this model is presented and first simulation results, compatible with empirical data from the literature are reported.

Keywords: Speech, gesture, generation, cognitive modeling, semantic coordination.

1 Introduction

Intelligent virtual agents are required to be able to express themselves convincingly and autonomously. There is a growing body of evidence demonstrating the importance of nonverbal expressivity for this purpose. Especially co-speech gestures play a major role in providing agents with convincing communication skills, because they are an integral part of human communication, inseparably intertwined with speech [24]. It has, for instance, been demonstrated that a virtual agent's use of gestures might enhance the perceived verbal eloquence [7] as well as users' ratings of the agent, mostly in terms of competence [3]. State of the art agent systems, however, still "remain a long way from matching the complexity and subtlety of real-life nonverbal behavior" [21, p. 80]. One major reason for this is that many aspects of how humans use and produce speech and gestures in combination are not fully understood, yet.

There is considerable theoretical disagreement regarding the process by which semantic coordination between speech and gesture is achieved. This involves two major issues, information distribution and information packaging. Information distribution means that the two modalities, although expressing the same underlying idea, do not convey necessarily identical aspects of it: Gestures can be found to be *redundant* with the information encoded verbally (e.g., 'round cake'

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 203–216, 2013.

+ gesture depicting a round shape), to *supplement* it (e.g., 'cake' + round gesture), or even to *complement* it (e.g., 'looks like this' + round gesture). Closely related to how information is distributed across modalities, is information packaging: How much information is put into a multimodal utterance? When are messages rather split into several parts? With regard to our example this means that the above utterance might also be split into two clauses, e.g., 'there is a cake' followed by 'and it is round', whereby both clauses might be accompanied by a gesture, e.g. a supplementary pointing gesture with the first clause, and a redundant shape-depicting gesture accompanying the second clause.

Empirical evidence suggests that both cognitive and linguistic constraints are involved in the process of meaning coordination. However, a concrete and comprehensive picture of how variations in meaning coordination arise under those constraints is still missing. In recent work we developed a cognitive model for the production of speech and gestures [19]. This model already covers *formulation* processes determining the surface form of speech and gestures as well as the *conceptualization* process by which meaning is structured, portioned and distributed across the two modalities. The latter is based on activation-spreading within dynamically shaped multimodal memories. We have shown how this model can simulate and explain different cases of information distribution under cognitive constraints.

In this paper we present an extension of the model to also consider linguistic constraints and, consequently, information packaging. We start with a review of empirical findings, followed by a discussion of related work. We then present our model which provides a detailed cognitive account of how meaning is dynamically organized and coordinated across speech and gesture. Finally, we present new modeling results demonstrating that the model can account for empirical findings as reported in the literature.

2 Background and Related Work

2.1 Information Distribution across Speech and Gesture

A couple of studies have investigated how the frequency and nature of gesturing, including its coordination with speech is influenced by *cognitive factors*. Bavelas et al. [2], for instance, found that speakers are more likely to produce non-redundant gestures when their addressees could see them, as opposed to when their gestures are not visible and hence less essential for their partners. Bergmann and Kopp [4] report results from an analysis of natural co-verbal gesturing in direction-giving, indicating that supplementary gestures are more likely in cases of problems of speech production (e.g. disfluencies) or when the information conveyed is introduced into the dialogue (and thus conceptualized for the first time). In line with this, recent work has suggested that speakers indeed produce more gestures at moments of relatively high load on the conceptualization process for speaking [17], in particular on the linearization and the focusing components of conceptualization [25].

Hostetter and Alibali [13] report findings suggesting that speakers who have stronger visual-spatial skills than verbal skills produce higher rates of gestures than other speakers. In a later study, Hostetter and Alibali [14] found that the speakers with high spatial skills also produced a higher proportion of non-redundant gestures than other speakers, whereas verbal-dominant speakers tended to produce such gestures more in case of speech disfluencies. The authors hypothesize that "*non-redundant gesture-speech combinations occur because mental images are more active in speaker's minds at the moment of speaking than are verbal codes*" [p.45]. Taken together this suggests that non-redundant gesture-speech combinations are the result of speakers having both strong spatial knowledge and weak verbal knowledge simultaneously, and avoiding the effort of transforming the one into the other.

2.2 Information Packaging for Speech and Gesture

Empirical evidence investigating speech-gesture information packaging suggests that gestures are influenced by *linguistic constraints* in terms of conceptual, syntactic, and lexical structure of concomitant speech. In a cross-linguistic study Kita and Özyürek [15] demonstrated that the packaging of content for gestures parallels linguistic information packaging. Speakers of Japanese, Turkish and English had to re-tell cartoon events for which their languages provide differing means of encoding. English speakers, for example, used the verb 'swing' for a character's action, encoding an arc-shaped trajectory, while Turkish and Japanese speakers employed a trajectory-neutral, change-of-location predicate such as 'move'. Gestures followed this packaging in a way that Japanese and Turkish speakers were more likely to produce straight gestures, whereas most English speakers produced arced gestures. In another cartoon-event, the character rolled down a hill. Again, speakers of English typically described this by combining manner and path of the movement in a single clause (e.g. "he rolled down"), accompanied by a single gesture encoding both semantic features. In contrast, Turkish and Japanese speakers encode manner and path separately in two clauses (e.g. "he descended as he rolled") and also used two separate gestures for these two features.

Evidence along the same line comes from a study on language acquisition [27]: Advanced L2 speakers of English typically encoded manner and path information in one clause and their gestures followed, whereas speakers at lower proficiency levels typically used two-clause constructions in speech, accompanied by separate gestures for manner and path. A subsequent study [16] showed that this effect also occurs when L1 speakers of English are forced to produce one- or two-clause descriptions of manner and path.

Kita and Özyürek [15] proposed an explanatory models for these empirical findings. This account explicitly incorporates the idea that language shapes iconic gestures such that the content of a gesture is determined by three factors: (1) the speaker's communicative intention, (2) action schemata selected on the basis of features of imagined or real space, and (3) bidirectional interactions between speech and gesture production processes. The latter takes place at the level of conceptualization, i.e. the organization of meaning. An additional interaction between the speech formulator and the (preverbal) message generator is assumed to allow for feedback from grammatical or phonological encoding to the conceptualizer and thus to gesture.

2.3 Computational Models of Speech and Gesture Production

Computational modeling of speech-gestural communicative behavior for virtual agents mostly focused on how gesture use is constrained by linguistic features, while cognitive constraints remained completely disregarded. Existing approaches basically fall into two groups depending on how they bring speech and gestures together: *rule-based* and *data-driven* approaches. Among the rule-based models, the BEAT system [10], for instance, was based on behavior generators in which generation rules extracted from empirical data were implemented. This way the framework considered linguistic information such as information structure for the selection of predefined gesture specifications. A similar approach was taken in the Nonverbal Behavior Generator (NVBG) [22]. The system analyzes the syntactic and semantic structure of surface texts and takes the affective state of the virtual agent into account to generate appropriate nonverbal behaviors. Based on a study from the literature and a video analysis of emotional dialogues, the authors developed a list of nonverbal behavior generation rules. In the REA architecture [9] gestures were lexicalized like words and selected using a lexical choice algorithm and incorporated directly into natural language generation. In particular, it implemented rules of information distribution to account for the fact that speech and gestures are sometimes redundant and sometimes complementary. The NUMACK account [20] followed the same strategy by using an integrated microplanner (SPUD) to compose multimodal utterances. Extended with a flexible gesture planner (instead of using a static set of predefined gestures), gestures were dynamically incorporated into SPUD's resources and utilized in grammatically pre-determined ways.

Data-driven models have been adopted mainly to account for individual or cultural style of co-verbal gesturing. Neff et al. [26] developed a system generating gesture animations for novel text by using speaker-specific gesture profiles that were created from a corpus of communicative behavior. Based on these profiles, the system made probabilistic generation choices conditioned upon the previously performed gesture and the input text tagged with theme, rheme, and focus. Similarly, Endrass et al. [12] proposed a corpus-driven method of generating gestures in a culture-specific way that accompany virtual agent's verbal utterances. The frequency of gestures and gesture-types, the correlation of gesture-types and speech-acts as well as the expressivity of gestures have been analyzed in the two cultures of Germany and Japan and integrated into a generation model.

Apart from models developed for virtual agents, [8] have proposed a cognitive modeling attempt for the production of speech and gestures, using the cognitive architecture ACT-R [1]. This account draws on two major assumptions: (1) on the claim that language representations include some irreducibly spatial components; (2) on the idea that language processing is based on constructions which consist of both semantic and syntactic components. The authors assume these constructions to prescribe spatial representations for what they call *linguistic spatial gestures* and which they assume to provide only "*little information not included in the accompanying language*" [p.14].

In sum, the fact that the virtual agent community has undertaken multiple different efforts to develop production accounts in which gestures are constrained by their linguistic context clearly shows the relevance of such models. This is particularly emphasized by the recent trend of cross-culturally (or cross-linguistically) employed agents. However, existing generation systems basically solve this by combining or switching between static models, rule-based or data-based, explicitly developed for a specific language, culture or individual. What we present in this paper is a flexible cognitive model that simulates natural semantic coordination of speech and iconic gesture, accounting for information distribution and packaging under dynamically arising or changing linguistic and cognitive constraints.

3 A Cognitive Model

In order to investigate to what extent information distribution and information packaging across modalities can be explained by cognitive and linguistic constraints we developed a model based on activation-based processing on multimodal memory [19]. This account is embedded in a larger production model that comprises three stages: conceptualization, where a *message generator* and an *image generator* work together to select and organize information to be encoded in speech and gesture, respectively; formulation, where a *speech formulator* and a *gesture formulator* determine appropriate verbal and gestural forms for this; *motor control* and *articulation* to finally execute the behaviors. Motor control, articulation, and formulation have been subject of earlier work [5].

3.1 Multimodal Memory

The central component in our model is a multimodal memory which is accessible by modules of all processing stages. We assume that language production requires a preverbal message to be formulated in a symbolic-propositional representation that is linguistically shaped [23] (SPR, henceforth). During conceptualization the SPR, e.g. a function-argument structure denoting a spatial property of an object, often needs to be extracted from visuo-spatial representations (VSR), e.g. the mental image of this object. We assume this process to involve the invocation and instantiation of memorized supramodal concepts (SMC, henceforth), e.g. the concept 'round' which links the corresponding visuo-spatial properties to a corresponding propositional denotation. Fig. 1 illustrates the overall relation of these tripartite multimodal memory structures.

To realize the VSR and part of the SMC, we employ a model of visuo-spatial imagery called *Imagistic Description Trees* (IDT) [28]. The IDT model was designed, based on empirical data, to cover the meaningful visuo-spatial features in shape-depicting iconic gestures. Each node in an IDT contains an imagistic description which holds a schema representing the shape of an object or object part. Important aspects include (1) a tree structure for shape decomposition, with abstracted object schemas as nodes, (2) extents in different dimensions as an approximation of shape, and (3) the possibility of dimensional information to be underspecified. The latter occurs, e.g., when the axes of an object schema cover less than the three dimensions of space or when an exact dimensional extent is left open but only a coarse relation between axes like "dominates" is given. This allows to represent the visuo-spatial properties of SMCs such as 'round', 'left-of' or 'longish'. Applying SMC to VSR is realized through graph unification and similarity matching between object schemas, yielding similarity values that assess how well a certain SMC applies to a particular visuo-spatially represented entity (cf. Fig. 1). SPR are implemented straight forward as predicate-argument sentences.

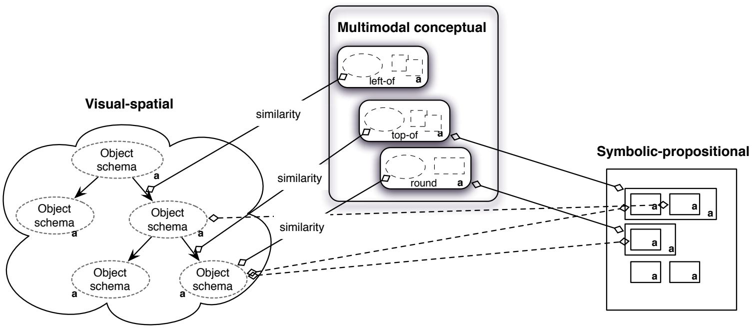


Fig. 1. Tripartite multimodal memory structure involved in speech and gesture production

3.2 Formulators and Generators

The *message generator* has to pre-package the activated SPR information in a way that the *speech formulator* can produce an appropriate sentence. We employ an LTAG-based sentence planner for speech formulation. To make sure that all facts necessary to generate a verbal utterance are available, the *message generator* applies networks that reflect the options of verbalization provided in the *speech formulator*'s LTAG grammar (this conforms the view that the conceptualizer learns to anticipate the formulator's abilities [23]).

The *image generator* retrieves visuo-spatial information about the object to be described, from activated and salient VSR and SMC entries in memory. It is in charge of unifying this information into an imagistic description, from which the *gesture formulator* can derive a gesture form specification (based on Bayesian decision networks learned from empirical data [5]). For instance, information about shape might be combined with information about the object's size or position (as encoded in the IDT representation of VSR entries). Depending on the knowledge encoded here, the *gesture formulator* is able to plan a shape-depicting gesture or rather a localizing gesture.

3.3 Overall Production Process

Fig. 2 shows an outline of the overall production architecture. We assume that conceptualization consists of cognitive processes that operate upon the multimodal memory structures and are constrained by principles of memory retrieval, which can be modeled by principles of activation spreading [11]. As assumed in cognitive architectures like ACT-R [1], activations float dynamically, spread across linked entities (in particular via SMCs), and decay over time. Activation of more complex SMCs are thereby assumed to decay slower than activations in VSR or SPR. For implementation details see [19].

Formulation-based Reinforcement. Information distribution is explained *via* a mechanism of formulation-based reinforcement: Propositions encoded in the speech formulators' first formulation suggestions result in reinforced activation of this concept in SPR memory, and thus increased activation of the associated concept in VSR. This is possible due to the fact that the speech formulator links words and semantics in terms of SPR entries. Similarly, imagistic representations encoded by the *gesture formulator* also result in respectively reinforced activations in VSR memory, spreading over to SPR memory. In result, multimodal coordination in terms of information distribution emerges from the local choices the generators and formulators take based on the activation dynamics in multimodally linked memory representations. Note that as activation is dynamic, which features are selected becomes dependent on the time of retrieval and, thus, available cognitive resources. Redundant speech and gesture, then, result from focused activation of supramodally linked mental representations, whereas non-redundant speech and gesture arise when activations in VSR and SPR scatter over entries that are not connected via SMC concepts.

Goal-based Reinforcement and Activation Decrease. Our account of information packaging rests upon two ideas: (1) goal-based intensification and (2) activation decrease when information is successfully conveyed. Production always starts with the *image generator* and *message generator* inducing activations in modal entries, evoked by a communicative intention (goal) such as "introduce churchwindow-1". Upon retrieval, the *generators* independently select features and pass them to the respective *formulator*. The communicative intention is held available in the generators as long as it has not been fulfilled, i.e., until all memory entries involved have been put into words or gestures. Notably, this need not take place in only one utterance. It might as well be that the communicative intention is split into two or more clauses. As soon as an utterance is passed to realization, the activation of associated concepts in VSR and SPR is decreased. Only those parts of the communicative intention that have not been realized so far are kept activated. As this might be a single concept, the process of goal-based reinforcement now ensures that enough contextual information is available for the generators to initiate the generation of an entire clause or sentence.

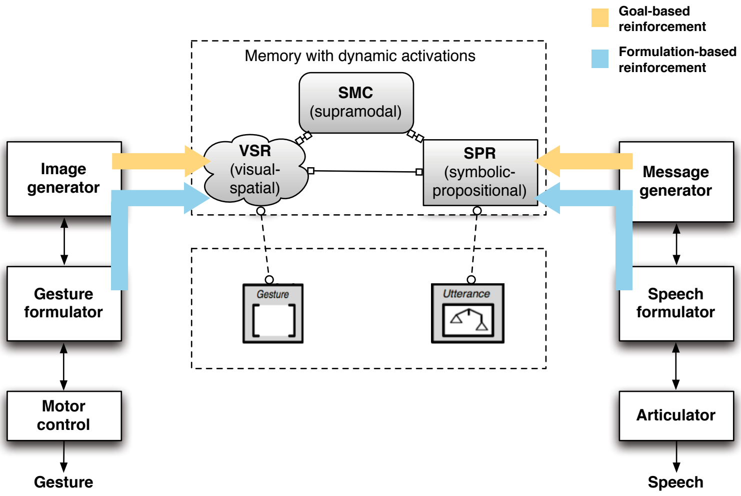


Fig. 2. Overall production architecture

4 Modeling Results

Our model has been implemented into our speech-gesture production architecture as a fully operational prototype that can directly be connected to standard behavior realization engines driving a virtual character. As a first exploration, we report results from a study on how (1) cognitive resources, specifically processing time, and (2) linguistic constraints in terms of limited verbalization ability affect meaning coordination in the produced multimodal utterances. We first describe the emerging production process in detail and then report quantitative results.

The production process is always initiated by setting the communicative intention "introduce churchwindow-1". Upon receiving this goal, the *image generator* activates visuo-spatial imagery of the church window in VSR, and the *message generator* activates symbolic representations of non-spatial semantic concepts in SPR. These activations spread through memory and lead to invocation of SMCs for, e.g., 'round' (bound to churchwindow-1) and 'at-top-of' (the church-tower), as well as instantiation of the corresponding SPR entries.

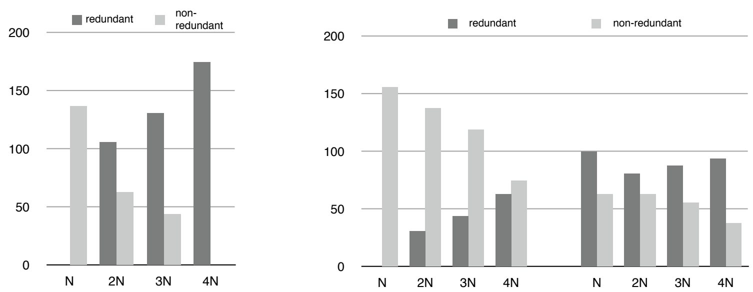
SMCs along with their linked entries in VSR and SPR attain highest and most slowly decaying activation values. Multimodal memory processing is based on so-called update cycles. Per cycle entry activations are being calculated, spread across associations and new associations are created if connections are sufficiently active. The frequency with which a memory cycle is being triggered can be modified with respect to the speed at which the system is required to run. After a preset number of processing cycles, both generators retrieve modality-specific information from memory with a probability depending on current activation values, leading to 'round' and 'at-top-of' concepts being encoded in speech and gesture in a less coordinated/aligned way. If there is enough processing time available, the *message generator* starts to collect concepts for the preverbal message and re-activates those entries being retrieved and used for *speech formulation*: the contents expressed either verbally or gesturally tend to converge. This results in well-coordinated multimodal representations when the modality-specific formulators finally start with their generation work. Thus, it is more likely that both generators receive the same information about shape and position of the entity. Accordingly, the *speech formulator* is enabled to plan a sentence like "The church has a round window at the top" accompanied by a shape depicting gesture like drawing the shape of the window in the air, or a static posturing gesture where the hands becoming a model of the circular shape. As the position of the entity is also available, the gesture would be performed in that part of gesture space.

Generation under cognitive constraints If the cognitive processing time available is short, goal-reinforced content coordination is restricted and the multimodal representations are less well coordinated when the modality-specific formulators start with their generation work. So the *message generator* may retrieve only information about the salient shape of the window, but not about its position relative to other entities. Thus, a sentence like "The church has a round window" gets formulated. The *image generator*, on the other hand, may receive information about the entity's position as well. This might result in gestures like the ones described previously–encoding both, shape and position information.

Generation under linguistic constraints To simulate cross-linguistic variation as described in [15], we assume as linguistic constraint that a speaker's grammar does not provide a noun phrase construction of the type 'DET ADJ NN', i.e., we modify the linguistic abilities embodied in the *message generator* and the *speech formulator*. The *speech formulator* is hence not able to return a one-sentence solution for all propositions. Instead, it proposes a solution like "The church has a window (at the top)" which does not convey 'round' bound to churchwindow-1. Depending on whether the accompanying gesture encodes the feature 'round', which is possible but not very likely as there is no formulation-based reinforcement of the SMC 'round', the deactivation process after utterance realization leaves the 'round' feature at a high activation level so that the generators again initiate an goal-based reinforcement of memory structures. Now the *speech formulator* builds a message plan for an utterance containing the 'round'-feature, resulting in a verbalization like 'the window is round'. Due to formulation-based activation of the shape property 'round' by the *message generator*, a shape-depicting gesture is likely to be planned to accompany this second utterance.

4.1 Quantitative Results

To quantify these observations, we ran simulation experiments in which we manipulated the time available (in terms of memory update cycles) and analyzed the resulting gestures for their semantic content and semantic coordination with speech in terms of redundancy/non-redundancy. In the simulations we constrained, first, processing time as a cognitive resource by forcing the system after a particular amount of cycles (N, 2N, 3N and 4N cycles) to realize an utterance based on the current memory state. Second, we manipulated the system's verbalization capabilities. In the 'full grammar' (FG) condition the 'DET ADJ NN' construction was available; in the 'limited grammar' (LG) condition the verbal ability of the *message generator* and *speech formulator* was impaired by making the 'DET ADJ NN' construction unavailable. In this condition the communicative intention had to be realized in two clauses. We ran the model 100 times in each condition.



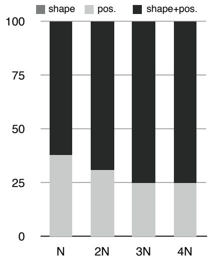
(a) Full grammar (one clause)

(b) Limited grammar (two consecutive clauses)

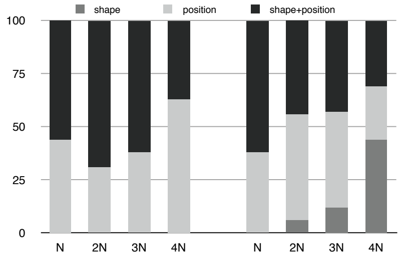
Fig. 3. Amount of semantic gesture features encoded redundantly vs. non-redundantly with speech in 100 simulations under manipulation of time available (N, 2N, 3N and 4N memory update cycles)

Figure 3 shows that non-redundant (supplementary) gestures dominate in those runs with stricter cognitive constraints, while redundant ones become more likely when processing time available is increased. As in the LG condition the communicative intention is realized *via* two separate clauses, the amount of verbally encoded information per clause is lower than in the FG condition. Accordingly, the accompanying gestures are more likely to be non-redundant perse.

Figure 4 shows the effect verbalization ability on semantic coordination. In the FG condition, gestures tend to convey both verbally encoded semantic features, position and shape. The more processing time is available, the more likely it is that gestures express both semantic features (in redundancy with speech). By contrast, in the LG condition, conflated gestures become less likely over time. Rather, gestures accompanying the first clause (*without* verbally encoded shape information) tend to express position information only, while the gestures accompanying the second clause (*with* verbally encoded shape information) tend to express shape information only, the more processing time is available. Moreover, the amount of conflated gestures conveying both semantic features is higher in the FG condition as compared to the LG condition, especially when utterances are produced with more processing time.



(a) Full grammar (one clause)



(b) Limited grammar (two consecutive clauses)

Fig. 4. Semantics encoded by gestures in 100 simulation under manipulation of time available (N, 2N, 3N and 4N memory update cycles)

5 Conclusion

We have presented an extended model to explain semantic coordination of speech and gesture in terms of (1) how visuo-spatial and symbolic-propositional memory structures are dynamically organized, (2) how these multimodal memory structures affect activation-spreading, and (3) how this interacts with modality-specific processes of conceptualization and formulation. The mechanisms of formu-lation-based and goal-based reinforcement of memory entries in combination with an activation decrease of successfully conveyed information enable to simulate two major empirical findings reported in literature. First, our model allows to simulate cross-linguistic data according to which gestural representations follow linguistic encoding patterns. When verbal capabilities allow for a compact verbalization of the communicative intention, gestures also tend to conflate the very semantics. When, by contrast, linguistic constraints require semantics to be separated into two clauses, this is reflected in gesture. Second, we have shown in simulation that the model also offers a natural account for the finding that non-redundant gestures are more likely when conceptualization load is high, based on the assumption that memory-based cross-modal coordination consumes resources (memory, time), and is reduced or compromised when such resources are limited.

This exemplifies how a flexible, cognitive account like ours can help to go beyond explicit modeling of the speech-gesture relationship. This way, we can improve human-agent interaction such that it progresses towards intuitive and human-like communication, and we can contribute to understanding the cognitive phenomena by making hypotheses testable in terms of predictions that can be explored in computational simulations as well as in appropriately set up empirical experiments. Future work will therefore be directed to further extend and evaluate the model regarding several respects.

First, we need to extend the model towards evaluation principles to decide upon the moment when utterance planning can be terminated. In other words, when is a multimodal utterance 'good enough' to be realized? This issue is closely related to the question whether a speech-gesture ensemble encodes the communicative intention adequately. Is it, for instance, sufficient to encode shape information only gesturally (as in ‘there is a church window’ + shaping gesture)? Or do we need a follow-up clause encoding the shape information verbally? A cost-based “good-enough” generation will enable the system to make such decisions autonomously, e.g., contingent upon addressee presence (cf. [29]) or addressee feedback.

Second, we aim to investigate speech-gesture use beyond single ensembles, by considering larger discourse units including sequences of contributions as well as multi-part gestures. Here we need to consider alignment effects: Speakers adapt their utterances with respect to their own or their interlocutor’s previous communicative behavior. We have already analyzed such gestural alignment empirically in [6] and found that a speaker’s own gestures influence each other more than the gestures the interlocutor performs. Moreover, not all gesture features seem to be equally subject to this effect. For a first sketch of how to account for these findings in a larger perception-action architecture see [18]. Moreover, there is the necessity of planning sequences of several consecutive utterances which also requires to consider the context in terms of what has been said and gesticulated before.

Finally, one might consider cognitive constraints not only as processing time, as we did in our initial modeling attempt here, but also in terms of other factors occurring in human cognition like spatial skills (cf. [13,14]) or conceptualization load (cf. [25]). A systematical manipulation of such model parameters with analyses of the resulting behavior will provide further insights into the production of multimodal utterances and a more comprehensive picture of how different factors and mechanisms act in concert to produce overt communicative behavior.

Acknowledgements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 “Alignment in Communication” and the Center of Excellence 277 “Cognitive Interaction Technology” (CITEC).

References

1. Anderson, J., Bothell, D., Byrne, M., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111(4), 1036–1060 (2004)
2. Bavelas, J., Kenwood, C., Johnson, T., Philips, B.: An experimental study of when and how speakers use gestures to communicate. Gesture 2(1), 1–17 (2002)
3. Bergmann, K., Eyssel, F., Kopp, S.: A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 126–138. Springer, Heidelberg (2012)
4. Bergmann, K., Kopp, S.: Verbal or visual: How information is distributed across speech and gesture in spatial dialog. In: Proceedings of SemDial 2006, pp. 90–97 (2006)
5. Bergmann, K., Kopp, S.: GNetIc – Using Bayesian decision networks for iconic gesture generation. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 76–89. Springer, Heidelberg (2009)

6. Bergmann, K., Kopp, S.: Gestural alignment in natural dialogue. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2013), pp. 1326–1331. Cognitive Science Society, Austin (2012)
7. Bergmann, K., Kopp, S., Eyssel, F.: Individualized gesturing outperforms average gesturing – evaluating gesture production in virtual humans. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 104–117. Springer, Heidelberg (2010)
8. Breslow, L., Harrison, A., Trafton, J.: Linguistic spatial gestures. In: Proceedings of Cognitive Modeling 2010, pp. 13–18 (2010)
9. Cassell, J., Stone, M., Yan, H.: Coordination and context-dependence in the generation of embodied conversation. In: Proceedings of the First International Conference on Natural Language Generation (2000)
10. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The behavior expression animation toolkit. In: Proceedings of SIGGRAPH 2001, New York, NY, pp. 477–486 (2001)
11. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychological Review 82(6), 407–428 (1975)
12. Endrass, B., Damian, I., Huber, P., Rehm, M., André, E.: Generating culture-specific gestures for virtual agent dialogs. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 329–335. Springer, Heidelberg (2010)
13. Hostetter, A., Alibali, M.: Raise your hand if you're spatial—relations between verbal and spatial skills and gesture production. Gesture 7, 73–95 (2007)
14. Hostetter, A., Alibali, M.: Cognitive skills and gesture-speech redundancy. Gesture 11(1), 40–60 (2011)
15. Kita, S., Özyürek, A.: What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language 48, 16–32 (2003)
16. Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., Ishizuka, T.: Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes 22, 1212–1236 (2007)
17. Kita, S., Davies, T.S.: Competing conceptual representations trigger co-speech representational gestures. Language and Cognitive Processes 24(5), 761–775 (2009)
18. Kopp, S., Bergmann, K.: Automatic and strategic alignment of co-verbal gestures in dialogue. In: Wachsmuth, I., de Ruiter, J., Jaecks, P., Kopp, S. (eds.) Alignment in Communication: Towards a New Theory of Communication, ch. 6. John Benjamins, Amsterdam (in press)
19. Kopp, S., Bergmann, K., Kahl, S.: A spreading-activation model of the semantic coordination of speech and gesture. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci 2013). Cognitive Science Society, Austin (in press, 2013)
20. Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K., Cassell, J.: Trading spaces: How humans and humanoids use speech and gesture to give directions. In: Nishida, T. (ed.) Conversational Informatics, pp. 133–160. John Wiley, New York (2007)
21. Kraemer, N., Bente, G.: Personalizing e-learning. The social effects of pedagogical agents. Educational Psycholoy Review 22, 71–87 (2010)
22. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)

23. Levelt, W.J.M.: Speaking: From intention to articulation. MIT Press (1989)
24. McNeill, D., Duncan, S.: Growth points in thinking-for-speaking. In: Language and Gesture, pp. 141–161. Cambridge University Press, Cambridge (2000)
25. Melinger, A., Kita, S.: Conceptualisation load triggers gesture production. Language and Cognitive Processes 22(4), 473–500 (2007)
26. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics 27(1), 1–24 (2008)
27. Özyürek, A.: Speech-gesture relationship across languages and in second language learners: Implications for spatial thinking and speaking. In: Proceedings of the 26th Boston University Conference on Language Development, pp. 500–509 (2002)
28. Sowa, T., Kopp, S.: A cognitive model for the representation and processing of shape-related gestures. In: Proc. European Cognitive Science Conference (2003)
29. Swets, B., Jacovina, M.E., Gerrig, R.J.: Effects of conversational pressures on speech planning. Discourse Processes 50(1), 23–51 (2013)

Modeling Multimodal Behaviors from Speech Prosody

Yu Ding[1], Catherine Pelachaud[1], and Thierry Artières[2]

[1] CNRS-LTCI, Institut Mines-TELECOM, TELECOM ParisTech, Paris, France
{yu.ding,catherine.pelachaud}@telecom-paristech.fr
[2] Université Pierre et Marie Curie (LIP6), Paris, France
thierry.artieres@lip6.fr

Abstract. Head and eyebrow movements are an important communication mean. They are highly synchronized with speech prosody. Endowing virtual agent with synchronized verbal and nonverbal behavior enhances their communicative performance. In this paper, we propose an animation model for the virtual agent based on a statistical model linking speech prosody and facial movement. A fully parameterized Hidden Markov Model is proposed first to capture the tight relationship between speech and facial movement of a human face extracted from a video corpus and then to drive automatically virtual agent's behaviors from speech signals. The correlation between head and eyebrow movements is also taken into account during the building of the model. Subjective and objective evaluations were conducted to validate this model.

Keywords: virtual agent, speech to motion synthesis, head motion synthesis, eyebrow motion synthesis, Hidden Markov model, speech driven.

1 Introduction

Embodied conversational agents, ECAs, are autonomous software characters that often have a human-like appearance and endowed with communicative and expressive capabilities. They are capable of using speech and multimodal behaviors to convey intentions and to express emotions as humans do. The prevalence of embodied conversational agents in interactive systems such as online web applications has been motivated by the development of computational models to generate realistic, natural and believable virtual agents. In video games or cinematographic applications the animation of virtual characters are reproduced from large motion capture datasets of an actor's performance. This approach induces two non-negligible disadvantages: data acquisition expenses and restriction to movement reproduction of the recorded scenarios [1]. On the other hand, the control of the behaviors of autonomous virtual characters is either based on psychology literature [2] or on statistical approaches such as Markovian model [1].

Human-to-human communication is a multimodal process involving speech, facial expression, body gesture and gaze, etc. Humans are sensitive to subtle

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 217–228, 2013.

expressions during face-to-face conversation. For example, they are skilled in inferring their interlocutor's affective and mental states from their accompanied facial expression or body gestures. [3] reported that natural head motion significantly facilitates auditory speech perception. Speech and behaviors production are tightly coupled [4, 5]. For example, [6] found a strong correlation between the raise of pitch contour (F0) and of eyebrow movements.

Our work is focused on investigating a statistical model to infer eyebrow and head motions from speech signals. This model can be parameterized from training samples and can then synthesize natural animation motion from speech features. In our model, head and eyebrow motions are not separately synthesized from speech signals; rather it takes into account the relationships between these two-modal motions.

In the remaining of this paper we first describe related works, we introduce our approach and finally we report on experimental results from objective and subjective evaluations.

2 Related Works

[7–9] proposed rule-based approaches to generate nonverbal communicative features, such as head motion and gesture. However, since human behaviors arise from and may be influenced by various factors, such as emotion, personality, gender, physiological state and social context [10], it is extremely difficult to define a large set of rules to fully capture the role of these factors onto human behaviors, even after decades of studies in psychology. Besides, multiple rules could result in synthesizing conflicting expressions [10].

Recently, data-driven models have been proposed as body or facial motion generators. Few predictive models have been investigated such as Conditional Restricted Boltzmann Machine (CRBM) [11] but the majority of systems are built using statistical models such as Hidden Markov models (HMMs) or more generally state space models. A state space model implements a probability density on observed sequences (e.g. a temporal series of head position). In a state space model the observed sequence is assumed to be produced in two steps: first a state sequence is chosen, second an observation sequence is produced given the state sequence. In both cases the choice is done by drawing a sample according to the corresponding model distribution. Such models are widely used to model speech, handwriting, etc. For instance to model speech phones one uses a state space model with one state for the beginning of the phone, one for the middle part and one for the end of the phone [12]. Within each state observation are produced according to a particular probability density (usually a Gaussian mixture).

A common approach for designing a speech to motion synthesis system relying on such statistical models consists in learning two state space models. There is one model for the speech stream and another model for the motion stream. Both models have the same number of states that are learnt jointly in such a way that these states are paired. Let us illustrate this approach with a simple case where the system consists in learning two HMMs where the first model is trained to model speech observation sequence whereas the second model is trained to model head movement observation sequences. First the speech HMM is learnt on speech observation sequences. Then this model is used to infer the most likely state sequence for every training speech observation sequence. Second the head move HMM model is trained by considering that head move observation sequence have been produced along the state sequences determined by the speech HMM. Doing so one learns which head movements are likely to occur for a given speech signal. The speech/motion mapping is somehow modeled by pairing the states. In presence of speech only, the speech model is first used to infer a state sequence then the head move model is used to syntesize a series of head positions along this state sequence via synthesis techniques as those proposed in [13]. This approach has been implemented in various ways. [14] used two Gaussian Mixture Models; while [15, 16, 1, 17–19] used two HMMs and [20] used a Conditional Random Field (CRF) for the speech and a HMM for motion.

Although this line of works has brought significant results it suffers from the weak interdependency between the input stream (e.g. speech) and the output stream (e.g. head movements) that is actually taken into account through the state space sharing strategy mentioned above.

3 Contextual Markovian Models

We have developed few variants of HMMs to simulate facial animation from speech described in a previous work [21]. In particular we proposed a new model that we name here fully parameterized Hidden Markov model (FPHMM) for learning a mapping function between speech and motion signals, where speech signals can influence directly the synthesized motion signals. We use such models here to simultaneously generate head and eyebrow motions from speech input. We first briefly introduce this model and we show how it can be used to synthesize a motion stream from a speech stream. More details on the models may be found in [21].

A FPHMM is an extension of a contextual HMM (named CHMM hereafter), which has been initially proposed in [22] for modeling and recognizing gestures. The idea behind CHMM is to exploit some contextual information that we know the observation we want to model depends on. Contextual variables may stand for the physiology of a person realizing the gesture, the amplitude of the gesture, the emotional state of a person speaking... [22, 23]. In CHMM contextual variables are used to alter probability density functions in states. More precisely, a CHMM is a HMM whose means and covariance matrices of Gaussian distribution depend on a set of contextual variables, noted by θ, that may vary with time [22, 23]. The contextual variables corresponding to a particular observation sequence are assumed to be known in the training stage as well as in the test stage.

A FPHMM is an extension of a CHMM where in addition to means and covariance matrices, transition probabilities and initial state distribution are also parameterized and depend on θ instead of being fixed at particular values.

In a FPHMM the transition probability a_{ij} from the ith state to the jth state at time t is defined as:

$$a_{i,j}(t) = \frac{e^{W_{ij}^{tr}\theta_t}}{\sum_{j'} e^{W_{ij'}^{tr}\theta_t}} \tag{1}$$

where θ_t is the c-dimensional vector of contextual features at time t and W's are matrices associated to transitions. Using such a modeling framework transition probabilities vary with time according to the values of contextual variables, meaning that a transition may be more likely to occur or not occur at some time according to the contextual information. As any statistical model a FPHMM is trained via likelihood maximization with a Generalized EM algorithm. To ease learning it is initialized with a trained CHMM.

In our work, a FPHMM is used to synthesize the motion stream from the speech stream as follows. We first learn a FPHMM that takes speech features as contextual variables and that produces motion features observation. During the training phase, motion and speech streams are both used to learn a FPHMM. During the motion synthesis phase (i.e. the animation generation phase), only the speech stream θ is known. It is used to compute the time dependent transition probabilities and the time dependent altered emission probability distributions in states (cf. e.g. equation (1)). Once all the parameters of the model are set, one can compute the most likely state sequence, or to get even a more accurate result one can infer the probability distribution over all state sequences. Finally, from this single state sequence or from the distribution over state sequences, one can synthesize a trajectory using techniques such as in [13].

4 Experiments

In this section, we present the corpus used in our experiments and how we extracted facial and prosodic features. Then we describe the conducted objective and subjective evaluations. At last, we report the results and discuss them.

4.1 Datasets

Experiments were performed on the Biwi 3D AudioVisual Corpus of Affective Communication database [24]. We used 240 sequences from 3 subjects extracted from this corpus. In this corpus, each subject tells 80 short English sentences. Each sentence lasts 4.67s long on average. 3D face geometries were captured at 25Hz, which comprise of a total of 23370 facial points including 3 head rotations.

From such a large set of facial points, we extracted a subset of facial motion features that correspond to 3 head rotations and 8 eyebrow features (see Figure 1). These features coincide with the Facial Animation Parameters as defined by the norm MPEG-4 [25]. For sake of simplicity, we assume both eyebrows move identically and we take the mean of the right and the left eyebrows as the eyebrow

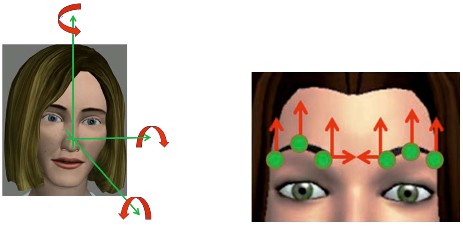


Fig. 1. Facial motion features - Left: 3 head rotations. Right: Eyebrow animation parameters (arrows illustrate displacements).

motion features. At the end a motion signal is transformed in a sequence of 7-dimensional (3-dimensional head and 4-dimensional eyebrow) feature vectors at a rate of 25 frames (i.e. feature vectors) per second (fps).

Concerning the speech features we consider 2 prosodic features (pitch and RMS energy), which were extracted with PRAAT software [26] at the same sample rate as for motion feature extraction (25 fps).

We call static features, the 7-dimentional motion features and the 2 prosodic ones. On top of these features, we include also the first and second order derivatives of static features (i.e. velocity and acceleration of the dynamic features).

4.2 Objective Evaluation

We first performed experiments for modeling head and eyebrow features separately with 2 PFHMMs, one for each motion stream. Then experiments were performed to jointly model head and eyebrow features with a single PFHMM, where the joint features of head and eyebrow were considered as FHMM's observation features. We compare our results with the baseline approach proposed in [18] that we have implemented and that we have tuned for the task at hand.

These two sets of experiments were evaluated by computing the reconstruction error defined as the mean square error between the synthesized motion signal (from the speech signal) and the real motion signal (MSE criterion). Table 1 reports the experimental performances with different numbers of states based on the averaged results and the standard deviation over 50 random splits of the dataset into 80% for training and 20% for testing. The experiments are configured using full covariance matrix and ergodic topology, which fully guarantee the model specified by samples of data training. As can be seen in Table 1, the

Table 1. Performance of the models with respect to the synthesis quality (MSE). Performances are averaged results gained on 50 experiments (standard deviations are given in brackets).

Model	10 states	20 states	30 states	50 states
[18]	0.57 (0.054)	0.53 (0.049)	0.49 (0.061)	0.46 (0.053)
separate PFHMM	0.45 (0.069)	0.41 (0.066)	0.39 (0.059)	0.38 (0.051)
joint PFHMM	0.39 (0.071)	0.34 (0.059)	0.30 (0.045)	0.30 (0.036)

joint model outperforms the baseline [18] as well as the other two using separate models.

In these experiments, a combination of covariance matrices were exploited for the observation function in FPHMM. In the joint model, the relationship between eyebrow and head features can be learned using the full covariance matrices. In synthesis phase (trajectory computation), generated eyebrow and head motions are defined not only from the speech features but also from their mutual influence. This influence is captured during the training phase through the combination of covariance matrices. On the other hand, when using two separate models, there is an underlying hypothesis implying that head and eyebrow movements are independent from each other that is carried out during the training phase and that produces poorer results.

As can be seen from Table 1, the joint model with 30 states achieves the best result. We use this model to compute the animation of the virtual agent. In the next section we present the subjective evaluation study we conducted where we looked at qualitative measures. It is a necessary complement of the objective measure we just presented. Indeed objective measure does not allow us to measure how motions are perceived by human eyes [27].

4.3 Subjective Evaluation

In this section, we detail the subjective evaluation we conducted with human participants to evaluate the qualitative aspects of FPHMM as generator of head and eyebrow motion from speech signals. The evaluation was done through an online web application.

Hypothesis. The subjective evaluation was conducted to investigate two hypotheses: 1) the perception of the virtual agent displaying head and eyebrow motions synthesized by FPHMM is similar to the perception of the virtual agent animated directly by human data; 2) the two-modal motions (head and eyebrow) outperform single model motion (either head or eyebrow) at a perceptual level. Through the first hypothesis we aim to measure if FPHMM is capable of capturing and of rendering the sophisticated relationship between motion and speech streams and that the animation of a virtual character offers similar results when driven by FPHMM and by real human data. The second one is to verify that multimodal motions facilitate human perception over monomodal motions. To verify the first hypothesis, we compare the perceptions resulted from real and generated motions. To answer the second hypothesis, we compare how animations of the virtual agent driven either by one of the modal motions (either head or eyebrow motions) or by two modal motions (head and eyebrow motions) are perceived.

Our work focuses on nonverbal communication and not on the appearance of the virtual agent. Therefore, motions from both, FPHMM types and human data, are displayed through the identical virtual agent.

Protocol. The participants went on a web page where after answering few questions about themselves, they have to view videos of the virtual agents. Their task consisted in answering few questions. We provide elements of the protocol we follow for the perceptive evaluation study.

(1) Participants: in total, there were 280 participants consisting of 136 males and 144 females with age ranging from 18 to 65 (M=32.89 years, SD=7.99 years).

(2) Stimuli: 7 spoken utterances were randomly selected from the testing database. They were given as input to the trained FPHMM. Then the synthesized motions (sequences of MPEG-4 FAPs frames) and the corresponding WAV file of the spoken utterances are used to drive a virtual agent. In all the animations the lip shapes and body movements of the virtual agents are reproduced from real data. The final animations including eyebrow and head motion as well as lip shape and body movements are stored as video clips.

To study both hypotheses, 7 versions (conditions) of the virtual agent animations were created for each selected sentence. In all conditions the lip shapes and body movements remain constant and are duplicated from real human data:

1st condition (cond1): No eyebrow and head motion;
2nd condition (cond2): Only human eyebrow motion (no head motion);
3rd condition (cond3): Only synthesized eyebrow motion (no head motion);
4th condition (cond4): Only human head motion (no eyebrow motion);
5th condition (cond5): Only synthesized head motion (no eyebrow motion);
6th condition (cond6): Human eyebrow and head motion;
7th condition (cond7): Synthesized eyebrow and head motion;

Therefore, there are a total of 49 video clips (7 sentences $\times$ 7 conditions), each of which lasts about 4.5s.

(3) Design and Procedure: Subjective evaluations were conducted online. At first, each participant fills out a demographic questionnaire concerning their age, gender, education level, occupation and country in which participant spent the majority of his/her life. Then, the participant watches 7 randomly selected video clips out of 49. The 7 video clips watched by any participant are comprised of the 7 sentences and of the 7 conditions. After watching each video clip, each participant is invited to answer the following questions using a 5 point Likert scale:

1. Do you think the animation of the virtual character is intelligible?
2. Do you think the animation of the virtual character is natural?
3. Do you think the correlation between the speech and the facial expression of the virtual character is coherent?
4. Do you think the correlation between the speech and the facial expression of the virtual character is synchronized?
5. Which emotion(s) does the virtual character display? You should grade each emotion separately.

The same 12 emotional states as used in the BIWI experiments are considered [24]: anger, sadness, fear contempt, nervousness, disgust, frustration, stress, excitement, confidence, surprise and happiness. Each video clip has been evaluated 40 times (i.e., by 40 participants).

Table 2. Results of F-statistic from repeated measures ANOVA

	intelligible	natural	coherent	synchronized
only eyebrow	F=20.6,p<.001	F=1.79, p>.05	F=2.02, p>.05	F=2.57, p>.05
only head	F=1.7, p>.05	F=0.39, p>.05	F=0.02, p>.05	F=2.57, p>.05
eyebrow&head	F=3.51, p>.05	F=0.93, p>.05	F=1.9, p>.05	F=0.01, p>.05

Result. To investigate the differences of participants perception from human and from synthesized motions, we conducted three pairwise comparisons in term of intelligibility, naturalness, coherency and synchronization: only eyebrow motion (2nd and 3rd conditions), only head motion (4th and 5th conditions) and both (6th and 7th conditions). The comparison results based on repeated measures ANOVA are shown in Table 2. The results show no significant differences between human and generated motions in almost all pairwise conditions except for the only eyebrow condition in term of intelligibility. In this latter case, the mean scores of this pairwise condition are 2.67 and 2.27 for only human and synthesized motions, respectively. While the difference is significative, they are still not too highly different.

Then, to test our second hypothesis (the two-modal motions outperforms single model motion), we compare the 4 different conditions involving synthesized motions with each other, namely: no head and eyebrow motions (1st condition), only synthesized eyebrow motion (3rd condition), only synthesized head motion (5th condition) and both synthesized (head and eyebrow) motions (7th condition). The comparison results presented in Figure 2 show that when eyebrow and head motions are modeled together, the human perception improves in the term of intelligibility, naturalness, coherency and synchronization. The results based on repeated measures ANOVA show significant differences between any two different types of motions among the 4 cases (synthesized eyebrow or head, both synthesized eyebrow and head, no motion of eyebrow and head). The same conclusions are supported by the similar pairwise for the animations from human data (see details in Figure 2).

At last, we investigated how synthesized motion conveyed emotional information. To do this, we extracted the scores in term of 12 emotions from both synthesized head and eyebrow motions (7th condition) and from both human motions (6th condition), respectively. Moreover, the BIWI corpus provides the emotion recognition rate for the videos of real humans for these 12 emotions using a 5 point Likert scale. We consider these results as reference when evaluating the virtual agent's performance. Figure 3 reports the recognition rate of participants for the virtual agent when driven from human data and from synthesized model, as well as for the videos of real humans. The results are average over the recognition rate for 5 of the 7 sentences spoken by the virtual agent. Two of the sentences have to be disregarded for this test as no result is provided for them in the case of the evaluation of the real human videos. The number of participants differ in the case of the study made with videos of real humans as with videos made with virtual agents. In case of the BIWI study, the first set of videos was evaluated when the BIWI corpus was built. There is a very low number of participants that eval-

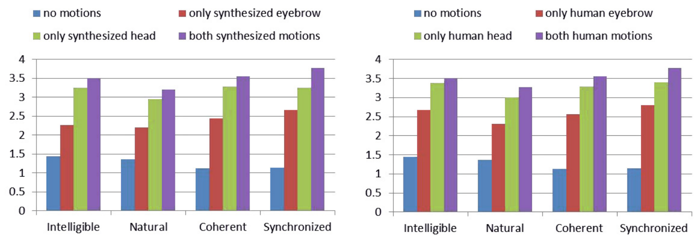


Fig. 2. Comparison results between animations from synthesized data (left) and human data (right)

uated each video (often around 4 participants per sentence) while in our study we have 40 participants evaluating all the sentences in average. As can be seen in Figure 3, in all three cases, the faces, be of a real human or of a virtual agent, are able to convey some emotional communication. However, in term of the perception of emotional expressiveness, the videos of the real humans outperform the videos of the virtual character driven from our statistical model; in turn, the videos of the virtual character driven from our statistical model outperform the videos of the virtual character driven from human data.

Due to the too big difference between participants number, we only compare results between the conditions of the virtual agent driven from our model and from human data. The animations with synthesized motions are perceived as showing more emotions in a statistically significant way for the emotion: fear, contempt, nervousness, stress, excitement, surprise, happiness. There is no significant difference for the emotions: anger, sadness, frustration. For the emotions disgust and confidence, the animations from human data are ranked higher than the animations from synthesized data.

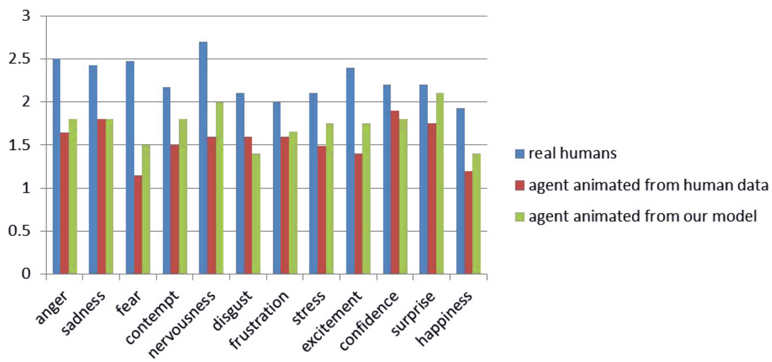


Fig. 3. Perceived emotions: average values over 5 sentences

Discussion. The post-hoc pairwise comparisons between identical modal motions show no significant difference between the human and synthesized motions. In the training phase, the mapping between human audio-visual signals is captured and recorded in FPHMM, and thus rendered in the output synthesized animations. Therefore the perception of the animation of the virtual agent with synthesized motions is similar to the perception of animation with human motions.

The post-hoc pairwise comparisons of conditions show that there are significant differences among the monomodal and multimodal conditions. The animations created in the multimodal conditions are perceived as more intelligible, natural, coherent and synchronized than the animations created with one modality only. Humans are very skilled in reading nonverbal signals. So the lack of either eyebrow or head motions are negatively perceived along these 4 qualitative dimensions. The comparisons of the perceptions between only eyebrow and only head motions reveals that head motion plays a more important role than eyebrow motion at the perception level. Similar results have also been reported in [18].

Rather than examining the recognition rate of emotions from the videos in different cases, we look at the level of emotional expressiveness. Indeed even in the reference case, namely the videos of real humans speaking the various utterances in emotionally-colored fashion, we remark that the recognition rate level of emotion is rather low and that there are a lot of confusion. That is human participants did not show a strong agreement in their perception of which emotion the human actor aimed to convey. Such a result is reproduced with the virtual agent. In both cases, animation of the virtual character from real human data and from our model, a lot of confusion can be noticed. However we can remark that the virtual agent driven by our model is perceived as speaking with emotions with a higher level than when the virtual agent is driven by human data. This can be interpreted as the virtual agent driven by our model is able to exhibit more expressive behaviors; that is, it can communicate in a more emotionally colored manner. Our statistical model relying on FPHMM is capable of capturing the speech/motion relationship. It is also able to render the quality of emotional behaviors: not only its types of movements (i.e. which head movements and eyebrow shapes) but also the dynamism of the movements. Our model computes the types of the visual cues but also their trajectory that carries out dynamics characteristics.

The results of both evaluation studies, the objective and subjective studies, show that our model is able to capture pertinent information that are conveyed through the nonverbal behavior animation of the virtual agent. Considering the link between prosody and both head and eyebrow motions together allows seizing their tight coupling. This is in link with results found in studies from psychology domain [28, 5].

We can conclude that the machine learning approach (FPHMM) can capture the link between speech prosody and facial movements and can reproduce the movement dynamism in the output synthesized animations. Thus our animation model based on FPHMM is able to gather these both aspects that are important to compute when animating a virtual agent.

5 Conclusion

In this paper, we have presented a data-driven approach to generate head and eyebrow motions for a virtual agent from speech prosody. The full parameterized HMM is used to capture the direct mapping between audio and visual information. The trained PFHMM allows defining visual animation as a function of the speech signal. The objective evaluation study shows that considering that simultaneously eyebrow and head motions increases the precision of the resulting animation. It also confirms that eyebrow and head motions are not independent from each other but rather are connected; the multimodal signals reinforce the communicative meaning. On the other hand, the subjective evaluation shows that our proposed model enhances the perception of the virtual agent animation at the level of emotional expressiveness.

References

1. Busso, C., Deng, Z., Neumann, U., Narayanan, S.: Natural head motion synthesis driven by acoustic prosodic features. Journal of Visualization and Computer Animation 16(3-4), 283–290 (2005)
2. Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., Pelachaud, C.: GRETA: Towards an Interactive Conversational Virtual Companion. In: Artificial Companions in Society: Perspectives on the Present and Future, pp. 1–17 (2010)
3. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Bateson, E.V.: Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological Science 15(2), 133–137 (2004)
4. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press (2004)
5. Ekman, P.: About brows: Emotional and conversational signals. In: von Cranach, M., Foppa, K., Lepenies, W., Ploog, D. (eds.) Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium, pp. 169–248. Cambridge University Press, Cambridge (1979)
6. Bolinger, D.: Intonation and Its Uses: Melody in Grammar and Discourse. University Press (1989)
7. Pelachaud, C., Badler, N.I., Steedman, M.: Generating facial expressions for speech. Cognitive Science 20, 1–46 (1996)
8. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Bechet, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: Ruled-based generation of facial expression gesture and spoken intonation for multiple conversational agents. In: Computer Graphics, pp. 413–420 (1994)
9. Beskow, J.: Rule-based visual speech synthesis. In: 4th European Conference on Speech Communication and Technology ESCA-EUROSPEECH 1995, Madrid (September 1995)
10. Lee, J., Marsella, S.: Modeling speaker behavior: A comparison of two approaches. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 161–174. Springer, Heidelberg (2012)

11. Chiu, C.-C., Marsella, S.: How to train your avatar: A data driven approach to gesture generation. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 127–140. Springer, Heidelberg (2011)
12. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
13. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: ICASSP, pp. 1315–1318 (2000)
14. Costa, M., Chen, T., Lavagetto, F.: Visual prosody analysis for realistic motion synthesis of 3d head models. In: Proc. of ICAV3D, pp. 343–346 (2001)
15. Dziemianko, M., Hofer, G., Shimodaira, H.: Hmm-based automatic eye-blink synthesis from speech. In: INTERSPEECH, pp. 1799–1802 (2009)
16. Hofer, G., Shimodaira, H., Yamagishi, J.: Speech driven head motion synthesis based on a trajectory model. In: ACM SIGGRAPH 2007 Posters (2007)
17. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Trans. on Audio, Speech & Language Processing 15(3), 1075–1086 (2007)
18. Mariooryad, S., Busso, C.: Generating human-like behaviors using joint, speech-driven models for conversational agents. IEEE Trans. on Audio, Speech & Language Processing 20(8), 2329–2340 (2012)
19. Xue, J., Borgstrom, J., Jiang, J., Bernstein, L., Alwan, A.: Acoustically-driven talking face synthesis using dynamic bayesian networks. In: 2006 IEEE International Conference on Multimedia and Expo, pp. 1165–1168 (2006)
20. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. ACM Trans. Graph. 29(4) (2010)
21. Ding, Y., Radenen, M., Artières, T., Pelachaud, C.: Speech-driven eyebrow motion synthesis with contextual markovian models. In: ICASSP, pp. 3756–3760 (2013)
22. Wilson, A.D., Bobick, A.F.: Parametric hidden markov models for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 884–900 (1999)
23. Radenen, M., Artières, T.: Contextual hidden markov models. In: ICASSP, pp. 2113–2116 (2012)
24. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-D Audio-Visual Corpus of Affective Communication. IEEE Transactions on Multimedia 12(6), 591–598 (2010)
25. Pandzic, I., Forcheimer, R.: MPEG4 Facial Animation - The standard, implementations and applications. John Wiley & Sons (2002)
26. Boersma, P., Weeninck, D.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
27. Lee, J., Marsella, S.: Predicting speaker head nods and the effects of affective information. IEEE Transactions on Multimedia 12(6), 552–562 (2010)
28. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)

An Examination of Whether People Prefer Agents Whose Gestures Mimic Their Own

Pengcheng Luo[1], Victor Ng-Thow-Hing[2], and Michael Neff[1]

[1] UC Davis, USA
{pcluo,neff}@ucdavis.edu
[2] Honda Research Institute USA, Inc.
vng@honda-ri.com

Abstract. Do people prefer gestures that are similar to their own? There is evidence that in conversation, people will tend to adopt the postures, gestures and mannerisms of their interaction partners [1]. This mirroring, sometimes called the "chameleon effect", is associated with affiliation, rapport and liking. It may be that a useful way to build rapport in human-agent/robot interaction is to have the agent/robot perform gestures similar to the human. As a step towards that, this study explores if people prefer gestures similar to their own over gestures similar to those of other people. Participants were asked to evaluate a series of agent motions, some of which mimic their own gestures, and rate their preference. A second study first showed participants videos of their own gesturing to see if self-awareness would impact their preference. Different scenarios for soliciting gesture behavior were also explored. Evidence suggests people do have some preference for motions similar to their own, but self-awareness has no effect.

Keywords: Human Agent Interaction, Gesture Mimicry.

1 Introduction

With the emergence of more intelligent agents and robots, there is a growing need for improved human-agent/robot interaction. Along with speech, non-verbal behavior plays an important role in communication and interaction. Numerous researchers have been working on developing gestures for agents [1–3] and robots [4, 5]. However, how an agent/robot can most effectively use gestures to interact with people is not well studied. Mimicry is one important phenomenon that is found in human-human interaction. Called "chameleon effect" in some literature, previous researchers have found that mimicry can increase liking among interaction partners [6, 7]. This phenomenon suggests that people may like gestures more similar to their own, and by extrapolation, since humanoid agents/robots share a similar morphology to people, having the agent copy the user's gestures may increase the user's comfort. We conducted a set of experiments to see whether people prefer virtual agents that perform gestures similar to their own. Before the second round of experiments, participants were shown recorded video segments of their movements to see if the increased self-awareness this generated affected their ratings. These studies provide evidence for the efficacy of potential future systems that perform mimicry in real time.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 229–238, 2013.

Our contributions lie in several parts: first, we conducted a pilot study on people's preference for agents that gesture like them. The experiments showed some evidence supporting that 1) People appear to show preference for agents that mimic their own gestures, although it may not be their favorite one. A post-experiment analysis raised the possibility that whether people prefer agents that mimic their own motions might be related to people's personality, but this requires further investigation. 2) Self-awareness does not affect people's preference. Secondly, several methods for soliciting gesture behavior have been tested and compared. Analysis shows the type of gestures people perform is affected by content of the prompt they are given.

The structure of the paper is as follows. Section 2 provides relevant literature. Section 3 describes an overview of the experiment design. In Sections 4 and 5, we will discuss the experiment preparation and results. We will conclude with our results a discussion of future directions.

2 Background

Nonverbal behavior of embodied agents and robots, especially gestures, has a strong impact on communication. Krämer et al. [8] found that self-touching gestures have positive effects on user' evaluations. The agent was rated as more natural, warmhearted, agile and committed when presenting self-touching gestures. Salem et al. [5] found that a robot is evaluated more positively when non-verbal behaviors such as hand and arm gestures accompany speech. Neff et al. [9, 10] conducted experiments to understand how the Big Five traits of emotional stability and extraversion correlate with changes in verbal and nonverbal behavior. The perception of these traits is varied by adjusting a virtual agent's gesture rate, whether self-adaptors are present and movement style parameters.

The "chameleon effect" impacts interpersonal interactions. Lakin and Chartrand [7] showed that mimicry can be used to build liking among interlocutors. Chartrand and Bargh [6] found that mimicry facilitates the smoothness of interactions and increases liking between interaction partners. Lee [11] found mimicry to be predictive of liking between participants instead of trust. Several behaviors have been observed to be mimicked in previous literature. Chartrand and Bargh [6] found rubbing one's face or shaking one's foot is mimicked. Cappella and Planalp [12] found that people have a tendency to mimic the way they speak, matching features such as rhythm and pauses. Gestural mimicry [13–15], the recurrence of gestural features across speakers, indicates one speaker's gesture is influencing the others. This inspired our work on examining gestural mimicry as a mechanism for human agent interaction.

For human-agent/robot interaction, mimicry has been used before in several places. The effect of agent mimicry, however, is still an open question. Bailenson and Yee [16] studied the effect of head mimicry using an embodied artificial agent. Their results showed that mimicking agents were more persuasive and received more positive trait ratings than non-mimickers, despite participants inability to explicitly detect the mimicry. Gratch and his colleagues [17] designed a listening agent that would try to create rapport by tying listening feedback to shallow features of a speaker's voice and bodily movements. In contrast to that research, we are looking at communicative gestures for a speaking agent. Kopp [18] designed a framework for human agent interaction by transferring several coordination mechanisms, such as mimicry, alignment, and synchrony, from face-face interaction to human agent interaction. The primary focus of this paper is to study people's attitude towards agents that gesture like them.

3 Overview

The objective of this research is to study people's attitude towards an agent that copies their own gestures. With this objective in mind, we have two hypotheses:

- Hypothesis1: People prefer gestures that are similar to their own.
- Hypothesis2: Self awareness affects peoples preference for agent motions.

While a real time mimicry system is under development, we implemented "gesture mimicry" by copying gestures performed by subjects during test sessions and applying these to a virtual agent afterwards. To achieve this, we first collected subjects' motion while they discussed different topics. This yielded a gesture database including gesture and key word information for each person. To generate new motion for a virtual agent, we first constructed new text with key words that appeared in the previous topics. Based on the mapping between gestures and key words, we selected gesture motion that is associated with key words from the database. More details can be found in section 4.1.

A pre-test was performed to confirm that the agent gestures accurately replicated the gestures of each subject. An agent video was made based on each subject's data. Side-by-side videos were then made of every combination of agent motion with videos of the original subject motion. These were posted to Amazon Mechanical Turk and subjects were asked to rate the similarity between the original videos and the agent motion. This ensured that the generated agent motion was able to demonstrate mimicry by confirming that the copied motions were the best match for the originals (Section 4.3). The newly generated virtual agents were then shown to the original subjects to see if they prefer the virtual agent that has similar motion to their own.

4 Experiment Preprocess

4.1 Elicitation of Motion Examples

Our first task was to build a motion database that will allow us to generate novel virtual agent utterances that make appropriate use of the captured gestures. This requires satisfying several objectives: 1) we need to collect enough gesture samples for each person, 2) participants should feel comfortable so they perform natural gestures, 3) for the easiness of synthesizing new motions based on same text, we need solicit motions from different people using similar key words. It remains unclear how to best solicit such natural gestures, so we experimented with several approaches. We suspect that speech content is an important factor that affects gesture quality, therefore we designed 5 tasks based on different content constraints.

1. FiC: Fixed Content with emphasis specified for several key words. Participants are told to speak a short piece of text with emphasis on specified key words.
2. FrC: Free Content. We let participants tell a story freely related to "catching a fish".
3. CV: Constrained by Video. We showed a piece of short video in which the robot Asimo is pouring a drink into a cup and let people describe what happened in the video.
4. CA: Constrained by Audio. We first broadcast a short audio clip about the tale of the"three little pigs" and then ask people to rephrase the story.
5. CI: Constrained by Image. We show people one image and let them describe the scenario in the image.

Eight regular university students participated in this long term research project. Each students was paid $25 for participating in the whole project. During the capture session, we used Kinect to capture 3-d motion and a video camera for image and voice. The motion data was then segmented into gesture sequences using Anvil software [19]. Segmented Kinect data were stored in a database for future use.

We looked at several motion features in the gesture sequences as shown in Table 1. The data shows that task CV (constrained by video) might be the most efficient and stable way to produce motion over different subjects as it has highest average MPD (duration of movement/duration of complete video sequence) and lowest std of MPD. FiC would generate a consistent number of gestures as it has lowest std in terms of number of gestures. CV has second lowest variance in terms of NOG.

Table 1. Motion Statistics. NOG: number of gestures; MPD: duration of movement/duration of complete video sequence; MD: sequence duration(seconds).

experiment type	mean of NOG	std of NOG	mean of MPD	std of MPD	mean of MD
FiC	8.75	1.49	0.43	0.15	15.8
FrC	14.5	11.65	0.75	0.16	51.9
CV	10.5	5.15	0.77	0.10	39.6
CA	34.75	11.55	0.49	0.13	160.4
CI	14.375	9.62	0.66	0.21	53.9

In [20], McNeill found a phenomenon - "the saliency of dimension" - when people describe cartoons in different contexts. Iconic gestures dominate in narrative contexts (promoting the development of the story) while metaphoric gestures predominate in extra-narrative clauses (description of the setting and characters, summaries, etc.). We examined the gesture types performed with in different tasks (Table 2) and found that iconic gestures dominate in CV. The reason might be that people tend to copy the robot's action while describing the video. If the text is fixed, people tend to use more beats. In a post-experiment survey, 3 of 8 participants mention that task FiC is difficult as they need to read the text and act it out. This could partially explain why there are more beats in this task. No one said that task CV was difficult. 6 of 8 participants agree that task CI, describing a picture, is the easiest one. Therefore we see a relatively balanced number of different gesture types, except emblems which are consistently rare in our motions. When the task is easy, participants might be more expressive using different kinds of gestures.

Table 2. Motion Dimension

	Emblems	Iconic	Metaphoric	Deictic	Beat	Total Number
FiC	0.00%	15.87%	15.87%	19.05%	**49.21%**	63
FrC	4.07%	48.78%	13.01%	4.07%	30.08%	123
CV	2.25%	**69.66%**	6.74%	10.11%	11.24%	89
CA	2.92%	22.99%	19.71%	8.39%	45.99%	274
CI	2.56%	38.46%	21.37%	16.24%	21.37%	117

4.2 New Motion Generation

It is necessary to have a common script for the virtual agent in order to generate our stimuli for comparison, but people made different speeches during the capture session. To develop a script, we first picked several typical words from FiC and CI based on two considerations: 1) the easiness of finding similar words across different people, 2) the easiness of detecting gesture shape differences given similar key words. We then created two new pieces of text for generating new motions using these key words. Voice was generated using NaturalReader [21], a text to speech software. To limit gender bias, we displayed the motion on a virtual wooden model.

The two new texts used in our generation tests are listed as follows (italicized words are key words that received gestures):

1. "We will *cut* taxes and *raise* the standard of living in this country. For *all* of you hurting out there, I *feel* your pain. "
2. "In the *picture*, there is a *big* house in the upper corner with thousands of *balloons*. A *rope* is coming out of it."

The captured motions in the database were segmented and labeled with associated key words. They were retargeted to the virtual character before motion reuse. Post-processing tools were applied to make sure the motions look similar to the original motion and this similarity was ensured through the pre-test discussed in Section 4.2. To generate personalized, new motions, we align motion segments in a personalized database with their associated key words in the text. During alignment, we make sure the end of the stroke is aligned with the end of key word audio. Motion segments are connected to create continuous sequences. In total we have produced 8 personalized agent motion clips for each utterance.

4.3 Similarity Evaluation

We want to ensure that there is clear "mimicry" between a subject's original performance and the sequence designed to match it. It is not an easy task to quantify the similarity between a pair of motion sequences, thus we employ a perceptual study on Amazon Mechanical Turk to evaluate the similarity between the original video and new motion. We showed participants motions side by side: left, original video (OV) and right, the generated motion (GM)(Fig. 1). After viewing the motion, they are asked to provide a rating for each of the following 5 prompts on a 5-point Likert-scale "Strongly disagree/Disagree/Neutral/Agree/Strongly agree" with values from 2 to -2.

- Character in video B seems to have the same personality as the person in video A.
- Character in video B appears to have the same natural rhythms (timing) as the person in video A.
- Character in video B is trying to copy the person in video A.
- Character in video B is different in attitude from the person in video A.
- Character in video B has a similar expressiveness to video A.

The average of the five questions (note the rating of the fourth item is reversed) will be used to indicate the similarity level between the videos and generated motion sequence.

Let (OV_i, GM_i) define a pair of motion sequences. The new motion GM_i is generated from original motion captured along with OV_i for subject i. Since there are 8 motion pairs, every participant needs to compare OV_i with $\{GM_k | k = 1, 2, \cdots 8\}$. For OV, we only showed video clips that are associated with the key words used in GM. In order to minimize potentially contaminating effects, we blacked out the face and turned off the voice. The motion clips are connected in a cross-dissolve format as we want to let people focus on the gestures while avoiding the abruptness of sudden cuts. To verify the similarity of 16 generated agent videos, we designed 16 perceptual experiment with 8 pairs of video in each. 10-15 participants were collected for each experiment and each was paid

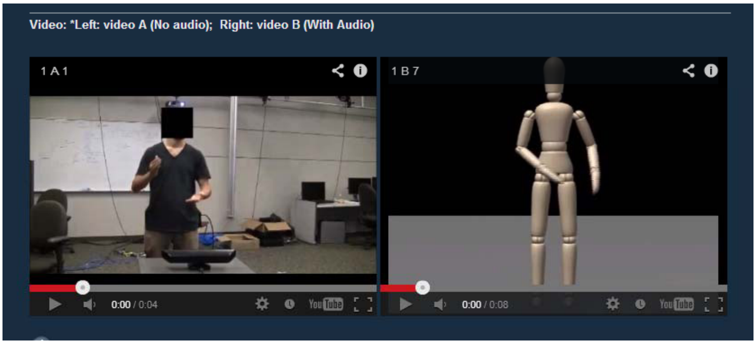


Fig. 1. The snapshot of similarity evaluation experiment

\$0.5/test. In order to collect high quality surveys, we only recruited "master workers"(qualified participants on Amazon Mechanical Turk) and reviewed the results before approving the payment. The results showed that the average similarity rating of the pair (OV_i, GM_i) is always the highest compared with other pairs (OV_i, GM_j) where $j \neq i$. This means that we have correctly produced new motions that have noticeable mimicry. The average ratings for each clip are shown in the following Table 3.

Table 3. Average similarity rating based on 5 questions above (first utterance)

original vs generated motion	1	2	3	4	5	6	7	8
1	**0.36**	0.00	-0.18	-0.45	-0.45	-0.27	0.00	-0.18
2	-0.43	**0.29**	-0.79	-0.50	-0.79	-0.50	-0.14	0.07
3	0.00	-0.75	**0.33**	-0.58	-0.50	-0.58	-0.08	-0.08
4	-0.43	-0.57	-0.79	**0.14**	-0.71	-0.43	-0.57	0.07
5	-0.22	-0.44	-0.56	-0.78	**0.78**	0.11	-0.67	-0.22
6	-0.36	-0.45	-0.55	-0.55	0.27	**0.45**	-0.45	-0.18
7	-0.17	-0.25	-0.08	-0.67	-0.58	-0.42	**0.42**	-0.25
8	-1.00	-0.40	-0.80	-0.70	-0.30	-0.20	-0.50	**0.60**

5 Experiment and Result Analysis

5.1 Overview

We designed two experiments to test our two hypotheses: Hypothesis1: people prefer the gestures that are similar to their own; Hypothesis2: self awareness affects peoples preference for agent motions.

The first experiment tests whether people prefer the motion that is similar to their own. To achieve this goal, we show two groups of generated agent motions to participants whose motions were recorded before. Each group has 8 videos for one of the utterances which were played in random order (only one animation is based on the participant's data). In order to achieve a more accurate rating on affinity, we designed questions along seven different character traits (persuasive/likable/charismatic/excited/competent/trustworthy/friendly) with ratings form 1 to 9 (least to most) in the format "How persuasive does this character seem?". A post analysis shows that these characteristics are highly correlated (the Pearson correlation coefficients range from 0.69 to 1.0). Therefore we used the average rating as an indicator of their affinity rate.

The second experiment examines whether self-awareness will affect people's preference for self-similar motion. To evaluate this, after showing the first set of motions, we showed participants own source videos in the format used in section 4.3. After that, we replayed the generated motions, again in random order, to see if there were any differences in ratings.

The original participants worked on this experiment 3 months after we recorded their motions. We also recorded the time duration that participants spent on each page, with which we can detect invalid data (if the time was too short, it is unlikely that participants were honestly evaluating the clips). 8 students participated in our research and 6 of them had valid data.

5.2 Experiment 1: Preference for Self-similar Motion

To test whether people prefer the motion that is similar to their own, we first rank the ratings of all the agent motions that they watched. If the rank of the agent motion that mimics their own motion ranked 1st, then it indicates that people prefer mimicry motions.

As we only have limited data and the distribution of ranks are not clear, we used a nonparametric test, the sign test [22]. The p-value of the test is 0.92 (n=24). This indicates that the median of the targeted clip rank is greater than 1 and the agent motion that has mimicry is not their favorite one. However if we relax the hypothesis to compare median rank to 4 (as we only have 8 agent motions), p-value < 0.05. The result indicates that the agent motion that has mimicry is on their preferable motion lists.

We also compare the rating of GM_i and the average rating of GM_j where $j \neq i$ without self-awareness. The results show that subject 2 and subject 5 particularly do not like their own motion, but other subjects show preference to their own motion compared with other averaged ones (Fig. 2). A follow up big 5 personality test [23, 24] shows that these two subjects are less extroverted as their test scores are in percentile 24% and 8% in a data groups with 711 users examined in [25]. This might suggest that whether people prefer mimicry is related to individual personality which supports von der Pütten et al.'s findings [26]: participants' personality traits influence their subjective feeling and evaluation of virtual characters.

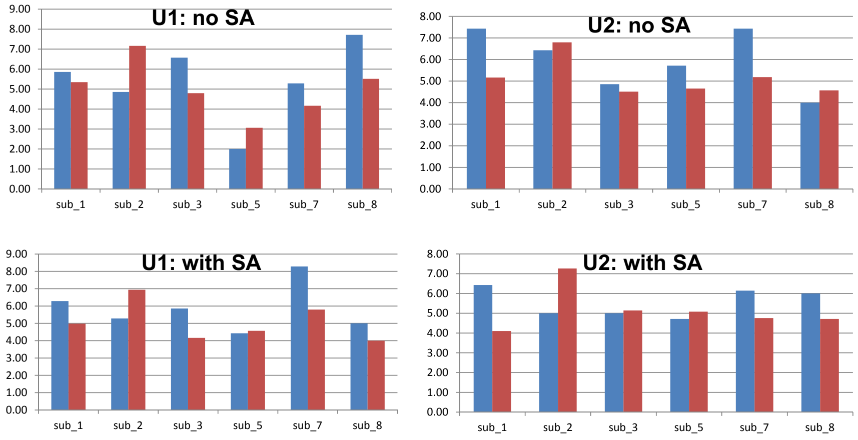


Fig. 2. Horizontal axis: subject i; Vertical axis: rating; Blue: rating of GM_i; Red: average rating of other motions; SA: self-awareness; U1: first utterance; U2: second utterance

5.3 Experiment 2, The Factor of Self-awareness

The next experiment examined the factor self-awareness. Due to the small data size, we used signed rank test [22] to see if there is any difference on rating of GM_i before and after showing subject i his own videos. The signed rank test shows that the self-awareness effect is not significant (p=0.46). We also used the t-test to look at the real rating of the targeted video before and after watching their own videos. t - test shows that there is no significant difference ($p = 0.88$).

6 Conclusion and Discussion

This paper describes a pilot study on whether people prefer gestures that look like their own, providing a potential way of improving human-agent interaction. Through experiments we found that 1) people appear to show a preference for gestures that copy their own but they are not their favorite choice, 2) self-awareness does not affect people's preference for self-similar gestures.

We found several interesting phenomena by analyzing the data which shed light on interesting future directions. 1) In our study, two subjects did not prefer mimicked gestures, but four did. This suggests that whether people prefer mimicry may be quite personality/person dependent. 2) Gesture types seem to be related to content. Task CV seems to have a high chance to solicit iconic gesture as participants tend to copy actions in the video; task CI is able to solicit gestures of balanced distribution in different categories. Guidelines on selecting appropriate content for gesture productions are as follows: if iconic gestures are the highest priority, describing a video with actions might be a good task; asking people to act out a piece of text would probably induce more beat gestures; describing an image would be likely to generate gesture types of balanced distribution.

Acknowledgments. Financial support for this work was provided in part through NSF grant IIS 0845529. Thanks to all the participants.

References

1. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: Beat: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 477–486. ACM (2001)
2. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1), 39–52 (2004)
3. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) 27 (1), 5 (2008)
4. Ng-Thow-Hing, V., Luo, P., Okita, S.: Synchronized gesture and speech production for humanoid robots. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4617–4624. IEEE (2010)
5. Salem, M., Rohlfing, K., Kopp, S., Joublin, F.: A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In: RO-MAN, pp. 247–252. IEEE (2011)
6. Chartrand, T., Bargh, J.: The chameleon effect: The perception–behavior link and social interaction. Journal of Personality and Social Psychology 76(6), 893 (1999)

7. Lakin, J., Chartrand, T.: Using nonconscious behavioral mimicry to create affiliation and rapport. Psychological Science 14(4), 334–339 (2003)
8. Krämer, N.C., Simons, N., Kopp, S.: The effects of an embodied conversational agent's nonverbal behavior on user's evaluation and behavioral mimicry. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 238–251. Springer, Heidelberg (2007)
9. Neff, M., Toothman, N., Bowmani, R., Fox Tree, J.E., Walker, M.A.: Don't scratch! self-adaptors reflect emotional stability. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 398–411. Springer, Heidelberg (2011)
10. Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 222–235. Springer, Heidelberg (2010)
11. Lee, J.: Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions. PhD thesis, Massachusetts Institute of Technology (2011)
12. Cappella, J.N., Planalp, S.: Talk and silence sequences in informal conversations iii: Interspeaker influence. Human Communication Research 7(2), 117–132 (1981)
13. Kimbara, I.: On gestural mimicry. Gesture 6(1), 39–61 (2006)
14. Parrill, F., Kimbara, I.: Seeing and hearing double: The influence of mimicry in speech and gesture on observers. Journal of Nonverbal Behavior 30(4), 157–166 (2006)
15. Kimbara, I.: Gesture form convergence in joint description. Journal of Nonverbal Behavior 32(2), 123–131 (2008)
16. Bailenson, J.N., Yee, N.: Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science 16(10), 814–819 (2005)
17. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, L.-P.: Virtual rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 14–27. Springer, Heidelberg (2006)
18. Kopp, S.: Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52(6), 587–597 (2010)
19. Heloir, A., Neff, M., Kipp, M.: Exploiting motion capture for virtual human animation. Proceedings of the Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality at LREC 2010 (2010)
20. McNeill, D.: Gesture and thought. University of Chicago Press (2008)
21. NaturalSoft: `http://www.naturalreaders.com/`
22. Wolfe, D.A., Hollander, M.: Nonparametric statistical methods (1973)
23. John, O.P., Naumann, L.P., Soto, C.J.: Paradigm shift to the integrative big five trait taxonomy. In: Handbook of Personality: Theory and Research, vol. 3, pp. 114–158 (2008)
24. John, O.P., Donahue, E.M., Kentle, R.L.: The big five inventoryversions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research, Berkeley (1991)
25. Benet-Martínez, V., John, O.P., et al.: Los cinco grandes across cultures and ethnic groups: Multitrait multimethod analyses of the big five in Spanish and English. Journal of Personality and Social Psychology 75, 729–750 (1998)
26. von der Pütten, A.M., Krämer, N.C., Gratch, J.: How our personality shapes our interactions with virtual characters - implications for research and development. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 208–221. Springer, Heidelberg (2010)

Impact of Varying Vocabularies on Controlling Motion of a Virtual Actor

Klaus Förger[1], Timo Honkela[2], and Tapio Takala[1]

[1] Department of Media Technology
[2] Department of Information and Computer Science,
School of Science,
Aalto University,
Espoo, Finland
{klaus.forger,timo.honkela,tapio.takala}@aalto.fi

Abstract. An ideal verbally controlled virtual actor would allow the same interaction as instructing a real actor with a few words. Our goal is to create virtual actors that can be controlled with natural language instead of a predefined set of commands. In this paper, we present results related to a questionnaire where people described videos of human locomotion using verbs and modifiers. The verbs were used almost unanimously for many motions, while modifiers had more variation. The descriptions from only one person were found to cover less than half of the vocabulary of other participants. Further analysis of the vocabularies against the numerical descriptors calculated from the captured motions shows that verbs appeared in closed areas while modifiers could be scattered to disconnected clusters. Based on these findings, we propose modeling verbs with a hierarchical vocabulary and modifiers as transitions in the space defined by the numerical qualities of motions.

Keywords: motion capture, natural language, virtual actors.

1 Introduction

Animations and computer games have characters that act out scenes which an animator has designed. When creating these scenes, animators need believable human motion and ways to control the motion. To satisfy this need for human motion, many collections of captured motion have been made available [1]. Word based searching can be used to find suitable motions without a need to browse through the whole database. This way of searching corresponds to an ideal situation in which an actor would be always ready to act out motions based on short descriptions. In this paper, we concentrate on the effects of varying vocabularies on the motion searches.

In addition to words, human motion databases could also be searched by giving example motions or giving numerical requirements as search expressions. However, we limit the scope of the paper to collections of human motion where every motion clip is annotated with at least one written search term. The annotations can be the instructions given to an actor or opinions of persons viewing

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 239–248, 2013.

the motions. A potential problem is that a third person might not use or even understand the same vocabulary which was used in the annotations.

To find out how much variation there is in the vocabularies of people describing human motion, we constructed a questionnaire containing several different kinds of human locomotion. We asked people to describe the animated motion with one verb and from zero up to three modifiers which were adjectives or adverbs. Data from the questionnaire shows that variation between vocabularies of different people is large enough to cause potential misunderstandings.

We also present further analysis of the vocabularies against the numerical descriptors calculated from the captured motions. This analysis shows that verbs appear in closed areas whereas modifiers can be scattered to disconnected clusters. Based on these finding, we discuss what are the best ways to model the vocabularies.

2 Related Work

Controlling virtual actors with natural and unrestricted language requires creating links between the describing words and physical motions. A simple approach for creating the links is manual annotation which means writing labels for every motion. The task can be made easier by calculating descriptor values which reflect the quality of the motion [2]. The motion descriptors allow generalizing annotations as we can assume that two motions that are numerically close to each other are likely to be annotated in the same way. In this paper, we use motion descriptors when comparing motions.

Many methods and systems designed for controlling virtual characters assume that there is a small selection of allowed commands [3–5]. More fine grained control of both style and length of motions performed by a virtual character could be desired. This can be achieved with real-time interaction rules between two virtual characters, as the rules are based on continuous parameters [2]. However, the set of parameters can feel artificial to the end user, especially if the parameters are derived from the numerical qualities of the motions. Motion analysis frameworks such as Laban Motion Analysis (LMA) assume that the user knows a set of expert terms for describing human motion such as the Laban notation [6]. It has been found that systems allowing the use of natural language can reduce the amount of expertise and time needed in controlling virtual actors [7]. A challenge in natural language processing is that people can have subjective views on the meaning of words [8]. Our interests are in finding out how much manual annotation and analysis of motion is needed to enable controlling a virtual actor with natural language.

The assumption that, a small amount of motion classes is enough, does not appear only in systems that control virtual characters. Commonly used motion databases are often based on a selection of words given to the actors who perform the motions [9]. This can result in databases with plenty of motions, but where all the motions belong to stereotypical categories. A reason for taking shortcuts in annotation is that manual annotation can take a lot of time and effort [10].

As a motion database with annotations by several persons was not available, we decided to create one.

There are methods for creating new motions with different styles by using a selection of parameters which may be stylistic and emotional [11] or related to the trajectories of the motions [12]. These methods enrich a motion database as they create new motions by blending existing ones. We decided to use motion blending as it allows producing motions between stereotypical classes.

Three questions of interest were left open by the related works. How sufficient annotations from a single person are when building natural language descriptions? Do people describe the same motions with several synonyms? Do people have different opinions about the meaning of the used words? To answer these questions, we created a motion collection and a questionnaire which are presented in the next sections.

3 Motion Data Generation

To study natural language descriptions of human motion, we first needed a collection of motions to be described. We chose locomotion as it appears commonly in animations and it also allows displaying many motion styles. In order to create a set of motions that would have variation in both verbs and modifiers, we decided to use a mix of acted motions and interpolations between those motions. We recorded short locomotion sequences with two actors using Optitrack motion tracking system. The actors were asked to perform walking and limping with styles 'sad', 'slow', 'regular', 'fast' and 'angry'. Running was recorded with only the styles 'slow' and 'fast' as the limited capture area made recording running challenging. To make the motions easy to interpolate, the actors were instructed to always start from the same position with their right leg and to perform the motions towards the same direction.

The blended motions were produced with three steps which were initial alignment, time warping and interpolation. In the first step, the supported and lifted phases of the feet were detected and aligned among the motions. The second step was time warping the motions to make them synchronized. The aligned frames between the supported and lifted phases were matched and the rest of the frames were re-sampled to get a smooth frame rate. As the last step, the coordinates of the root joints were interpolated linearly and the joint rotations were interpolated as quaternions with the slerp algorithm [13].

We used two-way and three-way interpolation to create the blended motions. In the two-way case, three new motions are created with steps of 25%. In the three way case, we created all the two-way combinations, three motions with the percentages 70%-15%-15% and one motion with an even split of 33%-33%-33%. Ideally, we would have created blends from all possible combinations of the original motions, but that would have resulted in too many to be viewed reasonably. Also, some motions like fast running and slow walking were too different to be interpolated. We ended up creating blends between the motions that had a similar style and also between motions that had the same intended verb. The combinations used in the blends are shown in Figure 1.

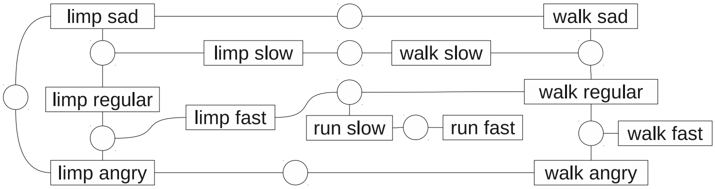


Fig. 1. The boxes show the original captured motions with the instructions given to the actors, the circles represent the combinations used in motion blending

4 Questionnaire and Methods for Analysis

The idea of the questionnaire was to collect verbs and modifiers that describe the motions. The questionnaire was web based and all the motions were shown as videos with a stick figure character as shown in Figure 2. The duration of the videos ranged from 3 seconds (fast running) to 12 seconds (slow limping). Finnish language was used in the questions and the answers. The participants were gathered through work contacts and social media.

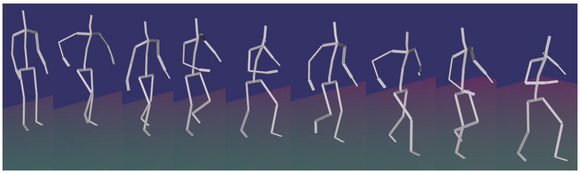


Fig. 2. An example of the stick figure representation portraying an angry walk

The task given to the participants was to describe the seen motion with one verb or phrase (such as 'swimming' or 'mountain climbing') and from zero up to three modifiers (such as 'colorfully' or 'very colorfully'). To make the answering easier we divided the videos into three sets and the participants could answer as many sets as they liked. Set A included all unmodified motions and had 24 videos, set B had 40 motions which were 50%-50% interpolations and the set C had the rest 60 motions. The total amount of videos was 124.

Our first research question is: Is the collective vocabulary used by a group of annotators larger than the target vocabulary given to actors of the motions and larger than the vocabulary of a single annotator? An answer to this question helps in deciding how much effort should be put into developing better search terms for motion databases. A way to find an answer to this question is to calculate how much of the collective vocabulary would be covered by the terms given to the actors and the words used by a single participant.

The second research question is: Can the variation in the collective vocabulary be decreased by finding synonyms? An answer to this question is important as joining search terms requires only a small change in a motion database. This question requires qualitative grouping and analysis of the vocabulary. Comparison of the distributions of the words over the motion samples can also help recognizing synonyms. We use FinnWordNet [14] as the source of the definitions of the words. The translations of the Finnish words into English are also based on the FinnWordNet as it contains professionally made translations.

The third research question is: Do people have different opinions about the meaning of the used words? If there are large variations in how people use the same words, it would make building an optimal motion search much harder as the subjectivity would have to be taken in to account. Answering this question calls for plotting the distributions of the describing words on a space that is defined by numerical qualities of the motions.

To form a space which is based on the qualities of the motion, we calculate describing values called motion descriptors which include coordinates, velocities, accelerations and rotations as quaternions of each joint. From the velocities we used both absolute values and the velocities separately along the x, y and z axes. Also, we included the distances between pairs of body parts in a set that includes hips, neck, head, elbows, hands, knees and feet. To remove the variation caused by physical differences between the actors, we removed the personal means of descriptors as that has been found to help classification of motions [15].

5 Results

The participants of the questionnaire consisted of 9 females and 13 males with ages between 21 to 70 years. For the participants, the previous experiences with human motion were mainly linked to sports related hobbies. All 22 participants completed the set A, 10 also completed the set B and 2 participants did all the three sets of videos. Varying inflections which do not affect the meaning in this context such as 'walk' and 'walking' were cleaned from the data.

In the analysis, we have two points of view to the vocabularies. The first is the plain vocabulary where all the used words are considered equally important. The second is the shared vocabulary in which a word used by N persons is N times more important than a word used by one person. The distribution of the shared vocabulary is shown in Figure 3. From the figure we can see that 88 unique verbs and 233 unique modifiers were used by the participants. It also shows that the most common words explain a large part of the word usage, but there is also a long tail of rarely used words. For example nine most used verbs explain 50% of the shared vocabulary, but in order to reach to 90% one must consider 65 verbs.

Coverage of the words which were given to the actors and the words used by an average annotator are shown in Figure 4. Analysis of the vocabularies in Figure 4 is limited to the 24 videos in the set A as we needed to have annotations from all the participants to make a fair comparison. For the other analyses all the motion sets were used. Acted verbs plotted in Figure 4 have only coverage of 3% in the plain vocabulary as the three verbs given to actors are only a small part of the total 88 used unique verbs. However, when considering the shared vocabulary the three words have coverage of 29%. This comes from the fact that walking (kävelee) was used by all the 22 participants, running (juoksee) by 19 and limping (ontuu) by 14, while the total sum of usage counts was 190.

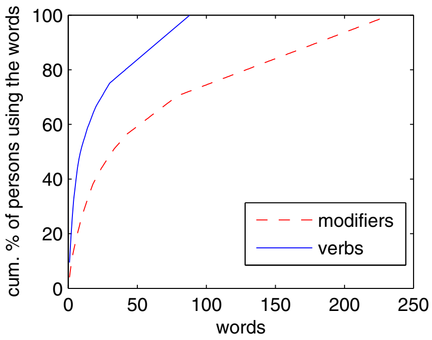


Fig. 3. Cumulative percentage of coverage of the shared vocabulary. The words are sorted from most used to the least used.

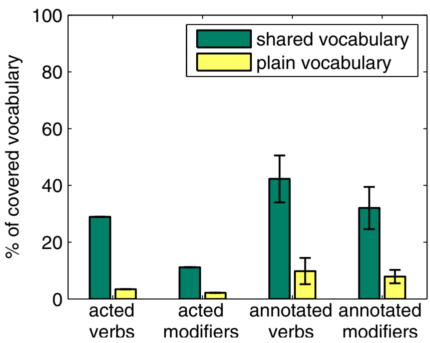


Fig. 4. Coverage of the plain vocabulary and shared vocabulary for verbs and modifiers given to the actors and average coverage of annotations from a single person. Standard deviation is shown for the averages.

Our first research question was related to how much the words given to the actors or the words of a single annotator cover of the overall vocabulary. The answer based on Figure 4 is that in the best case the words given to the actors cover a third of the vocabulary. Therefore, we can say that the set of words given to the actors of the motions would not enable making motion searches with natural language. The vocabulary of an average annotator does work better as it covers nearly 50% of the verbs in the shared vocabulary. Still, there is room for improvement. The coverage of the plain vocabulary is less than 10% which shows that having only one annotator will cause missing many rarely used words.

For finding synonyms, we used dictionary definitions of words and their translations to English as provided by FinnWordNet [14]. The words 'ontuu', 'nilkuttaa', and 'linkuttaa' are synonyms based on dictionary definitions and they all translate to 'limps' in English. This also shows that they could be considered to be alternative labels for the exactly same motions. From the modifiers we could not find synonyms as easily as from the verbs. Modifiers such as 'nopeasti' – 'fast' and 'kiirehtien' – 'hurriedly' can be considered to be similar, but whether they are synonyms is uncertain based on the data from the questionnaire.

For seeing the relationship between the numerical qualities of the recorded motions and the words used in the descriptions, we plotted the nine most frequent verbs (Fig. 5) and nine most frequent modifiers (Fig. 6) onto the PCA (principal component analysis) space based of the motion descriptors. To make the figures more readable we added small offsets to the overlapping pies to separate them. Web based versions of the two figures that also show the related animated motions are available at: http://research.ics.aalto.fi/cog/mglt/

Figure 5 shows that for many motions vast majority of the annotators are unanimous about the verbs. The three alternative words for 'limping' appear in the same area of the map and cause division between the annotators, but joining those words as synonyms would clean up the division. Two subjective divisions which cannot be accounted to synonyms are visible in the verbs. The first is between 'jogging' and 'running'. It seems that the participants could not agree where to draw a line between the two actions. The second subjective division is between 'walking' and 'limping'. While 'walking' has an area that is almost unanimously 'walking', almost all of the 'limping' motions have also a small share of 'walking' in them.

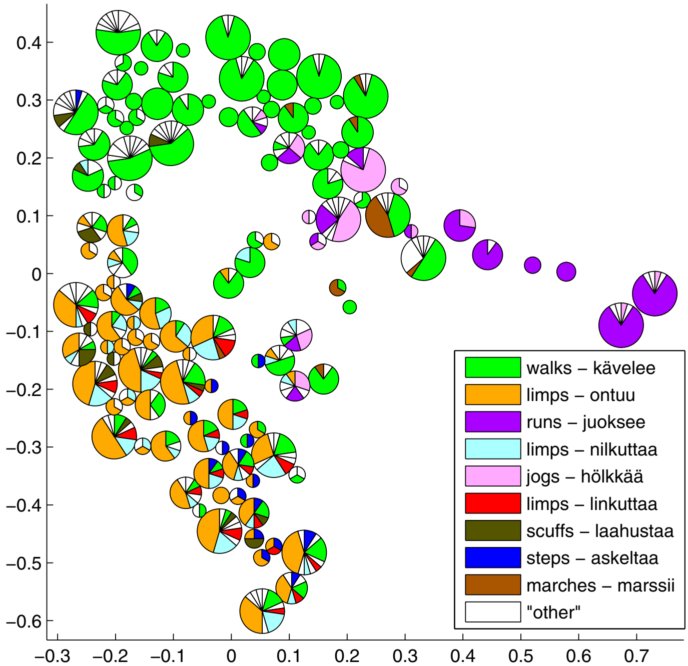


Fig. 5. Distributions of most common verbs for each motion mapped on the first and second normalized PCA components. The surface area of the pies is proportional to the number of answers and the position of the pies reflect the style of the motions.

Modifiers plotted in Figure 6 show that the participants were less unanimous in their answers than with verbs. There are even cases where almost all the participants gave different modifiers. Part of the variation can be explained by the fact the participants could give up to three modifiers. Still, even limiting the analysis to the first given modifiers, there would be no videos where one word would cover more than 50% of the answers if the video got more than two answers.

Many of the words are limited to a part of the PCA space. Verbs in Figure 5 form connected areas while modifiers can have disconnected distributions. For example the modifier 'slowly' appears mostly in the left side of Figure 6 where are the verbs 'walking' and 'limping', but also a few times near the center where the motions are described as 'jogging' or 'running'. The greater variation of modifiers is visible as the greater amount of the class 'other' than in the verbs.

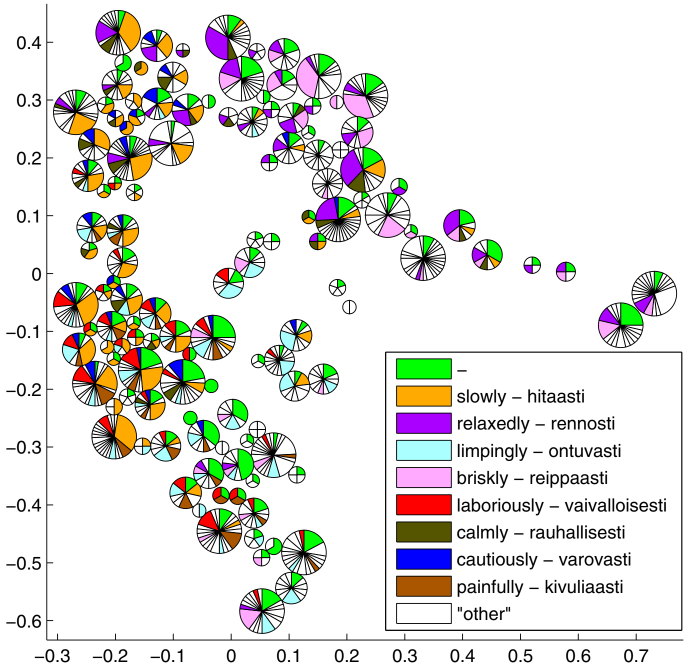


Fig. 6. Distributions of most common modifiers for each motion mapped on the first and second normalized PCA components. The surface area of the pies is proportional to the number of answers and the position of the pies reflect the style of the motions.

6 Discussion

How do the results of the questionnaire guide building a virtual actor that could be controlled with natural language? The first lesson is that relying only on the words given to the actors is not likely to cover the required vocabulary. Having one person annotate all the motions works better. However, the annotations of a single person are not enough in the cases where the borders between different verbs are subjective or when several synonyms exist. Modifiers are more challenging than verbs as the participants were far from unanimous and the modifiers did not always form continuous areas in the descriptor space.

For the verbs, hierarchical style of description could be beneficial as that would allow using words in a general sense and in a more specific sense. For example a parent category 'walking' could be divided into subcategories 'limping' and 'walking'. This way part of the subjectivity could be taken into account without needing more than one annotator. In practice, this could be achieved by giving the annotators two motions and a task to describe the motions with one verb.

Verb-modifier combinations could act as the most specific level of the description hierarchy. However, this would mean annotating a large amount of verb-modifier combinations. A more practical approach to handle modifiers could be to treat them as transitions in the descriptor space instead of areas of the space.

For a user instructing a virtual actor, this would mean first saying a verb and then saying a modifier to adjust the style of motion towards a desired direction. This approach could fix the problems caused by discontinuities in the distributions of modifiers. For example starting from walking and moving repeatedly towards a faster motion style would end up in a running motion. To find out what transitions correspond to which modifiers, a comparative task such as 'motion A is more X than motion B' should be given to the annotators.

While the questionnaire could always be made better, the main factor that speaks for the questionnaire is that the participants were able to freely select the words they used. If a selection of possible words had been given, it would have distorted the vocabularies of the participants. The decision to analyze the vocabularies as words-per-person instead of words-per-video makes our results more general. The counts for words-per-video are closely tied to the selection of videos, but the counts for words-per-person should not change dramatically even if part of the videos would be shown more times than others. One shortcoming in the questionnaire is the lack of repetitions. From data with repetitions, we could analyze how much of the variation in the descriptions is caused by difficulty of deciding between possible alternatives.

7 Conclusions and Future Work

In this paper, we presented results from a questionnaire in which participants were asked to describe videos of human locomotion with one verb and from zero up to three modifiers which were adjectives or adverbs. We analyzed the vocabulary as such and also in connection with numerical motion descriptors calculated from the motions. The results show that the original words given to the actors of the motions did not cover the used vocabulary of the participants viewing the motions. The vocabulary of a single annotator had better coverage, but the data would not help in cases where several synonyms exist for a verb or when the exact definition of a verb is not shared between the participants. The results also show that the modifiers used in describing the motions contain more variation than the verbs.

The main use case we considered was a virtual actor that can be controlled with natural language. Based on our results, we conclude that just linking each motion with the describing words would not allow controlling a virtual actor accurately. The linking would not take into account that meaning of verbs can be subjective and that modifiers are used variedly. The improvements we are planning include building a hierarchical vocabulary for verbs and modeling modifiers as transitions in the space defined by the numerical qualities of the motions. Realizing these improvements requires changing the annotation method from annotation of one motion at a time to annotation where similarities and differences are described between two motions.

Acknowledgments. This work has received funding from the Hecse doctoral program and the project Multimodally grounded language technology (254104) which is funded by the Academy of Finland.

References

1. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Action dataset - A survey. In: Proc. of SICE Annual Conference 2011 (SICE 2011), pp. 1650–1655 (2011)
2. Förger, K., Takala, T., Pugliese, R.: Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 341–354. Springer, Heidelberg (2012)
3. Blumberg, B., Galyean, T.: Multi-level direction of autonomous creatures for real-time virtual environments. In: Mair, S.G., Cook, R. (eds.) Proc. of SIGGRAPH 1995, pp. 47–54. ACM, New York (1995)
4. Perlin, K., Goldberg, A.: Improv: a system for scripting interactive actors in virtual worlds. In: Proc. of SIGGRAPH 1996, pp. 205–216. ACM, New York (1996)
5. Vilhjálmsson, H.H., et al.: The behavior markup language: Recent developments and challenges. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 99–111. Springer, Heidelberg (2007)
6. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. In: IEEE International Workshop on Robot and Human Interactive Communication 2005 (ROMAN 2005), pp. 294–299. IEEE (2005)
7. Talbot, C., Youngblood, G.: Spatial Cues in Hamlet. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 252–259. Springer, Heidelberg (2012)
8. Honkela, T.: Raitio, j., Lagus, K., Nieminen, I., Honkela, N., Pantzar, M.: Subjects on objects in contexts: Using GICA method to quantify epistemological subjectivity. In: Proc. of International Joint Conference on Neural Networks (IJCNN 2012), pp. 2875–2883 (2012)
9. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
10. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently Scaling up Crowdsourced Video Annotation. International Journal of Computer Vision 101(1), 184–204 (2012)
11. Rose, C., Bodenheimer, B., Cohen, M.F.: Verbs and Adverbs: Multidimensional Motion Interpolation Using Radial Basis Functions. IEEE Computer Graphics and Applications 18(5), 32–40 (1998)
12. Kovar, L., Gleicher, M.: Automated Extraction and Parameterization of Motions in Large Data Sets. In: Marks, J. (ed.) Proc. of SIGGRAPH 2004, pp. 559–568. ACM, New York (2004)
13. Shoemake, K.: Animating rotation with quaternion curves. ACM SIGGRAPH Computer Graphics 19(3), 245–254 (1985)
14. Lindén, K., Carlson, L.: FinnWordNet - WordNet på finska via översättning (In English: FinnWordNet - Finnish WordNet by Translation). LexicoNordica - Nordic Journal of Lexicography 17, 119–140 (2010)
15. Bernhardt, D., Robinson, P.: Detecting affect from non-stylised body motions. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 59–70. Springer, Heidelberg (2007)

Conversational Gaze Aversion for Virtual Agents

Sean Andrist, Bilge Mutlu, and Michael Gleicher

Department of Computer Sciences, University of Wisconsin–Madison
1210 West Dayton Street, Madison, WI 53706, USA
{sandrist,bilge,gleicher}@cs.wisc.edu

Abstract. In conversation, people avert their gaze from one another to achieve a number of conversational functions, including turn-taking, regulating intimacy, and indicating that cognitive effort is being put into planning an utterance. In this work, we enable virtual agents to effectively use gaze aversions to achieve these same functions in conversations with people. We extend existing social science knowledge of gaze aversion by analyzing video data of human dyadic conversations. This analysis yielded precise timings of speaker and listener gaze aversions, enabling us to design gaze aversion behaviors for virtual agents. We evaluated these behaviors for their ability to achieve positive conversational functions in a laboratory experiment with 24 participants. Results show that virtual agents employing gaze aversion are perceived as thinking, are able to elicit more disclosure from human interlocutors, and are able to regulate conversational turn-taking.

Keywords: Gaze aversion, virtual agents, conversational behavior, intimacy, disclosure, turn-taking.

1 Introduction

Engaging in mutual gaze with others has long been recognized as an important component of successful social interactions. People who exhibit high amounts of mutual gaze are perceived as competent, attentive, and powerful [4]. In the same way, virtual agents that use eye contact to exhibit some degree of mutual attentiveness have been shown to achieve a number of positive social and conversational functions, including building rapport with people [20] and increasing positive perceptions of affiliation [2].

Similarly, engaging in gaze *aversion* in conversation also serves a number of communicative functions. Gaze aversions are used to *signal cognitive effort* [4], *modulate intimacy* [1], and *regulate turn-taking* [14]. While social science literature has highlighted the positive functions of gaze aversion, it does not provide the precise temporal measurements required to synthesize a model of gaze aversion for virtual agents that could achieve these functions.

In this work, we enable virtual agents to use gaze aversions to more effectively engage in conversations with people. We first present an analysis of a video corpus of human dyadic conversations from which we obtained temporal parameters of gaze aversion. From these temporal parameters, we designed a

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 249–262, 2013.

Fig. 1. The four agents used in our experiment: Norman, Jasmin, Lily, and Ivy. Norman, Jasmin, and Lily are performing gaze aversions in different directions, while Ivy is maintaining mutual gaze with her interlocutor.

gaze controller that can generate appropriately timed gaze aversion behaviors for virtual agents. We also present an experimental evaluation of these behaviors to demonstrate their effectiveness in achieving their intended conversational functions. In this experiment, human participants interacted with four different virtual agents in four conversational tasks, each of which was designed to test a different conversational function of gaze aversion (Figure 1).

2 Background

In this section, we present an overview of relevant social and cognitive science research on human gaze aversion. We then review related work on designing effective gaze behaviors for virtual agents.

2.1 Gaze Aversion in Humans

Previous social science research has identified a number of underlying mechanisms to explain human gaze aversion and the social functions it achieves. One such mechanism relevant to our work is the "cognitive interference hypothesis" [5] [8] [9] [12]. This hypothesis posits that gaze aversions facilitate cognitive activity by disengaging the speaker from the environment and limiting visual inputs. Research to support this hypothesis has shown that mutual gaze significantly interferes with the production of spontaneous speech [5]. Research also shows that forcing oneself to look away from a conversational partner while retrieving information from long-term memory or when planning a response to a challenging question significantly improves performance [12] [18].

Previous research has also shown that eye contact is a significant contributor to the intimacy level of an interaction, such that reducing eye contact can decrease the perceived intimacy of a conversation [4]. For example, people generally engage in less eye contact while responding to embarrassing questions than while responding to less objectionable questions [10]. Other work has examined how topic intimacy and eye contact interact over the course of a conversation [1].

Another primary function of gaze aversion is to facilitate turn-taking. Just as making eye contact while listening can serve as a signal that the conversational floor is requested, breaking eye contact while speaking can serve as a signal that the conversational floor is being held and that the speaker has more to say [21]. Kendon [14] found that speakers often look away from their addressees at the beginning of utterances to claim the speaking turn and then look back toward their addressees at the end of their utterance, yielding the turn.

In this work, we group the social-scientific findings discussed above into three broad conversational functions: the *cognitive*, *intimacy-modulating*, and *turn-taking* functions of gaze aversion. These groupings informed our empirical investigation to develop a more computational understanding of how gaze aversions are temporally employed in conversation.

2.2 Gaze Aversion in Virtual Agents

An agent's gaze behavior plays a key role in achieving rich interactions. Well-designed gaze mechanisms—e.g., shifting gaze at turn boundaries during conversation—result in increased task performance and more positive subjective evaluations [13]. Coordinating the head and eyes to maintain a high degree of attention toward human interlocutors has been shown to increase feelings of affiliation with virtual agents [2]. Poor gaze behavior can be worse than the absence of gaze behavior. The positive effects of having an embodied agent—as opposed to only audio or text—can be completely lost if gaze is very poor or random [11].

Previous work has studied different conversational functions of gaze in human-agent interactions, e.g., the use of gaze in facilitating turn management [7] [17]. Wang and Gratch [20] have shown that a virtual agent that continuously gazes toward a human interlocutor is able to increase perceptions of rapport when the gaze is accompanied by nonverbal indicators of positivity and coordination. Continuous gaze without these accompanying behaviors had a negative social impact. Lee et al. [15] developed a statistical model of quick saccadic eye movements for a virtual agent to employ while speaking and listening. This work does not consider the strategic deployment of longer gaze aversions that can be used to achieve specific interactional goals.

While previous research has explored how agents can use gaze to achieve positive social outcomes, a precise account of when agents should avert their gaze from human conversational partners and what social functions these aversions might achieve is still needed. Our work seeks to address this knowledge gap from both theoretical and empirical perspectives through the application of existing social-scientific knowledge and a study of human dyadic conversations to design gaze aversion behaviors for virtual agents.

3 Interaction Design

As outlined above, research in the social sciences has identified a number of conversational functions of gaze aversion. To extend this knowledge to include temporal patterns that will be directly implemented on virtual agent systems, we collected video data from 24 dyadic conversations and derived statistical parameters for the length, timing, and frequency of gaze aversions in relation to speech and conversational functions. We addressed three primary conversational functions of gaze aversion in this analysis, which are defined and described below.

Cognitive – These gaze aversions serve to disengage a speaker's attention from the face of their interlocutor in order to facilitate thinking and remembering [12]. With these aversions, people signal that cognitive processing is occurring while creating an impression that deep thought or creativity is being undertaken [4].

Intimacy-modulating – Gaze aversions also serve to moderate the overall intimacy level of the conversation. Periodic gaze aversions while listening can serve to make speakers more comfortable and reduce negative perceptions associated with staring [1].

Turn-taking – These gaze aversions serve to regulate conversational turn-taking. By looking away at the beginning of an utterance, the speaker strengthens his or her claim over the speaking turn. Looking away during a pause in speech indicates that the conversational turn is being held and that the speaker should not be interrupted [14].

3.1 Data Collection and Analysis

We recruited 24 females and 24 males, aged 18 to 28 and previously unacquainted, for our study. Each dyad engaged in a structured conversation for approximately five minutes. One participant was instructed to learn about the other participant's taste in movies, with the goal of making a movie recommendation. We counterbalanced all conversations for both gender—female and male—and conversational role—recommender and recommendee. We also counterbalanced gender concordance—there was an equal number of gender-matched and gender-mismatched dyads.

Using VCode,[1] we analyzed the recorded videos of the participants' gaze and speech. Video coding was carried out by two independent coders with partial overlap. Sequences of time spent speaking and averting gaze were annotated. Cognitive events were marked as discrete points in time where the participants appeared to be thinking or remembering, commonly occurring at the beginning of responses to questions.

Gaze aversions were coded for the conversational function that they were perceived to be supporting: cognitive, intimacy-modulating, or turn-taking. This coding took place in three passes. In the first pass, the coder was instructed to mark gaze aversions as cognitive if they occurred near labeled cognitive events, e.g., when a participant appeared to be thinking of something new to say. In the

[1] `http://social.cs.uiuc.edu/projects/vcode.html`

Table 1. Gaze aversion parameters in relation to conversational functions and coordinated with (before, after, or within) speech and cognitive events

Conversational Function	Coordinated With	Parameter	Value
Cognitive	Cognitive Event	Length (sec)	3.54 ($SD = 1.26$)
		Start time (sec)	1.32 before ($SD = 0.47$)
		End time (sec)	2.23 after ($SD = 0.63$)
Intimacy	Speaking	Length (sec)	1.96 ($SD = 0.32$)
		Between (sec)	4.75 ($SD = 1.39$)
	Listening	Length (sec)	1.14 ($SD = 0.27$)
		Between (sec)	7.21 ($SD = 1.88$)
Turn-taking	Utterance Start	Frequency (%)	73.1
		Length (sec)	2.30 ($SD = 1.10$)
		Start time (sec)	1.03 before ($SD = 0.39$)
		End time (sec)	1.27 after ($SD = 0.51$)
	Utterance End	End time (sec)	2.41 before ($SD = 0.56$)

second pass, gaze aversions were marked as turn-taking if they occurred near the beginning of a speaking turn and were not previously labeled as cognitive. In the third pass, all remaining gaze aversions were labeled as intimacy-modulating. An inter-rater reliability analysis showed substantial agreement on the identification of gaze aversions and their conversational function (Cohen's $\kappa = .747$).

From our analysis, we obtained timing statistics for different kinds of gaze aversions, including the frequency, length, and temporal placement of these gaze aversions relative to speech (Table 1). We also labeled each gaze aversion for its direction, categorized as *up*, *down*, and *side* (Table 2).

3.2 Designing Gaze Aversion for Virtual Agents

Findings from the data analysis were synthesized into a gaze controller for virtual agents that automatically plans and performs gaze aversions to accomplish the conversational functions previously discussed. This controller takes as inputs the current conversational state, the start time and length of upcoming planned utterances, and the time of upcoming cognitive events, and then supplies as outputs the start and end times of planned gaze aversions to be executed by the agent. The exact timings of the gaze aversions are drawn from the parameter distributions shown in Table 1. These distributions are modeled as Gaussian functions in the current implementation.

Source of inputs – Recognized speech from the user is passed to a dialogue manager that associates a semantic tag with the utterance and plans the agent's speech accordingly. For example, if the dialogue manager receives a recognized question, it will produce the associated answer. The dialogue manager sends

Table 2. Frequency of gaze aversions up, down, and to the side for each conversational function

Conversational Function	Frequency Up	Frequency Down	Frequency Side
Cognitive	39.3%	29.4%	31.3%
Intimacy-modulating	13.7%	28.8%	57.5%
Turn-taking	21.3%	29.5%	49.2%

upcoming cognitive events, speech events, and the current conversational state to the gaze controller. Cognitive events could alternatively be passed to the gaze controller from a dedicated cognitive architecture, but in our implementation, cognitive events were created by labeling some of the agent's utterances as "cognitively difficult" and generating a cognitive event at the beginning of those utterances.

Gaze controller – Cognitive events are represented with a single timestamp, t_c. Planned speech events are represented as a vector containing start and end times, $[t_s, t_e]$. Conversational state, CS, indicates that the agent is currently in either *speaking* or *listening* mode. As the gaze controller receives these inputs from the dialogue manager, it continuously plans future gaze aversions in real-time. The first priority is to plan gaze aversions around upcoming cognitive events, t_c. The start and end times of the gaze aversion, $[GA_s, GA_e]$, are computed by drawing from the cognitive parameter distributions shown in Table 1. The controller next looks for upcoming speech events and calculates first if a turn-taking gaze aversion will be performed. If a gaze aversion will be performed, the controller then calculates $[GA_s, GA_e]$ around the start of the utterance, t_s, by drawing from the turn-taking parameter distributions provided in Table 1. Finally, the controller calculates the next intimacy gaze aversion according to CS. These gaze aversions are only planned for times when cognitive and turn-taking aversions are not already planned. Also, intimacy gaze aversions are prohibited near the end of utterances, t_e, so that virtual agents can appropriately pass the floor by maintaining mutual gaze.

Example simulation – Figure 2 illustrates a simulation of the gaze aversion behaviors produced by our controller. In this example, two agents, A1 and A2, are having a conversation. Both are using the gaze aversion controller. A1 asks a question constructed from two utterance parts with a pause in between. A turn-taking gaze aversion is planned and executed around the start of the second utterance in order to hold the conversational floor. While A1 is listening, it occasionally looks away to regulate the intimacy of the conversation. Upon recognizing A1's question, A2 plans to give its response, which has been tagged with a cognitive "thinking" event at its beginning. The gaze controller plans and executes a cognitive gaze aversion around the beginning of the utterance to express this thinking. All other gaze aversions in the example have been similarly produced by the controller to achieve one of the three conversational functions.

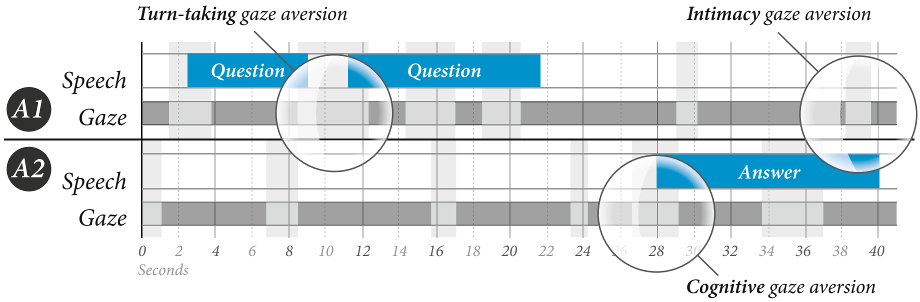


Fig. 2. Gaze aversions created by our controller for two agents in conversation. Dark gray intervals on the gaze stream indicate periods of gazing toward the interlocutor, and light gray intervals indicate gaze aversions.

4 Experimental Evaluation

We developed four hypotheses to test how agents might use the gaze aversion behaviors generated by our controller to achieve conversational functions. The first two hypotheses focus on the cognitive function, the third on the intimacy-modulation function, and the fourth on the turn-taking function.

Hypothesis 1 – A virtual agent averting its gaze while not currently speaking will be perceived as *thinking*, whereas an agent that does not avert its gaze will not elicit this impression.

Hypothesis 2 – Virtual agents that display gaze aversions at the start of utterances will be rated as being more *thoughtful* and creative than virtual agents that do not display gaze aversions.

Hypothesis 3 – Virtual agents that display periodic gaze aversions while listening will increase a human interlocutor's comfort and elicit more *disclosure* than agents that do not display gaze aversions.

Hypothesis 4 – Virtual agents that display gaze aversions during pauses will be perceived as *holding the floor* and will be interrupted less than agents that do not display gaze aversions.

4.1 Study Design

Twenty-four participants were recruited for this study (12 females and 12 males), aged between 18 and 45 ($M = 23$, $SD = 6.82$). All participants were native English speakers and were recruited from the University of Wisconsin–Madison campus.

The experiment involved a single independent variable, *gaze aversion condition*, with three conditions varying between participants. One condition involved the virtual agents using gaze aversions generated by the controller described in the previous section, which we call the *good timing* condition. The other two conditions were baselines for comparison. The first baseline was a *static gaze* condition in which the virtual agents did not employ any gaze aversions. The second baseline was a *bad timing* condition in which the virtual agent employed just as many gaze aversions as in the *good timing* condition but with reverse timings. When the gaze controller indicated that a gaze aversion should be made, the *bad timing* model engaged a mutual gaze shift, and vice versa. This third condition was included as a baseline to show that both the presence and the timing of gaze aversions are important for achieving positive social outcomes.

We created separate tasks to test each hypothesis, each using a different virtual agent (Figure 1). Participants were randomly assigned to one of the three gaze aversion conditions, which was held constant for all four tasks (8 participants per condition). Tasks were presented in random order.

Task 1 – The first task was designed to test Hypothesis 1. The participant was told that the virtual agent, Norman, was training to work at a help desk in a campus library. The participants were given five library-related questions to ask Norman. They were instructed to ask each question and listen to the response. Norman would pause for 4 to 10 seconds (randomly determined) before answering each question. Participants were instructed to ask a question again if they thought Norman did not understand or was not going to answer. The primary measure was the time participants waited for Norman to respond to questions before interrupting him to ask the question again. For this task, we deliberately chose an agent with an abstract design that minimally elicits attributions of intent or thought in order to ensure that the agent's gaze aversions were solely responsible for the impression of thinking, unconfounded from any other animation variables.

Task 2 – The second task was designed to test Hypothesis 2. For this task, participants were instructed to ask the agent, Jasmin, a series of five common job interview questions. Jasmin was programmed to respond with answers taken from real-world job interviews. Participants rated each response immediately after it was given on four seven-point rating scales. These scales measured the perceived thoughtfulness, creativity, disclosure, and naturalness of each response. In our analysis, we combined the scales into a single broad indicator of *thoughtfulness*. Internal consistency was excellent for this measure (Cronbach's $\alpha = .903$).

Task 3 – The third task was designed to test Hypothesis 3. In this task, participants spoke to an agent named Lily, who was introduced as training to be a therapist's aide who would conduct preliminary interviews with incoming clients. Lily asked the participant a series of five questions of increasing intimacy, and participants were instructed to respond with as much or as little detail as they wished. Questions ranged in intimacy, from "What do you like to do in your free time?" to "What is something you would like to accomplish before dying?" The primary measure for the third task was the *degree of self-disclosure*, specifically the breadth of disclosure. Breadth of disclosure was obtained using a word count of participants' responses to Lily's questions. Word count has been validated as an appropriate measure of disclosure in previous research on how computers can be used to elicit self-disclosure from people [16].

Task 4 – The fourth task was designed to test Hypothesis 4. Participants were provided with a list of five questions to ask a virtual agent named Ivy,

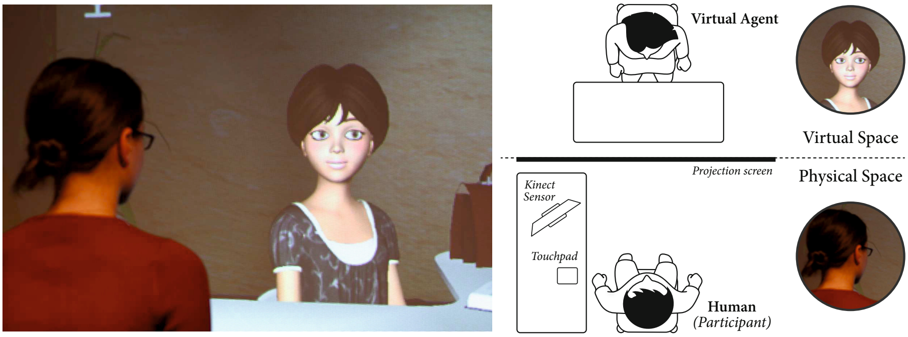


Fig. 3. An experimenter demonstrating the interaction with the virtual agent on a life-size projected display (left) and the physical setup of the experiment (right)

with the goal of getting to know each other. Participants were instructed to ask each question, listen to Ivy's response, and then reciprocate with their own response to the same question. Ivy's responses had two parts, separated by a pause between 2 and 4 seconds in length (randomly determined). If participants started speaking during the pause, Ivy refrained from giving the second part of her response. The primary measure of the fourth task was the time participants waited for Ivy to be silent during the pause in her speech before interrupting.

4.2 Setup and Procedure

The experiment was implemented using a custom character animation framework built on top of the Unity game engine.[2] The agent's behaviors were implemented as Unity scripts. In the second, third, and fourth tasks, the agents periodically smiled, blinked, and nodded their heads to achieve greater naturalness and humanness. In all tasks and conditions, the agent's eyes made small, periodic saccadic motions according to the model presented by Lee et al. [15]. The agents were created using commercially available parametric base figures. Audio and lip-sync animations were pre-recorded.

Gaze aversions were executed using the head-eye coordination model described by Andrist et al. [3] with a moderate amount of head movement. Head alignment was high as the agent oriented its gaze back to the interlocutor, in accordance with the finding that high head alignment increases people's feelings of affiliation with agents [2].

After giving informed consent, the experimenter led each participant into the study room and gave a brief introduction to the experiment. The participant sat in a chair approximately six feet away from a large screen on which the life-size virtual agent was projected (Figure 3). A wireless touchpad was used as a button to begin each conversational task, and a Kinect microphone was used for capturing speech. The Microsoft Speech Platform[3] was used for speech

[2] http://www.unity3d.com
[3] http://msdn.microsoft.com

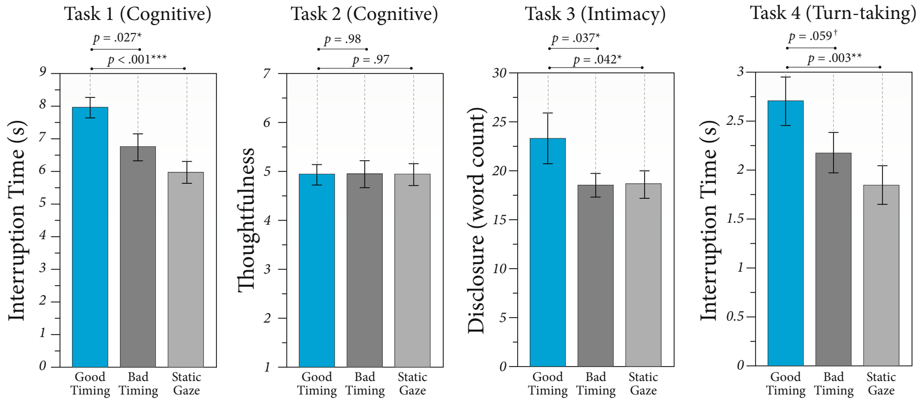


Fig. 4. The results of the evaluation. Virtual agents that displayed gaze aversions with appropriate timings successfully conveyed the impression that they were "thinking," elicited more disclosure from participants, and were better able to hold the conversational floor during breaks in speech. (†), (*), (**), and (***) denote $p < .10$, $p < .050$, $p < .010$, and $p < .001$, respectively.

recognition in combination with a custom dialogue manager specific to each task. After completing all four tasks, the participant responded to a survey of demographic characteristics and was debriefed. The study took approximately 30 minutes, and each participant was given $5 as compensation.

4.3 Results

We performed a mixed-design analysis of covariance (ANCOVA) to assess how agent gaze aversion behaviors affected the dependent variable for each task. Participant gender was included as a covariate to control for gender differences. Question ID was included as a covariate to control for learning effects. Planned comparisons were carried out as *apriori* contrast tests using Scheffé's method.

Hypothesis 1 – Our analysis supported this hypothesis. The time given to the virtual agent before interrupting was significantly higher when the agent used proper gaze aversion with good timing rather than bad timing, $F(1, 110) = 5.06, p = .027$, or with no gaze aversion at all, $F(1, 110) = 12.71, p < .001$.

Hypothesis 2 – Our analysis did not support this hypothesis. Participants' ratings did not differ for virtual agents using proper gaze aversion over agents using gaze aversion with bad timing, $F(1, 110) = 0.0004, p = .98$, or with no gaze aversion at all, $F(1, 110) = 0.002, p = .97$.

Hypothesis 3 – Our analysis supported this hypothesis. Virtual agents using gaze aversions with good timing elicited significantly more disclosure from participants than when their gaze aversions were badly timed, $F(1, 110) = 4.48, p = .037$, or when they used no gaze aversion, $F(1, 110) = 4.25, p = .042$.

Hypothesis 4 – Our analysis partially supported this hypothesis. The time given to the virtual agent during its pause before interrupting was marginally higher when the agent used properly-timed gaze aversion than when its gaze aversions were badly timed, $F(1, 110) = 3.64, p = .059$, and significantly higher than when it did not use gaze aversion at all, $F(1, 110) = 9.48, p = .003$. All of our primary results are illustrated in Figure 4.

5 Discussion

Virtual agents that displayed gaze aversion behaviors generated by our controller were partially successful in achieving the cognitive conversational function of gaze aversions. As shown in Task 1, virtual agents successfully used gaze aversion to indicate that they were engaged in a form of cognitive processing with a response forthcoming and thus delayed interruptions by a human interlocutor. However, as shown in Task 2, using gaze aversions before responses did not affect how thoughtful participants thought those responses were. A possible explanation for this result is that while participants respond *behaviorally* to an agent using gaze aversion to achieve conversational functions, these cues fail to elicit explicit attributions of thought when participants are asked to reflect on the interaction afterwards.

Virtual agents displaying gaze aversion behaviors generated by our controller were successful in eliciting more disclosure from participants. Measurements of the breadth of participants' responses in Task 3 show that participants disclosed more when the virtual agent periodically looked away from them with appropriate timings than when the agent did not look away or looked away at inappropriate times.

Finally, virtual agents displaying gaze aversion behaviors generated by our controller were successfully able to regulate conversational turn-taking. By averting their gaze at the appropriate time in Task 4, virtual agents more effectively held the conversational floor than when they used gaze aversion at inappropriate times or not at all.

Designers must consider gaze aversion as more than "lack of eye contact" and instead as a powerful cue that can achieve conversational goals. If the goal is to elicit disclosure from a human, the virtual agent should use gaze aversion to regulate the intimacy of the conversation. When virtual agents need to pause in their speech, e.g., to process information or plan their next utterance, gaze aversion is an effective strategy to hold the conversational floor and indicate to the human that a new utterance is forthcoming. This idea is similar to work by Shiwa et al. [19], which showed that robots can use conversational fillers to successfully alleviate users' negative perceptions to long system response times.

5.1 Limitations and Future Work

Although the gaze aversion strategies presented in this paper are closely tied to the conversational states of speaking and listening, future work should concentrate on connecting gaze aversions more closely with the content and structure of speech. Previous research by Cassell et al. [7] identified relationships between gaze behavior and information structure of utterances, specifically the theme and rheme of sentences. Integrating these findings with our gaze aversion controller would be a useful extension to the current work.

A limiting assumption of our controller is that the gaze aversion behaviors generated are stable over time, while in reality these behaviors likely change over the course of a conversation due to increasing familiarity with the interlocutor, changing emotions and level of comfort, and so on. In future work, we plan to develop models of gaze that dynamically adjust gaze aversion strategies over time as well as retain the significant edge cases of behavior that are potentially lost by our current statistical approach of collapsing data into averaged distributions.

Another limitation of our work is that the gaze aversion behaviors of the virtual agent do not take into account the gaze behavior of the user. By tracking the gaze of the user, a virtual agent could more effectively modulate the amount of mutual gaze exhibited in the interaction in order to better regulate intimacy. It could also assess whether it has the attention of the user before attempting nonverbal behaviors that have an associated conversational goal. Previous research has explored the development of interactive gaze models for virtual agents, such as work by Bee et al. [6]. Future work might develop interactive models of gaze *aversion* that more dynamically employ aversion behaviors in human-agent conversations.

6 Conclusion

Gaze aversions are commonly associated with negative social outcomes, including discomfort, inattention, and deceit, but in reality they serve a number of important positive conversational functions, including cognitive, intimacy-modulating, and turn-taking functions. In this paper, we demonstrated how to enable virtual agents to use gaze aversions to achieve these functions in conversations with people. We presented an analysis of human dyadic conversations that informed the development of a gaze aversion controller that can automatically plan and execute appropriately timed gaze aversions for virtual agents. We also presented an experiment that evaluated the gaze aversion behaviors generated by the controller for their effectiveness in achieving positive conversational functions. The experiment demonstrated that virtual agents using gaze aversions generated by our controller were perceived as thinking, elicited more disclosure from human interlocutors, and effectively managed turn-taking. Our findings suggest that gaze aversion is a powerful conversational cue that designers should draw on in order to create effective and natural human-agent interactions.

Acknowledgments. This research was supported by National Science Foundation award 1017952. We would like to thank Faisal Khan, Brandi Hefty, Allison Terrell, Danielle Albers, Irene Rae, and Brandon Smith for their help in data collection, video coding, and providing voices for our virtual agents.

References

1. Abele, A.: Functions of gaze in social interaction: Communication and monitoring. Journal of Nonverbal Behavior 10(2), 83–101 (1986)
2. Andrist, S., Pejsa, T., Mutlu, B., Gleicher, M.: Designing effective gaze mechanisms for virtual agents. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, pp. 705–714. ACM (2012)
3. Andrist, S., Pejsa, T., Mutlu, B., Gleicher, M.: A head-eye coordination model for animating gaze shifts of virtual characters. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, pp. 4:1–4:6. ACM (2012)
4. Argyle, M., Cook, M.: Gaze and mutual gaze. Cambridge University Press, Cambridge (1976)
5. Beattie, G.W.: A further investigation of the cognitive interference hypothesis of gaze patterns during conversation. British Journal of Social Psychology 20(4), 243–248 (1981)
6. Bee, N., Wagner, J., André, E., Vogt, T., Charles, F., Pizzi, D., Cavazza, M.: Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, pp. 9:1–9:8. ACM (2010)
7. Cassell, J., Torres, O., Prevost, S.: Turn taking vs. discourse structure: How best to model multimodal conversation. In: Machine Conversations, pp. 143–154 (1999)
8. Doherty-Sneddon, G., Phelps, F.: Gaze aversion: A response to cognitive or social difficulty? Memory & Cognition 33(4), 727–733 (2005)
9. Ehrlichman, H., Micic, D.: Why do people move their eyes when they think? Current Directions in Psychological Science 21(2), 96–100 (2012)
10. Exline, R., Gray, D., Schuette, D.: Visual behavior in a dyad as affected by interview content and sex of respondent. Journal of Personality and Social Psychology 95, 201–209 (1965)
11. Garau, M., Slater, M., Bee, S., Sasse, M.: The impact of eye gaze on communication using humanoid avatars. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 309–316. ACM (2001)
12. Glenberg, A.M., Schroeder, J.L., Robertson, D.A.: Averting the gaze disengages the environment and facilitates remembering. Memory & Cognition 26(4), 651–658 (1998)
13. Heylen, D., Van Es, I., Van Dijk, E., Nijholt, A., van Kuppevelt, J., Dybkjaer, L., Bernsen, N.: Experimenting with the gaze of a conversational agent. In: Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, pp. 93–100. Kluwer Academic Publishers (2005)
14. Kendon, A.: Some functions of gaze-direction in social interaction. Acta Psychologica 26(1), 22–63 (1967)
15. Lee, S., Badler, J., Badler, N.: Eyes alive. ACM Transactions on Graphics (TOG), 637–644 (2002)
16. Moon, Y.: Intimate exchanges: Using computers to elicit self-disclosure from consumers. Journal of Consumer Research 26(4), 323–339 (2000)

17. Pelachaud, C., Bilvi, M.: Modelling gaze behavior for conversational agents. In: Rist, T., Aylett, R.S., Ballin, D., Rickel, J. (eds.) IVA 2003. LNCS (LNAI), vol. 2792, pp. 93–100. Springer, Heidelberg (2003)
18. Phelps, F.G., Doherty-Sneddon, G., Warnock, H.: Helping children think: Gaze aversion and teaching. British Journal of Developmental Psychology 24(3), 577–588 (2006)
19. Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., Hagita, N.: How quickly should communication robots respond? In: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, pp. 153–160. IEEE (2008)
20. Wang, N., Gratch, J.: Don't just stare at me! In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2010)
21. Wiemann, J.M., Knapp, M.L.: Turn-taking in conversations. Journal of Communication 25(2), 75–92 (1975)

From a User-created Corpus of Virtual Agent's Non-verbal Behavior to a Computational Model of Interpersonal Attitudes

Brian Ravenet, Magalie Ochs, and Catherine Pelachaud

Laboratoire Traitement et Communication de l'Information, CNRS-LTCI
Télécom Paristech - 37/39 rue Dareau
75014 Paris, France
{brian.ravenet,magalie.ochs,catherine.pelachaud}@telecom-paristech.fr

Abstract. Human's non-verbal behavior may convey different meanings. They can reflect one's emotional states, communicative intentions but also his social relations with someone else, *i.e.* his interpersonal attitude. In order to determine the non-verbal behavior that a virtual agent should display to convey particular interpersonal attitudes, we have collected a corpus of virtual agent's non-verbal behavior directly created by users. Based on the analysis of the corpus, we propose a Bayesian model to automatically compute the virtual agent's non-verbal behavior conveying interpersonal attitudes.

1 Introduction

Nowadays, virtual agents are more and more used to endowed particular social roles during human-machine interaction, such as actors in video games [1], assistants [2], or tutors [3]. To perform successfully social roles, virtual agents should be able to convey different interpersonal attitudes, such as being friendly, hostile or dominant. Interpersonal attitude is an "affective style that can be naturally or strategically employed in an interaction with a person or a group of persons" [4]. One's interpersonal attitude is expressed through verbal but also non-verbal behavior [5]. For instance, a more dominant person tends to do wider gestures [6].

In this paper, we aim at endowing virtual agents with the capacity to express different interpersonal attitudes through their non-verbal behavior. In particular, we focus on the type of behavior the virtual agent should display to communicate a given attitude. In this paper, we do not focus on the strategy employed by the agent. In order to identify how a virtual agent may convey interpersonal stances through its non-verbal behavior, we have collected a corpus of virtual agent's non-verbal behavior directly created by users. Based on the collected corpus, we have developed a Bayesian network to compute the virtual agent's non-verbal behavior depending on its interpersonal attitude. This paper is organized as follows. In Sect. 2, we present the theoretical background highlighting how interpersonal attitude is expressed through human's non-verbal behavior.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 263–274, 2013.

In Sect. 3, existing virtual agents endowed with the capabilities to display social non-verbal behavior are presented. In Sect. 4, we describe the platform we have developed to collect a corpus, directly created by users, of virtual agent's non-verbal behavior conveying different attitudes. In Sect. 6, we explain a Bayesian model to automatically generate virtual agent's non-verbal behavior depending on its interpersonal attitude. We conclude in Sect. 7.

2 Theoretical Background on Interpersonal Attitudes

A common way to represent interpersonal attitudes or relations is to follow the Argyle's representation [7,8]. This representation is composed of two dimensions: *dominance* (also called status or agency) and *liking* (also called affiliation or communion). This is sometimes referred to as the *Interpersonal Circumplex* (Fig. 1). The axes are labeled with the dominant, submissive, friendly and hostile terms. An interpersonal attitude is then represented by a point in the circumplex, *i.e.* a value on the dominance axis and on the liking axis. Several researchers in Human and Social Sciences have explored how interpersonal attitudes are conveyed through non-verbal behavior. They show that interpersonal attitude is mainly conveyed by several modalities from the upper part of the body like facial expression, gestures or gaze [6,9]. In [6], they report that people with a more dominant attitude do more and wider gestures. A more submissive attitude leads to less expressive faces [9] and a friendly attitude to more smiles [5]. Head orientation is also different depending on the attitude; a more dominant or friendly person will use more upward orientation [6,9]. Mutual gaze conveys an attitude of dominance or friendliness as shown in [5]. Moreover some works highlighted the influence of the gender of the interactants on one's non-verbal behavior [10,11]. Based on these results, we have constructed a platform to collect data on the non-verbal agent's behavior associated to an interpersonal attitude. Before describing this platform, in the next section, we present existing works in virtual agents.

3 Related Works on Virtual Agents

In this section, we present the existing works on virtual agents with social abilities. They are usually referred to as *relational agents*. Laura[12] is one of the first relational agents. She is a fitness coach users had to interact with on a long-term duration. The relations evolve in the same way for every user. The more a user interacts with her, the friendlier is her non-verbal behavior (proximity, mutual gaze, higher hand gestures frequency and head nods). Eva [13] is expressing emotions and is also endowed with a social relations model, following the *Interpersonal Circumplex*. Social relations do not affect directly the behavior of Eva, but the emotions generated within the agent. In the project Demeanour [14], the authors also use the Interpersonal Circumplex to represent social relations between two agents. Gaze, gestures and postures of the agents change depending on their social relations. The behavior model is based on a psychology model and on informal observations of people. In [15], a study is conducted to assess

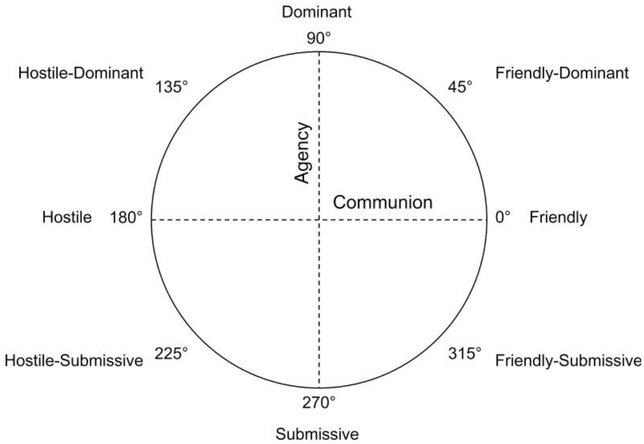


Fig. 1. Example of an Interpersonal Circumplex [8]

how users perceive attitude (hostile, friendly) and extraversion during the first seconds of an encounter. The differences in behavior (smile, gaze and proximity) are modeled from the literature in Human and Social sciences. In [16], they build a model of non-verbal behavior (gaze, posture and proximity) depending on the conversation role (speaker, addressee, side participant or bystander), the communicative act and the relations between the characters. This model was built by observing actors playing a scenario (the *Gunslinger*) involving the different roles and knowing the interpersonal relations between the characters. While several models of interpersonal attitudes have been proposed, they do not study specifically gender differences in conveying interpersonal attitudes. In our research, we aim to go beyond by considering a larger set of behaviors and gender differences.

We explore four different social attitudes, *Dominant*, *Submissive*, *Friendly* and *Hostile*. We are observing the following non-verbal cues: facial expressions, gazes, head shifts, head movements and arm gestures. Moreover, we also consider two expressivity parameters for the arm gestures, *Spatial* (linked to gesture amplitude) and *Power* (linked to gesture force). We are using a user perceptive approach to collect data about how interpersonal attitudes are encoded through these cues. This data, collected directly from participants in an online application we have developed, is used to generate a computational model of agent's non-verbal behavior depending on its interpersonal attitude.

4 GenAttitude: A Platform to Collect Virtual Agent's Non-verbal Behavior

As presented in Sect. 2, one's non-verbal behavior may convey particular interpersonal attitude. In order to identify the non-verbal behavior a virtual agent should perform to express social attitudes, we propose a user perceptive approach consisting in asking directly the user to configure the non-verbal behavior of an agent with particular social attitudes. This approach has already been used in [17] to gather data about the features of different types of smiles (amused, polite and embarrassed) and to generate a decision tree for these types of smiles. Whereas in [17], the focus was in the signal of smiles, in the presented work we explore the multimodal behavior of a virtual agent.

Illocutionary Acts. In our work, the attitudes are considered in the context of an interaction. We focus on the role of the locutor[1]. We base our work on the classification of illocutionary acts (or speech acts) proposed by Searle [18]. This classification considers five different categories: assertive, directive, commissive, expressive and declarative. Since research showed that gestures are tightly connected to the content of speech [19,20] and to cover the five categories of illocutionary acts, we consider a set of six illocutionary acts: *Inform* as an assertive, *Ask* as a directive, *Accept* as a commissive, *Agree* and *Disagree* as expressive and *Deny* as a declarative.

GenAttitude platform. We have developed *GenAttitude* using Flash technology and the virtual agent platform Greta [21]. Flash makes it easy to create an online graphical application and the architecture of Greta allowed us to develop additional modules we needed in order to generate animations. The interface of *GenAttitude*, illustrated in Fig. 2 is inspired by the one proposed in [17]. On the top of the screen, there is a description of the current task. It explains what attitude and what illocutionary act the user has to configure. There are four different possible social attitudes (randomly selected): Dominant, Submissive, Friendly and Hostile. On the right side, there are the different parameters for the non-verbal behavior. The user can change the value of these parameters using radio buttons. The list of parameters as well as their possible values has been designed following the research in Human and Social Sciences presented in Sect. 2. For each parameter, the user can select a value from a discrete set of values. The parameters are the following:

1. The type of facial expression: smile, frown or neutral
2. Gesture: none, head movement only, arm movement or both
3. Amplitude of gesture: small, medium or large
4. Power of gesture: small, medium or strong
5. Head position: straight, up, down or tilt on the side
6. Gaze: gaze at or gaze away

The types of head movements and of arm gestures cannot be modified by the participants. The head movements have been defined to be consistent with the communicative act. When the act is either *Deny* or *Disagree*, the head movement available is a shake. It is a nod otherwise. Concerning the arm gestures, we have chosen a neutral gesture (not specific to a communicative act). It is a two hands

[1] Note that the interpersonal attitude may also be reflected by the non-verbal behavior of a virtual listener.

palm open facing up placed in front of the agent. This way, the same gesture is used for each communicative act. When there is no gesture, both hands rest on the agent's belly. Also, it is not possible to select a value for the *Spatial* and *Power* parameters if is not activated (value *gesture only* or *both* for the *Gesture* parameter). In order to avoid the experiment to be biased by a default configuration, the parameters are randomly pre-selected. On the left side, a video plays in loop the animation resulting from the configuration of parameters. The video is changed each time a value for a parameter is changed. Finally, on the bottom of the screen, there is a Likert scale for the user to rate how satisfied he is with the resulting animation.

Procedure. Each participant has used *GenAttitude* to indicate the non-verbal behavior of a virtual agent in 6 different situations. Each situation corresponds to a particular social attitude, a particular communicative act and a male or female agent. To avoid an effect of the order of the presentation on the results, the order of the communicative act, the attitude and the gender presented to the participant was counter-balanced.

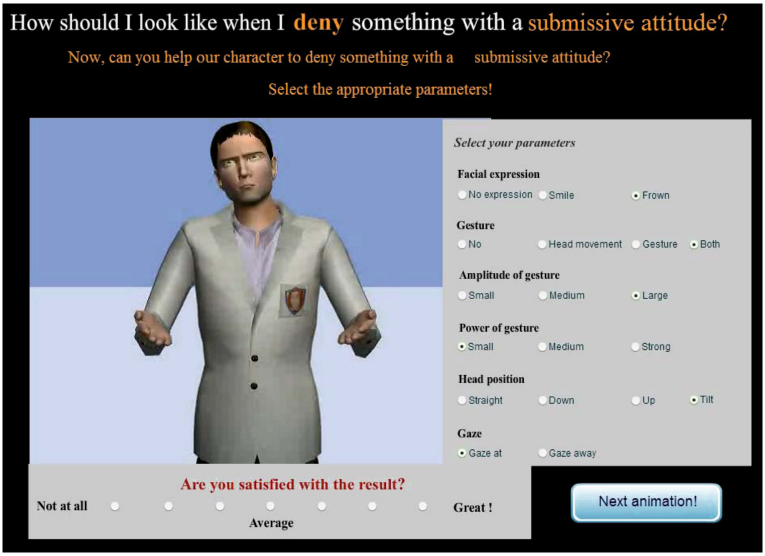


Fig. 2. Screenshot of the Interface of GenAttitude

Generating the videos. Due to the number of parameters and possible values, we have generated 1440 videos. For this purpose, we developed an additional module in the platform Greta. This module takes an FML [22] file as input, generates all combinations of signals (facial expressions, gestures, head movements and parameters) associated with the intentions specified in the FML file and capture a video of the resulting animation for each combination.

Before making *GenAttitude* available online, 6 pre-tests have been performed in order to ensure the interface was clear and that the tasks were well understood.

5 Statistical Results

5.1 Description

From this experiment, we collected 925 entries from 170 participants (311 entries from female participants) during a week. Each entry corresponds to an illocutionary act, an attitude, a gender and the set of values for each parameter. We had participants from different countries but most of them were from France, aged from 17 to 62 (M = 28.8, SD = 2.8). In average, the participants were globally satisfied with the created animation: the satisfaction of the participants is in average 5.35 (on a likert scale of 7 points).

5.2 Influence of the Attitude

In order to analyze the collected corpus and to study the correlation between the attitudes, acts, gender and the non-verbal parameters, we have computed Cramer's V χ^2 tests. The χ^2 test compares the distribution frequencies of the outcomes of two variables with a theoretical distribution where the variables are independent. The result of the χ^2 test indicates if two variables are dependent and the Cramer's V test provides the strength of this dependency. The results show significant correlations between the attitude and the *Facial expression* parameter ($\chi^2(6) = 451.65$, $p < 0.005$; Cramer's V= 0.49), the *Gestures* parameter ($\chi^2(9) = 24.0$, $p < 0.005$; Cramer's V= 0.09), the *Power* parameter ($\chi^2(6) = 69.9$, $p < 0.005$; Cramer's V= 0.26), the *Spatial* parameter ($\chi^2(9) = 24.0$, $p < 0.005$; Cramer's V= 0.09), the *Head position* parameter ($\chi^2(9) = 167.3$, $p < 0.005$; Cramer's V= 0.25) and the *Gaze* parameter ($\chi^2(9) = 81.69$, $p < 0.005$; Cramer's V= 0.29). In other words, all the non-verbal parameters are implied in the expression of social attitude.

We also looked at the correlations between the illocutionary act and the parameters. Only the *Facial expression* ($\chi^2(10) = 24.9$, $p < 0.05$; Cramer's V= 0.11) and the *Gestures* ($\chi^2(15) = 89.39$, $p < 0.005$; Cramer's V= 0.18) showed significant correlations. Indeed, positive facial expression was generally selected more for positive illocutionary acts (*Agree*, *Accept*) while the negative one was generally selected more for negative acts (*Disagree*, *Deny*). Regarding the gestures, head movements were more generally selected for negative and positive acts (respectively *Disagree*, *Deny* and *Agree*, *Accept*) and the gesture only (no head movement) were more often chosen for a neutral act (*Ask*, *Inform*). This could be explained by the fact that we introduced a shake for negative acts and a nod for positive and neutral acts.

Moreover, we have looked for correlations between the gender of the agents and the gender of the users on the selected parameters. The results did not reveal any significant correlations. In other words, the gender of the agent, or of the user, does not seem to have an impact on the selected parameters.

Table 1. Contingency table of the non-verbal parameter and the attitudes

Parameter	Value	Dominant	Friendly	Hostile	Submissive
Facial Expression	negative	**44,64%**	9,40%	**79,20%**	16,81%
	neutral	29,18%	8,55%	12,39%	**46,98%**
	positive	26,18%	**82,05%**	8,41%	36,21%
Gestures	gesture only	27,47%	**30,77%**	24,34%	20,69%
	head only	29,18%	28,21%	26,99%	**34,91%**
	both	**33,05%**	29,49%	**31,86%**	23,28%
	none	10,30%	11,54%	16,81%	21,12%
Power	normal	24,82%	39,01%	23,62%	26,47%
	small	26,95%	**41,84%**	19,69%	**53,92%**
	strong	**48,23%**	19,15%	**56,69%**	19,61%
Spatial	normal	29,08%	**35,46%**	22,05%	17,65%
	small	31,21%	29,08%	29,13%	**56,86%**
	large	**39,72%**	**35,46%**	**48,82%**	25,49%
Head position	down	22,32%	19,23%	**30,53%**	**58,62%**
	straight	30,47%	27,78%	23,45%	13,36%
	tilt	12,45%	**39,32%**	24,78%	18,10%
	up	**34,76%**	13,68%	21,24%	9,91%
Gaze	gaze away	15,88%	17,52%	28,32%	49,14%
	gaze at	**84,12%**	**82,48%**	**71,68%**	**50,86%**

5.3 Discussion

The statistical analysis presented in the previous section shows a significant relationship between attitudes and the selected parameters of the virtual agent's non-verbal behavior. We discuss these results in more details in the following by looking at the most selected values for each attitude.

Dominant. The dominant attitude is mainly characterized by a negative facial expression, the presence of head movements and arm gestures, the absence of gaze avoidance, and an upward position of the head. The gestures are characterized with a *large* spatial parameter and a *strong* power parameter.

Submissive. A submissive attitude is mainly characterized by a neutral facial expression, head movements only (no arm gestures), and an downward position of the head. For the gaze avoidance there is a little preference for the *gaze at* value. Both spatial and power parameters receive a *small* value when the gestures are activated.

Friendly. This attitude is characterized by gestures only, a positive facial expression, a tilt of the head on the side and no gaze avoidance. For the spatial parameter, the *normal* value and the *large* value are the most selected and for the power parameter, the *small* value is the most selected.

Hostile. Similar as the dominant attitude, hostile is characterized using the negative facial expression, both head movements and arm gestures and no gaze avoidance. The same parameters for gesture is selected (*large* and *strong*). However, the head position is *down*.

These results, described Table 1, show close similarities between the way people consider non-verbal behaviors in a human and in an agent. Indeed, people expect from a dominant person more wide gestures or an upward oriented head [6], from a submissive person less expressive faces [6] and from a friendly one more smiles [5]. However, our results contains some differences with the literature. No influence of the gender of the agent or of the user was significant. Also, more gaze avoidance from an hostile or submissive agent was expected but not found.

Based on these results, in the next section, we propose a Bayesian model to automatically generate non-verbal behavior of virtual agent given its social attitude.

6 Bayesian Network Based Model of Virtual Agent's Non-verbal Behavior

Some computational models of virtual agent's non-verbal behavior are based on a Bayesian network [23,24]. Bayesian networks are directed acyclic graphs where the nodes represent the variables and the edges between two nodes represent the conditional dependencies between two variables [25]. In [23], the parameters of the non-verbal behavior as well as the context of the interaction are represented by the nodes of the network. As input nodes to describe the context, the previous gesture, the discourse and verbal context and some features of the topic of conversation (for instance position and shape of objects) are used. As output nodes, parameters for the gestures (for instance hand orientation, movement and hand shape) are specified. In our work, we propose to use a similar representation based on a Bayesian network to compute the non-verbal behavior of a virtual agent given its social attitude. We present in details the structure of the network in the next section.

6.1 Structure of the Network

As input nodes for our model, we take the *Interpersonal Attitude* and the *Illocutionary Act*. As output nodes, we consider the non-verbal behavior parameters explored with the *GenAttitude* platform (Section 4). We placed an oriented arc from the input nodes to the output nodes depending on the correlation between input and output nodes. The correlations correspond to those identified in Sect. 5. The result is shown in Fig. 3.

6.2 Parameters of the Bayesian Network

Learning the parameters of a Bayesian network consists in learning the conditional probability distribution for each node. To compute the parameters of the model, we use the data of the collected corpus (Section 5). Indeed, the conditional probabilities can be easily computed from the 925 descriptions of non-verbal behavior associated to the attitudes.

In order to consider the satisfaction indicated by the user on the created non-verbal behavior (Section 4), we have done *oversampling* to give a higher weight to the entries with a high level of satisfaction: each entry has been duplicated n times, where n is the level of satisfaction associated with this entry. So, a non-verbal behavior describing an attitude with a level of satisfaction of 7 is duplicated 7 times whereas a non-verbal behavior with a level of satisfaction of 1 is not duplicated. The resulting data set is composed of 4947 entries.

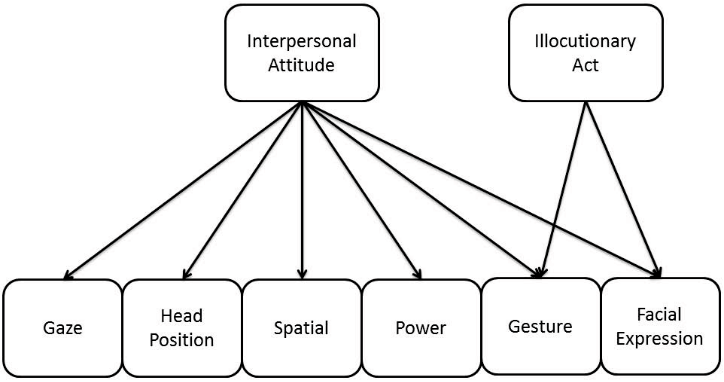


Fig. 3. The Bayesian Network

Outcome of the model. The Bayesian network enables two ways of using our data in the context of interpersonal attitudes of virtual agent. From this network, we can choose an attitude and a communicative act and obtain non-verbal behaviors with different probabilities to be perceived as corresponding to this attitude. This introduces variability in the generation of non-verbal behaviors. Or, from a captured non-verbal behavior (described with our parameters), we can infer which attitude might be conveyed by this behavior. Moreover, this model can be easily extended in the future with new parameters. It could be mixed with the model from [23] for instance. This would allow the model to be able to decide if gestures are activated depending on the communicative act and attitude, and then to decide which shape (hand orientation and hand shape) to use. Including the study in [24], it would extend the model with complex facial expressions.

6.3 Illustrations

In this section, we propose to illustrate the outputs of the proposed model on a concrete example. The agent has to perform an *Inform* illocutionary act with a dominant attitude. Following our Bayesian network, the combination which has the highest probability is a neutral face, gesture only (no head movements), large and strong gesture, an up-oriented head and no gaze avoidance. For the same illocutionary act but with a submissive attitude, the combination with the highest probability is a positive facial expression, no gesture at all and a down-oriented head. The non-verbal behaviors for these attitudes along with other variants are illustrated Fig. 4.

Fig. 4. Screenshot of a virtual agent expressing a social attitude: dominant attitude (top) and submissive attitude (bottom)

7 Conclusion

In this paper, we presented how we developed an online platform in order to collect data from participants about the non-verbal behavior of a virtual agent depending on its interpersonal attitudes. We analyzed these data in order to understand how a virtual agent endowed with interpersonal attitude should behave as a speaker using different illocutionary acts. Then we proposed a Bayesian network built upon these data in order to create a computational model of non-verbal behavior depending on the interpersonal attitude of the agent and on the illocutionary act.

The proposed model could be improved in several ways. For instance, the discrete values on the interface could be replaced by continous values that the user could manipulate through sliders. The description of the task to the user could also be more precise and integrate other elements from the context, such as the agent's emotions. In future works, we aim at analyzing differences in the expression of the attitude depending if the agent is speaking or listening and who his interlocutor is. In the presented study, we have focused on the interpersonal attitude expressed through a communicative intention. However, the attitude also emerges from the interaction depending, for instance, on the reactions of the agent to the user's behaviors. For this purpose, we aim at considering how the global non-verbal behavior of the agent may convey particular attitudes on the overall interaction. Then, in the Interpersonal Circumplex, we only considered the four edges of its two axes as potential attitudes. It would be interesting to expend the possible attitudes, by letting the users rate the value on each axis for instance. Finally, non-verbal parameters may not be independent with each other. We have not analyzed this in this work.

Acknowledgment. This research has been supported by the European projects NoE SSPNet and IP REVERIE.

References

1. Isbister, K.: Better Game Characters by Design: A Psychological Approach (The Morgan Kaufmann Series in Interactive 3D Technology). Morgan Kaufmann (June 2006)
2. Schmeil, A., Broll, W.: Mara: an augmented personal assistant and companion. In: ACM SIGGRAPH 2006 Sketches, SIGGRAPH 2006. ACM, New York (2006)
3. Bernardini, S., Porayska-Pomsta, K., Smith, T.J., Avramides, K.: Building autonomous social partners for autistic children. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 46–52. Springer, Heidelberg (2012)
4. Scherer, K.: What are emotions? and how can they be measured? Social Science Information (2005)
5. Burgoon, J.K., Buller, D.B., Hale, J.L., de Turck, M.A.: Relational Messages Associated with Nonverbal Behaviors. Human Communication Research 10(3), 351–378 (1984)
6. Carney, D., Hall, J., LeBeau, L.: Beliefs about the nonverbal expression of social power. Journal of Nonverbal Behavior 29, 105–123 (2005)
7. Argyle, M.: Bodily Communication. University paperbacks, Methuen (1988)
8. Gurtman, M.B.: Exploring personality with the interpersonal circumplex. Social and Personality Psychology Compass 3(4), 601–619 (2009)
9. Burgoon, J.K., Le Poire, B.A.: Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality. Communication Monographs 66(2), 105–124 (1999)
10. Briton, N.J., Hall, J.A.: Beliefs about female and male nonverbal communication. Sex Roles 32, 79–90 (1995)
11. Hess, U., Thibault, P.: Why the same expression not mean the same when shown on different faces or seen by different people. In: Tao, J., Tan, T. (eds.) Affective Information Processing, pp. 145–158. Springer, London (2009)
12. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Trans. Comput.-Hum. Interact. 12(2), 293–327 (2005)

13. Kasap, Z., Moussa, M.B., Chaudhuri, P., Magnenat-Thalmann, N.: Making them remember emotional virtual characters with memory. IEEE Computer Graphics and Applications, 20–29 (March 2009)
14. Gillies, M., Crabtree, I., Ballin, D.: Customisation and context for expressive behaviour in the broadband world. BT Technology Journal 22(2), 7–17 (2004)
15. Cafaro, A., Vilhjálmsson, H.H., Bickmore, T., Heylen, D., Jóhannsdóttir, K.R., Valgarðsson, G.S.: First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 67–80. Springer, Heidelberg (2012)
16. Lee, J., Marsella, S.: Modeling side participants and bystanders: The importance of being a laugh track. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 240–247. Springer, Heidelberg (2011)
17. Ochs, M., Niewiadomski, R., Brunet, P., Pelachaud, C.: Smiling virtual agent in social context. Cognitive Processing 13(2), 519–532 (2012)
18. Searle, J.R.: A classification of illocutionary acts. Language in Society 5(3), 1–23 (1976)
19. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press (1992)
20. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press (2004)
21. Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive eca system. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009, vol. 2, pp. 1399–1400. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2009)
22. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
23. Bergmann, K.: The Production of Co-Speech Iconic Gestures: Empirical Study and Computational Simulation with Virtual Agents. PhD thesis, Bielefeld University (2012)
24. Pelachaud, C., Poggi, I.: Subtleties of facial expressions in embodied agents. The Journal of Visualization and Computer Animation 13(5), 301–312 (2002)
25. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)

Explaining the Variability of Human Nonverbal Behaviors in Face-to-Face Interaction

Lixing Huang and Jonathan Gratch

Institute for Creative Technologies, University of Southern California
lxhuang1984@gmail.com, gratch@ict.usc.edu

Abstract. Modeling human nonverbal behaviors is a key factor in creating a successful virtual human system. This is a very challenging problem because human nonverbal behaviors inherently contain a lot of variability. The variability comes from many possible sources, such as the participant's interactional goal, conversational roles, personality and emotions and so on, making the analysis of the variability hard. Such analysis is even harder in face-to-face interactions since these factors can interact both within and across the participants (i.e. speaker and listener). In this paper, we introduce our initial efforts in analying the variability of human nonverbal behaviors in face-to-face interactions. Specifically, by exploring the Parasocial Consensus Sampling (PCS) framework [13], we show personality has significant influences on listener backchannel feedback and clearly demonstrate how it affects backchannel feedback. Moreover, we suggest that PCS framework provides a general and effective approach to analyze the variability of human nonverbal behaviors, which would be difficult to perform by using the traditional face-to-face interaction data.

Keywords: Parasocial Interaction, Nonverbal Behaviors, Variability, Personality.

1 Introduction and Background

Today, we have seen a few virtual human systems with natural and realistic behaviors in interactive scenarios such as training [1], health care [2] and education [3]. One of the key factors that makes these systems successful is that virtual humans can provide contingent and appropriate feedback to their interactional partners in real time. There have been many efforts in building nonverbal behavior models to predict when and how the virtual human should respond to his interactive partner accordingly. A lot of progress has been made. Originally, researchers depend on the findings from the social psychology literature. They [4] [5] [6] [7] usually derived a set of rules from the literature to drive the virtual human's behavior. However, such descriptive rules are more helpful as general theoretical points than to directly drive a virtual human's behavior as they typically describe general findings and do not precisely characterize the specific circumstance and timing information for when such behaviors should

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 275–284, 2013.

be employed. Recently, researchers [8] [9] [10] start to explore more advanced machine learning techniques to learn behavior models from large amounts of annotated human behavior data. Such approach usually generates quantitative models which can directly be used to drive the virtual human's behavior. More importantly, by changing the dataset that the algorithm learns from, it is feasible to train models that are in line with the context where the virtual human will be applied. However, there are still challenging and unsolved problems.

First, most of the virtual humans can only provide generic feedback. Bavelas et al. [11] proposed that nonverbal feedback can be classified into two classes. One is generic feedback, which is not closely connected to what is being said. Such generic behaviors don't convey any specific meanings, and would be appropriate in different scenarios. The other one is specific feedback, which is tied to a deeper understanding of, and reaction to, the personal relevance of what is being said. Such specific behaviors usually depend not only on the understanding of the semantic meanings but also on our own role and participatory goals, which may change as the conversation unfolds [15]. Currently, most of the virtual human systems address the first type of behavior. For example, the Rapport Agent [7] relies on low level analysis of the nonverbal signals of the human speaker and provides contingent feedback, such as head nod, accordingly. In order to apply the virtual human technology in more complex secnarios, it is inevitable that we need start building models for specific nonverbal behaviors. Second, virtual human needs not only respond to his interactional partner but also be able to reflect his own emotion, personality, and interactional goal. For example, the recent SEMAINE project built the Sensitive Artificial Listener [12]. By exhibiting different styles of audiovisual listener feedback, the listener is able to express four different personalities. However, it is still difficult to perform formal analysis on how personalities can affect nonverbal behaviors in face-to-face interactions.

Because of these reasons, nonverbal behaviors inherently contain a lot of variability. They are affected by many internal factors, such as emotion and personality, and external factors, such as the presence of others and others' responses. These factors interact both within and across participants - for example the emotions of one participant in a conversation can spill over and alter the behavior of other actors - making it difficult to isolate and model the variability of nonverbal behaviors. The problem is not insurmountable but it implies that we will have to collect large amounts of behavioral data. But the traditional way of recording face-to-face interaction data is very expensive and time-consuming. It usually takes months to recruit pairs of participants, followed by an extensive period of manually-annotating the resulting recordings.

To solve these problems, Huang et al. [13] proposed a new approach called Parasocial Consensus Sampling (PCS), where multiple independent participants experience the same social situation parasocially (i.e. act "as if" they were in a real dyadic interaction) in order to gain insight into the typicality (i.e. consensus view) of how individuals would behave within face-to-face interactions. Since multiple participants can now interact with the same social situation, usually pre-recorded videos (e.g. speaker video), we hold one side of the interaction consistent. This helps unpack the bidirectional causal influences that naturally occur in conversations. Moreover, by using pre-recorded speaker videos, we can dramatically increase the efficiency of the data collection process by having multiple participants interact with the same speaker simultaneously.

In this paper, we extend the original Parasocial Consensus Sampling work in two ways. First, we examine a more naturalistic approach (i.e. videotaping) to measure participants' behavior. In the original work, participants were guided to press a button whenever you feel like to respond. Although efficient, it has several limitations. For example, pressing a button demands an explicit conscious decision from a participant and it is difficult to measure multiple behaviors at the same time since pressing different buttons for different behaviors is likely to place too much cognitive load on the participants. In our study, we ask participants to interact with pre-recorded speaker videos and act as if they were in a real conversation (e.g. smile if they feel like smiling) and videotape the participants' nonverbal responses. Second, we take advantage of the efficiency of this framework to increase the number of participants. Our goal is to analyze how participant's personality affects their nonverbal behaviors in the interaction. As mentioned before, there are a lot of possible sources for the variability of human behavior. By exploring the PCS framework, it is possible to examine how each of these sources can affect human behavior independently (e.g. by assigning different interactional goals to different participants, we can investigate how interactional goal affects nonverbal behaviors). We examine personality first and use it as an example to demonstrate how PCS can help us tease apart the causalities.

The following section describes the data collection and data annotation process. Section 3 discusses the results. We conclude our work in Section 4.

2 Data Collection and Annotation

2.1 Data Collection

In the study, we recruited 28 participants via www.craigslist.com from the general Los Angeles area. Before beginning the study, the participants were required to read the instructions and ask questions about anything they do not understand. They were informed beforehand that they would be videotaped and instructed to pretend to show interest and create a sense of rapport with the

Table 1. The attributes of each coder we measured before they started interacting with the speakers parasocially

Big Five Personality Traits	Extroversion, Agreeableness, Conscientiousness, Neuroticism, Openness
Self-Consciousness	Self-directed, Other-directed
Parasocial Experience	Parasocial experience scale [18]
Other	Shyness, Self-monitoring, Gender

speaker in the video[1] by showing backchannel feedback such as head nod, head shake, and smile and so on. They first finished a 90-item personality inventory to measure their personality traits. Table 1 lists several individual traits that we are currently investigating. Next, they watched 8 speaker videos in sequence in a random order. Their nonverbal responses to the speakers were videotaped. At the end of the study, the participants were debriefed and each was paid 35 USD. Figure 1 shows an example of the parasocial interaction.

Fig. 1. An example of the parasocial interaction. The participant (right side) interacted with the speaker video (left side) parasocially, and her nonverbal behaviors were recorded by a camera. In this example, the speaker paused and tried to remember the details of the story he was supposed to tell. He had an embarrassed smile because it took him a relatively long time, and the participant smiled back, probably to reassure him. Although the participant was aware that the interaction was not real, she displayed such facial expressions seemingly automatically. We use the OKAO vision system from Omron Inc [14] to detect smiles, which can infer the level of smiling (continuous value from 0 to 100).

2.2 Results

At the end of the study, we collected parasocial responses from all 28 participants to each of the 8 speaker videos. Participants produced wide variety of behaviors including both generic feedback (e.g. head nod) and specific feedback (e.g. headshake and expressive facial expressions) [11]. The specific feedback is always triggered by certain events mentioned in the conversation. In this study, we chose head nods, headshakes and smiles, which are the mostly occurred behaviors in our dataset, as the target behaviors. We will leave other common behaviors, such as frowns, to the future work.

[1] The video set used in this study was previously collected and used for studying how humans create rapport during face-to-face interactions. Each video records a human speaker retold a story to another human listener. The dataset is available at `rapport.ict.usc.edu`.

2.3 Annotation

We are interested in three kinds of nonverbal behaviors: head nods, headshakes and smiles. A mix of manual and automatic annotation techniques were used to annotate these behaviors from the recorded videos. To annotate head nods and headshakes, we recruited native annotators from Amazon Mechanical Turk. To facilitate the annotation work, we developed a web-based annotation tool (as shown in Figure 2) that helps annotators go through the videos and annotate behaviors efficiently. Each annotator examined seven videos in sequence and only annotated a single type of behavior at one time. Each video was annotated by two independent annotators and each was paid 3 USD.

Smiles were annotated automatically using the OKAO vision system [14]. Briefly, it uses computer vision techniques to identify 16 facial landmarks. From this, it derives a variety of facial pose estimates including a smile intensity ranging from 0 (no smile) to 100 (full smile). By setting the threshold to 50, we can reliably determine whether the participant is smiling or not.

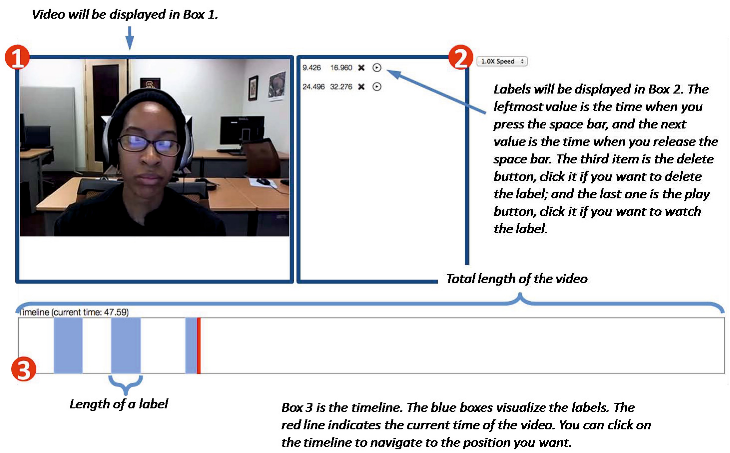


Fig. 2. This is the annotation interface. Coders press the space bar to start loading a video, and the loading progress will be shown in Component 1. After the video is loaded, coders press the space bar to start playing the video. At the beginning of the target behavior, coders press the space bar and hold it, and release the space bar when the target behavior ends. After finish labeling the video, coders can adjust the labels by dragging on their boundaries.

3 Data Analysis

For each of the 8 speaker videos, we aggregate the nonverbal behaviors (head nods, headshakes, and smiles) from all 28 participants to build the consensus view. Figure 3 shows an example of the consensus of head nod, headshake and smile. The peaks found in both the consensus of head nod and headshake are potential backchannel opportunities. But headshakes occur a lot less than head nods do, and they are usually associated with semantically negative events in the speech. There is a noticeable jump in the consensus of smile, where the speaker says the most dramatic part of the story. Interestingly, this phenomenon is observed in all 8 videos.

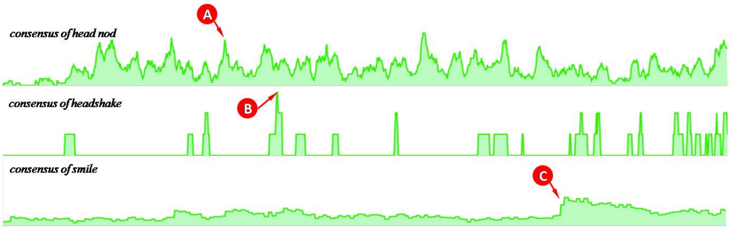


Fig. 3. An example illustrates the consensus of head nod (top), headshake (middle) and smile (bottom). The speaker video is from a sexual harassment training course. At point A, the speaker said "It's from Rick in accounting or Rick in legal or something, and [pause], he said 'oh, no' ..." and the nod is most likely to occur during the pause; at point B, the speaker said "and she says, 'you know, I gave him a ride once when his car broke down, now he won't leave me alone, it's been five weeks, I always get these emails, and e-cards, and he won't leave me alone'..." and the shake is most likely to occur when the speaker described the fact that Rick kept bothering the lady; at point C, the speaker said "and then she says 'oh, and next, I am gonna need a foot massage', and then she shuts the blinds..." and the smile is most likely to occur after mentioning the foot massage.

3.1 How Listener's Personality Traits Influence the Behavior

To examine the impact of personality on nonverbal behavior, we calculated personality scores from the 90-item personality inventory as described in Table 1. For each personality subscale (e.g. extroversion), we performed a median split and grouped participants into a high and low scoring group. For example, those scoring below the median for the scale of extroversion would be combined into an introverted group whereas those scoring above the median would be combined into an extroverted group. We then contrasted the consensus of the behaviors of these two partitions with Bonferroni Correction [16]. The results are shown as below.

The rows where significant differences are found are highlighted. Table 2 shows that listeners personality traits have significant influence on the number of head nods. The result is in line with previous research. For example, Chartrand and Bargh [19] found empathic individuals exhibit more mimicry behavior during the interaction to a greater extent than not empathic individuals; accordingly, our data suggests that extroversion, openness and consciousness all have similar influence on the number of head nods.

Besides head nods, we also examined headshake and smile. As Table 3 shows, headshake rarely happens during the interaction. In our data, we find that its

Table 2. Compare the average number of head nods between the *low_group* and *high_group* with respect to each attribute

Trait	Low	High	p-value
Extroversion	**39.7**	**64.1**	**p=0.002**
Agreeableness	65.0	65.7	p=0.92
Conscientiousness	52.0	68.2	p=0.026
Neuroticism	**68.9**	**31.8**	**p=0.005**
Openness	**33.3**	**59.3**	**p=0.003**
Self-consciousness	103	42.8	p=0.011
Other-consciousness	**31.6**	**81.4**	**p=0.0001**
Shyness	74.5	65.8	p=0.17
Self-monitor	77.0	58.0	p=0.016

occurrence is always associated with the semantically significant events (usually negative events) in the speech. Listeners personality traits dont have significant influence on headshake. However, smiling, although a kind of specific feedback, is significantly affected by the listeners personality traits (as shown in Table 4). This is similar to the results we found in literature. For example, Shiota et al. [17] showed that more extraverted, more conscientious, more agreeable, and less neurotic people are more likely to experience joy.

Table 3. Compare the average number of headshakes between the *low_group* and *high_group* with respect to each attribute

Trait	Low	High	p-value
Extroversion	2.2	1.2	p=0.017
Agreeableness	1.2	1.9	p=0.10
Conscientiousness	1.7	1.8	p=0.92
Neuroticism	2.0	1.2	p=0.12
Openness	0.375	1.375	p=0.027
Self-consciousness	0.7	3.5	p=0.21
Other-consciousness	2.25	0.46	p=0.11
Shyness	1.33	0.81	p=0.48
Self-monitor	2.41	1.21	p=0.058

3.2 How Speaker's Personality Traits Influence the Behavior

Each speaker video was watched by all participants, and we call the aggregation of their behaviors the "crowds' behavior". The crowds' behavior is different when interacting with different speaker videos. A natural follow-up question is whether or not the speaker's personality traits can influence the crowds' behavior. In a previous study [7], we measured each speaker's personalities. We compute the correlation coefficients between the speaker's personality measurements and the crowds' behavior (i.e. the number of head nods, the number of headshakes, and the amount of smiles). The results suggest that speaker's personality does not affect the listeners' head nods or smiles. Listeners' smiles are highly correlated with

Table 4. Compare the amount of smiles between the *low_group* and *high_group* with respect to each attribute. The number is calculated by dividing the duration of the listener's smiling by the duration of the whole interaction.

Trait	Low	High	p-value
Extroversion	0.13	0.21	p=0.018
Agreeableness	**0.15**	**0.26**	**p=0.001**
Conscientiousness	**0.13**	**0.40**	**p=0.0002**
Neuroticism	**0.40**	**0.11**	**p<0.0001**
Openness	0.14	0.12	p=0.3
Self-consciousness	**0.06**	**0.27**	**p=0.0001**
Other-consciousness	**0.08**	**0.04**	**p=0.006**
Shyness	0.29	0.16	p=0.03
Self-monitor	**0.02**	**0.21**	**p<0.0001**

Table 5. The correlation coefficients between speaker personality traits and the number of headshakes of crowds

Correlation Coefficient	Extroversion	Agreeableness	Neuroticism	Shyness
Number of Headshakes	0.63	0.75	-0.87	-0.81

the speakers' smiles (correlated coefficient = 0.80), indicating that the listener was mimicking the speaker's smile. However, some of the speaker's personality measurements are highly correlated with the listeners' headshakes (as shown in Table 5).

The result shows that the number of headshakes is positively correlated with the speakers extroversion and agreeableness measurements, and is negatively correlated with the neuroticism and shyness measurements. In our task, headshake always indicates negative emotions towards what the speaker said. That is, if the speaker is more extroverted and agreeable, the listeners are more likely to express their negative emotions; however, if the speaker is more neurotic and shyer, the listeners tend to hide their negative emotions.

3.3 Predicting Personality from Parasocial Responses

We investigate how well we can predict personality just from the listener backchannel feedback and how well we can explain the variability of listener backchannel feedback by only using the listeners' personality. We ran a stepwise linear regression analysis between backchannel feedback (including the number of nods, the number of shakes and the duration of smiles) and personality measurements.

First, we predict personality traits from the parasocial consensus. The dependent variable is each of the personality traits (e.g. extroversion), and the independent variables are the number of head nods, the number of headshakes, and the duration of smiles produced by PCS coders. We observed significant results for neuroticism and self-consciousness. Smile itself (correlation coefficient = -0.23) can predict about 12% of the variance of neuroticism (F=3.4, p=0.07); smile (correlation coefficient = 0.2) and nod (correlation coefficient = -0.17) together can predict about 20% of the variance of self-consciousness (F=3.11, p=0.062). Second, we run the same analysis reversely; that is, the dependent variable is the number of head nods, the number of headshakes and the duration of smiles respectively, and the independent variables are the personality traits. We only observed significant result for smile. Self-consciousness (correlation coefficient = 0.868) and neuroticism (correlation coefficient = -0.658) can predict about 28% of the variance of smile (F=4.95, p=0.015). Together, this suggests we can intuit something about a speakers personality simply by looking at the responses of their conversation partner, although this relationship is rather modest.

4 Conclusion and Future Work

In this paper, we introduced our initial efforts in analyzing and modeling the variability of human nonverbal behaviors in face-to-face interaction. We extended the Parasocial Consensus Sampling (PCS) framework to make such analysis possible. The results showed that personality has significant influences on backchannel feedback and clearly demonstrated how it affects backchannel feedback. In the future, we will integrate the results into nonverbal behavior models, and test whether the virtual human driven by such models can exhibit the corresponding personalities or not. Moreover, we will further explore the PCS framework to investigate how other factors, such as interactional goal, roles in the conversation and emotions, can influence nonverbal behaviors in face-to-face interaction.

References

1. Swartout, W.R., Gratch, J., Hill Jr., R.W., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Toward virtual humans. AI Magazine 27(2), 96–108 (2006)
2. Bickmore, T.W., Puskar, K., Schlenk, E.A., Pfeifer, L.M., Sereika, S.M.: Maintaining reality: Relational agents for antipsychotic medication adherence. Interacting with Computers 22(4), 276–288 (2010)
3. Rowe, J.P., Shores, L.R., Mott, B.W., Lester, J.C.: Integrating learning and engagement in narrative-centered learning environments. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part II. LNCS, vol. 6095, pp. 166–177. Springer, Heidelberg (2010)
4. Pelachaud, C.: Simulation of face-to-face interaction. In: Proceedings of the Workshop on Advanced Visual Interfaces, pp. 269–271 (1996)
5. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, pp. 413–420 (1994)
6. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)

7. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 125–138. Springer, Heidelberg (2007)
8. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, pp. 289–296 (2009)
9. Morency, L.-P., de Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 176–190. Springer, Heidelberg (2008)
10. Jonsdottir, G.R., Thorisson, K.R., Nivel, E.: Learning smooth, human-like turntaking in realtime dialogue. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 162–175. Springer, Heidelberg (2008)
11. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology, 941–952 (2000)
12. Bevacqua, E., Mancini, M., Pelachaud, C.: A listening agent exhibiting variable behaviour. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 262–269. Springer, Heidelberg (2008)
13. Huang, L., Morency, L.-P., Gratch, J.: Parasocial consensus sampling: combining multiple perspectives to learn virtual human behavior. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, pp. 1265–1272 (2010)
14. Lao, S., Kawade, M.: Vision-based face understanding technologies and their applications. In: Li, S.Z., Lai, J.-H., Tan, T., Feng, G.-C., Wang, Y. (eds.) Sinobiometrics 2004. LNCS, vol. 3338, pp. 339–348. Springer, Heidelberg (2004)
15. Wang, Z., Lee, J., Marsella, S.: Towards more comprehensive listening behavior: Beyond the bobble head. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 216–227. Springer, Heidelberg (2011)
16. Bonferroni, C.E.: Il calcolo delle assicurazioni su gruppi di teste. In: Tipografia del Senato
17. Shiota, M.N., Keltner, D., John, O.P.: Positive emotion dispositions differentially associated with Big Five personality and attachment style. The Journal of Positive Psychology, 61–71 (2006)
18. Hartmann, T., Goldhoorn, C.: Horton and wohl revisited: Exploring viewerss experience of parasocial interactions. In: The Annual Meeting of the International Communication Association
19. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 893–910 (1999)

Head Motion Analysis and Synthesis over Different Tasks

Atef Ben Youssef, Hiroshi Shimodaira, and David A. Braude

Centre for Speech Technology Research, University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
abenyou@inf.ed.ac.uk, h.shimodaira@ed.ac.uk, D.A.Braude@sms.ed.ac.uk

Abstract. It is known that subjects vary in their head movements. This paper presents an analysis of this variety over different tasks and speakers and their impact on head motion synthesis. Measured head and articulatory movements acquired by an ElectroMagnetic Articulograph (EMA) synchronously recorded with audio was used. Data set of speech of 12 people recorded on different tasks confirms that the head motion variate over tasks and speakers. Experimental results confirmed that the proposed models were capable of learning and synthesising task-dependent head motions from speech. Subjective evaluation of synthesised head motion using task models shows that trained models on the matched task is better than mismatched one and free speech data provide models that predict preferred motion by the participants compared to read speech data.

Keywords: head motion variety, head motion synthesis.

1 Introduction

Head movement demonstrates a wide variety of meaning. For example, nodding can be used not only for agreement, but also for emphasis, indicating attention and to indicate thinking during dis-fluencies [10,12]. Such motion can be different for the same speaker in other tasks or for others speakers doing the same task.

In recent years, the problem of driving head motion from speech has become a popular topic for research. Head motion may considered as complementary information for speech or other visual information (e.g. movements of mouth; lips, jaw and tongue, and also eyebrows, eyelids movements). This information increases speech intelligibility. Munhall *et al.* [12] found that the display of head motion also improves speech perception.

Research on speech-driven talking faces began with work on synthesis of lip and mouth motions that are synchronised with speech, (lip sync.) [11]. In contrast to the lip sync on which a significant number of studies has been done, automatic synthesis of head motion from speech has not been studied extensively, especially in terms of the use of machine learning techniques. However, existing speech-driven head motion system often ignore the variance of head movements over different situation and speakers.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 285–294, 2013.

Graf *et al.* [4] showed a link between the prosody expressed by the voice and that given by the head. Yehia *et al.* [17] proposed a frame-wise mapping based on a linear-regression model to estimate head rotation angles (Euler angles) from F0. They found that the linear model had to be separately trained on each utterance sample otherwise the correlation between F0 and head motion almost disappeared. A GMM-based simple frame-wise mapping has also been employed for a talking head [7], longer temporal information was used in [3,2] and [9]. In the former, HMMs were employed to map F0 and energy to a frame-wise VQ code of head rotation angles, whereas in the latter a discrete HMM was used to decode a sequence of animation cluster codes from the pitch and intensity features at every input syllable. Sargin *et al.* [13] developed a fully HMM-based approach for mapping the trajectory of F0 and intensity to the one of head rotation angles, in which parallel HMMs were used to cluster trajectories of speech and head motion separately. Hofer *et al.* [6,5] proposed the use of human-understandable head-motion units (e.g. nodding and shaking) as the model unit of HMMs. In their approach HMMs are trained with the combined streams of audio speech features (MFCC, F0, and energy) and head rotation angles. Despite the very low frame-wise correlations they found between the speech and head motion features, it was shown that head motion units were correctly recognised with an accuracy of approximately 70% on a free-speech data set, and reasonably natural head motions were synthesised. Lee *et al.* [8] evaluate 3 different machine learning techniques in head nods and eyebrow movements prediction. They found that the behaviors generated by the different models affect the human perception of the agent.

In this paper, we used an ElectroMagnetic Articulograph (EMA) corpus that contain articulatory and head movements recorded synchronously with audio. The rationale for considering articulatory features is that there is some evidence that articulatory movements, e.g. opening the jaws, contribute to the movement of the head [18]. The goal of the work described in this paper is to analysis the head motion variety over different tasks and speakers and their impact on head motion synthesis.

2 Data Set

In the present study, we used 12 English native speakers (4 males and 8 females denoted by $R00XX\_csX$) of the Edinburgh Speech Production Facility (ESPF) corpus [15]. This corpus contains articulatory and head movements over time synchronously recorded with audio and electropalatography. Using two Carstens AG500 electromagnetic articulometers positioned 8.5m apart to avoid electromagnetic inter-machine interference, the articulatory and head movements of English speakers in *dialogue* was recorded using 3D positions of sensors glued on the lips, tongue, jaw, and head. Communication among participants and experimenters is regulated via a talkback system.

Each speaker recorded different tasks. The recorded tasks were:

- Script reading: the speaker reads the script "Comma gets a cure"
- Map Giver: Map task, the speaker is the instruction giver.
- Map follower: Map task, the speaker is the instruction follower.
- Spot diff.: 3 Spot the difference picture tasks were recorded (Street, Diapix, Farm), the speakers was collaborating to find the differences.
- Repet. Teller: 2 Repetition task (Dance story, Loch story), the speaker is the reader.
- Repet. Shadower: 2 Repetition task, the speaker repeats what the teller reads out.
- Story Teller: Shadowing task, the speaker tell a story of his choice
- Story Shadower: Shadowing task, the speaker follow the the partner's story.

Depending on the speaker, the duration of the speech is between 11 and 38 minutes and the number of the available task is between 8 and 10.

2.1 Head Motion Data

Head motion is represented by the head correction of the articulatory trajectories in the data set. Four coils attached to the upper incisor, to the nose and to the left and right ears served as references to extract the head movements. Head translations and rotations were calculated in order to remove the contribution of head movements from the articulatory data. In this study, head motion are represented by head rotations (R_z, R_y, R_x) about the z, y and x axes, respectively. In order to use a common frame shift of 10 ms, the data was down-sampled to 100 Hz and their first derivatives was added.

2.2 Speech Data

Articulatory Data. Articulatory movements correspond to the horizontal and vertical midsagittal (x, y) coordinates of six coils attached to the speech organs. A jaw coil is attached to the lower incisors, three coils are attached to the tongue (tongue tip, tongue middle and tongue back), a coil is attached to the upper lip and another to the lower lip. The articulatory data (denoted by EMA and represented by 12 parameters) was down-sampled to 100 Hz to match the head motion data and their first derivatives were added. Note that audio-speech signal was recorded synchronously with EMA data (not used in this study).

3 Head Motion Variation

It is well-known that subjects vary greatly in their head movements. Although head movements is associated with many factors that can be explained by the settings of more/less speaking, the dialogue partner, the seating arrangement and also speaker's personality, social stance, physiological state and visual focus

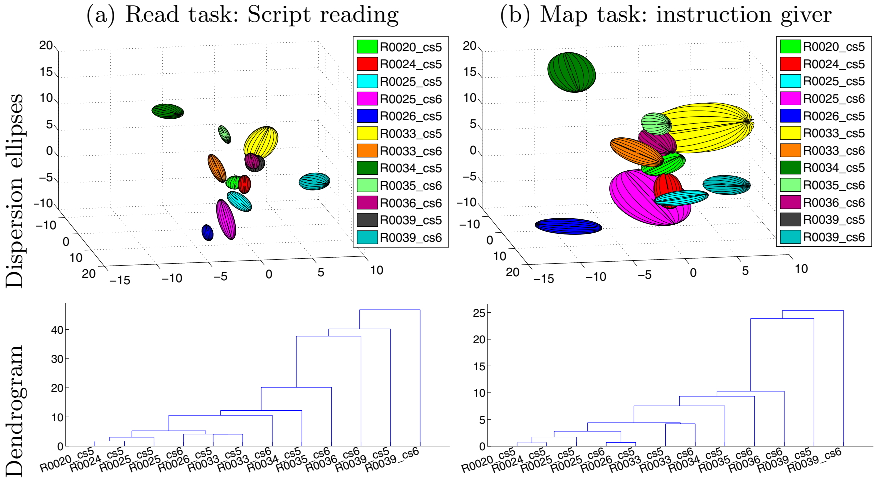


Fig. 1. Head motion variation of the same tasks over different speaker

of attention. Modelling the impacts of these factors and their dependency is a challenging problem.

This section discuss how the head motion varies between tasks and speakers. The dispersion ellipses of the head motion is represented in the $3D$ head space (i.e. x, y and z axes) by the mean and the full covariance matrix over the sets of the task instance (cf. Fig. 1) and speakers (cf. Fig. 2).

For a clear representation, we display all the task for two speakers in Fig. 2 and all speakers for two tasks in Fig. 1. Fig. 1 presents the distribution of two different tasks: (a) Read task (i.e. Script reading) when the speaker reads a script and (b) Map task (i.e. instruction giver) when the speaker is given the map instruction to the follower. This illustrates the very low variability of the head motion for the read task, as expected since the speaker's head is focusing on the script. The high variability of the head motion observed on the map task, can be explained by the free speech given by the speaker when he is collaborating with the follower to find the way. The variation of the head motion changes from speaker to another specially on the free speech (i.e. map task). Confusion trees have been built for speakers, based on the matrix of Mahalanobis distances of the head motion between each pair of speakers We confirm that the head motion is speaker dependent.

Fig. 2 displays the dispersion ellipses of head motion over all tasks for two speaker (i.e. male speaker $R0020\_cs5$ and female speaker $R0039\_cs6$), as well as the confusion trees that have been built for tasks, based on the matrix of Mahalanobis distances of the head motion between each pair of task. Using hierarchical clustering to generate dendrograms, we find different distance between tasks depending on the speaker. We can confirm that head motion is not only speaker dependent but also task dependent.

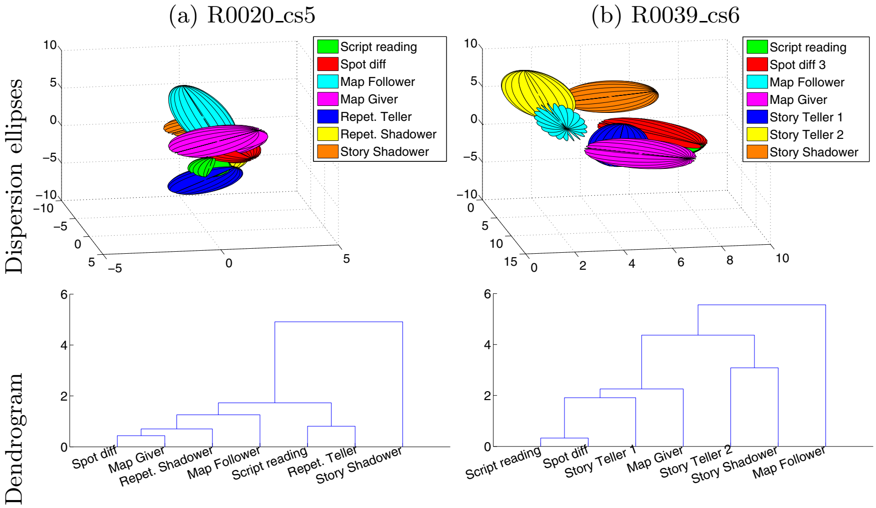


Fig. 2. Head motion variation of the same speaker over different tasks

Hierarchical clustering was performed based on Mahalanobis distances. The results are viewed in a dendogram, which displays the nodes arranged into their hierarchy and also shows how far apart the items were. Dendrograms of Fig. 1 and Fig. 2 that display the Mahalanobis distances between tasks and speakers, respectively, show that the distance between tasks (up to 6) is lower than the distance between speakers. This illustrate that head motion is depends more on speaker variation rather than tasks variation.

4 Speech Driven Head Motion Synthesis

The key idea of this paper is to model the head motion variation and its impact on head motion synthesis from speech.

We recall the experiments published previously in [1]. We found that canonical correlation analysis (CCA) on a free speech data shows that the articulatory features are more correlated with head rotation than prosodic and/or cepstral speech features. Therefore, we used measured articulatory features as input feature for speech driven head motion synthesis.

4.1 Clustering of Head Motion Data

Data annotation is an essential step in the HMM training process. However, manual annotation is often time-consuming and expensive. Furthermore, as head motion is concerned not only one segmentation will necessarily be right.

In our experiments, the training data of head motions were automatically labelled using an HMM-based clustering technique that may be able to provide both short and long segments that provide a statistical description of a particular motion. Algorithm 1 explain the instructions used to label automatically the data. We used GMM clustering to initialise the HMMs. Over the task data of a speaker, GMM with K distributions was trained using EM algorithm. Then, the data was clustered using the trained GMM into K clusters. Each cluster was used to initialise an HMM. The HMMs parameters were re-estimated using EM algorithm and then new cluster labels were decoded using Viterbi algorithm. This process was repeated until convergence was reached.

Algorithm 1. CLUSTERING of head motion data

```
Input: Head motion data of a task of one speaker
Output: Head motion cluster labels with their durations
Train GMM with K distributions using EM algorithm.
Cluster the data into K cluster labels using the trained GMM based on the
maximum likelihood.
Initialise K HMMs with K clusters
repeat
    Re-estimate the HMMs parameters using EM algorithm
    Decode new cluster labels using the re-estimated HMMs and Viterbi
    algorithm
until convergence is reached
return cluster labels
```

In order to find the optimal number of clusters and the optimal HMM topology that match best with the task of head motion synthesis, we varied the number of clusters, K. To define the best HMM configuration, we synthesise the head motion trajectories from the recognised sequence of clusters and the trained HMMs. Then, we evaluate it by a comparison with the original head motion trajectories.

Preliminary experiment shows that the optimal number of clusters variate between 11 and 15, although it varies across the speakers. Busso *et al.* [3] found that 16 clusters achieves the best result of generating head motion sequences from prosodic features.

A similar experiment was done for the number of states per HMM to confirm that there is no clear strategy for deciding the optimal number of states when clustering is concerned. Thus the number was fixed to 5 for the following experiments.

4.2 Head Motion Synthesis

An overview of the multi-stream HMM-based speech-driven head motion system is presented in Fig. 3.

In this experiment, we used 15 clusters to train task-dependent multi-stream HMMs. 5-state left-to-right no-skip context-independent HMMs were used to model speech and head motion streams of the task. The proposed technique is based on the joint modelling of articulatory and head motion features, for each cluster.

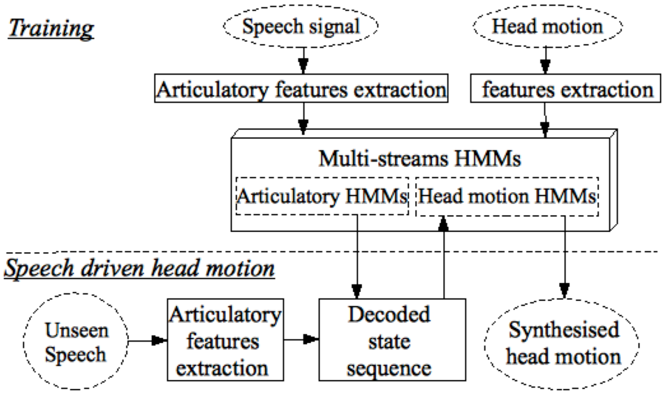


Fig. 3. Overview of the speech-driven head motion system

In training stage, streams of head motion and articulatory feature vectors are joined to train multi-stream HMMs, whose model units are determined by the HMM-based clustering technique [1]. For each stream, the emission probability density function of each state is modelled by a multivariate Gaussian distribution with a diagonal covariance matrix.

The speech driven head motion synthesis is achieved in 2 steps: 1) finding the most likely HMM state sequence from the articulatory observations; 2) inferring the head motion from the decoded state sequence. For a given speaker's articulatory feature vectors X, we predict the head motion features Y such as

$$\hat{Y} = \arg\max_{Y} \{p\left(Y|\lambda^{y,x}, Q^{y,x}\right) P\left(\lambda^{y,x}, Q^{y,x}|X\right)\} \quad (1)$$

where $\lambda^{y,x}$ is the parameters set of the head-motion cluster-size HMM and $Q^{y,x}$ the HMM state sequence. $\hat{Y}$ is obtained by maximizing separately the two conditional probability terms of Eq. 1. First, we decode the HMM state sequence by maximising $\left\{\left(\hat{\lambda}^{y,x}, \hat{Q}^{y,x}\right) = \arg\max_{\lambda,Q} \{P\left(\lambda^{y,x}, Q^{y,x}|X\right)\}\right\}$ using the Viterbi algorithm. Second, we synthesise the head motion trajectories by estimating $\left\{\hat{Y} = \arg\max_{Y}\{p(Y|\hat{\lambda}^{y,x}, \hat{Q}^{y,x})\}\right\}$, using the maximum-likelihood parameter generation algorithm (MLPG) algorithm [14].

4.3 Evaluation

To evaluate the impact of the head motion variation over tasks, we used data of two tasks recorded by the same speaker (i.e. $R0020\_cs5$): map as instruction giver and script reading. The data of each task was split in two partition:

1. Training partition: two-third of the task was used to train the models (mapHMMs trained from the map task training data and readHMMs trained from script reading training data).

Table 1. Pearson's correlation between the original head motion and the synthesised one using matched models and mismatched models

Task of speech input \ used models	mapHMMs	readHMMs
Map task	0.47	−0.34
Read task	−0.38	0.48

2. Test partition: the remaining third was used for test. In order to evaluate the impact of the task on the head motion, cross-task speech-driven head motion synthesis was tested.

The articulatory speech data of the test task was the input for the two trained models to form matched and mismatched models on synthesis stage.

Objective Evaluation. To evaluate the impact of the task difference, Pearson's correlation between the original head motion and the synthesised one using the matched models and mismatched models was calculated. As can be observed in Table 1, the correlation on the matched condition is higher than the mismatched one. This result suggests that the synthesised data follow the motion of the task used for training rather than following the speech input specially for the mismatched condition.

Mismatched condition gives high, even though negative correlation. This means that as one rotation angle increases in value, the synthesised one decreases in value (i.e. when the head moves from up to down, the estimated movements was from down to up). This confirm that the proposed models was capable of learning and synthesising task-dependent head motions from speech.

Subjective Evaluation. We performed a subjective A/B comparison test to measure the opinions on the naturalness of the synthesised head motion. The participant are asked to chose between two head motion on a scale of A better than B, no preference and B better than A. 6 side-by-side comparison pairs was used: 3 pairs for each task. The 3 comparison was between the measured head motion (org) and two synthesised ones from matched HMMs and from mismatched one. Each comparison pairs is 50 seconds video length. The subjective tests were performed by 11 participants. The average preference are shown in Fig. 4.

The original data is typically perceived as much more realistic, except for the Read task, for which the mapHMM appears better. By looking to the map task results, we found that matched HMMs are more preferred than mismatched ons. However for read task results, contrary to our expectation, synthesised head motions using the mismatched models that was trained on free speech (i.e. mapHMMs) are preferred by the participant rather than the matched models. This can be explained that the free speech data that have more variation compared to read speech (see Fig. 1) may provide more expressive and preferred motion compared to read speech.

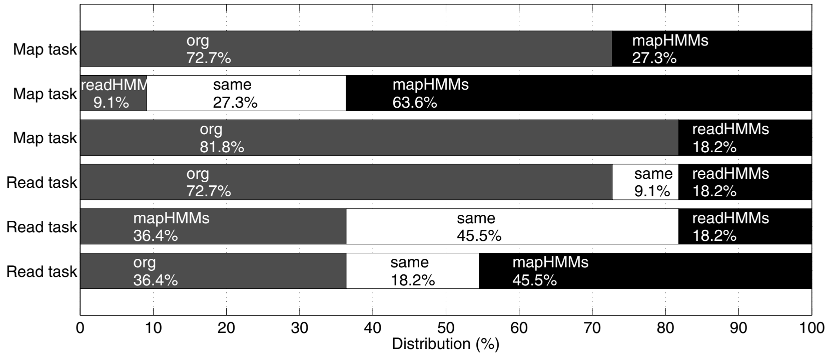


Fig. 4. Subjective A-B tests results over 11 participants

5 Conclusion

We have presented an analysis of head motion variation and the impact of this variety on synthesis. Over 12 speakers, we confirm that the head motion varies depending not only on the speaker but also on task. Articulatory features that have more correlation with head motion than acoustic features were used to drive head motion [1]. Experimental results confirmed that the proposed model was capable of learning and synthesising task-dependent head motions from speech. Synthesised head motion trajectories are more correlated with original motion when it was synthesised using a models trained on matched tasks than using mismatched ones. The subjective evaluation tests indicates that the free speech data may provide more expressive motion compared to read speech. A better head movement models could be trained using free speech data collected over similar tasks to the ones the avatar is supposed to perform.

This work could be extended in several ways. The advantage of HMM-based head motion synthesis is that a possible emotional and personalised motion can be achieved using adaptation techniques [16]. Further studies will include an extension to speaker-independent models with speaker adaptation.

In real-world head motion synthesis scenarios, it is not practical to assume the availability of articulatory measurements from a user. To address this challenge, acoustic-to-articulatory mapping system may be used to predict articulatory features from an acoustic signal. Another motivation of using acoustic-to-articulatory mapping is to use the predicted articulatory features for lip sync.

Acknowledgements. This work was supported by EU FP7 SSPNet (grant agreement no. 231287) The financial assistance of the National Research Foundation of South Africa (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

References

1. Ben Youssef, A., Shimodaira, H., Braude, D.A.: Articulatory features for speech-driven head motion synthesis. In: Proceedings of Interspeech, Lyon, France (2013)
2. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1075–1086 (2007)
3. Busso, C., Deng, Z., Neumann, U., Narayanan, S.: Natural head motion synthesis driven by acoustic prosodic features. Computer Animation and Virtual Worlds 16(3-4), 283–290 (2005)
4. Graf, H., Casatto, E., Strom, V., Huang, F.J.: Visual Prosody: Facial Movements Accompanying Speech. In: Proc. 5th International Conf. on Automatic Face and Gesture Recognition, pp. 381–386 (2002)
5. Hofer, G.: Speech-driven Animation Using Multi-modal Hidden Markov Models. PhD thesis, Uni. of Edinburgh (2009)
6. Hofer, G., Shimodaira, H.: Automatic head motion prediction from speech data. In: Proc. Interspeech 2007 (2007)
7. Le, B., Ma, X., Deng, Z.: Live speech driven head-and-eye motion generators. IEEE Transactions on Visualization and Computer Graphics 18(11), 1902–1914 (2012)
8. Lee, J., Marsella, S.: Modeling speaker behavior: A comparison of two approaches. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 161–174. Springer, Heidelberg (2012)
9. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. In: SIGGRAPH Asia 2009 (2009)
10. McClave, E.Z.: Linguistic Functions of Head Movements in the Context of Speech. Journal of Pragmatics 32(7), 855–878 (2000)
11. Morishima, S., Aizawa, K., Harashima, H.: An intelligent facial image coding driven by speech and phoneme. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1989, vol. 3, pp. 1795–1798 (1989)
12. Munhall, K., Jones, J., Callan, D., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: head movement improves auditory speech perception. Psychological Science 15(2), 133–137 (2004)
13. Sargin, E., Yemez, Y., Erzin, E., Tekalp, A.M.: Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Trans. Patt. Anal. and Mach. Intel. 30(8), 1330–1345 (2008)
14. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. 1315–1318 (2000)
15. Turk, A., Scobbie, J.M., Geng, C., Macmartin, C., Bard, E.G., Campbell, B., Diab, B., Dickie, C., Dubourg, E., Hardcastle, B., Hoole, P., Kainada, E., King, S., Lickley, R., Nakai, S., Pouplier, M., Renals, S., Richmond, K., Schaefer, S., Wiegand, R., White, K., Wrench, A.: An edinburgh speech production facility
16. Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., Nakano, Y.: Model adaptation approach to speech synthesis with diverse voices and styles. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV–1233–IV–1236 (2007)
17. Yehia, H., Kuratate, T., Vatikiotis-Bateson, E.: Linking Facial Animation, Head Motion, and Speech Acoustics. Journal of Phonetics 30, 555–568 (2002)
18. Zafar, H., Nordh, E., Eriksson, P.O.: Temporal coordination between mandibular and headneck movements during jaw opening closing tasks in man. Archives of Oral Biology 45(8), 675–682 (2000)

Using Virtual Agents to Guide Attention in Multi-task Scenarios

Philipp Kulms and Stefan Kopp

Sociable Agents Group, Center of Excellence 'Cognitive Interaction Technology' (CITEC)
Faculty of Technology, Bielefeld University, Germany
{pkulms,skopp}@techfak.uni-bielefeld.de

Abstract. Humans have the ability to efficiently decode human and human-like cues. We explore whether a virtual agent's facial expressions and gaze can be used to guide attention and elicit amplified processing of task-related cues. We argue that an emphasis on information processing will support future development of assistance systems, for example by reducing task load and creating a sense of reliability for such systems. A pilot study indicates subjects' propensity to respond to the agent's cues, most importantly gaze, but to not yet rely on them completely, possibly leading to a decreased performance.

Keywords: Virtual embodied assistant, attention, task-switching, social information, cognitive processes.

1 Introduction and Concept

Human-machine interaction can be very complex, especially when it involves executive functions (attention allocation, priority setting, response scheduling, working memory, etc.) operating on multiple tasks. For example, in car-driving, operating industrial facilities, or air traffic control, humans often have to perform multiple tasks with time-varying demands. While Cognitive Scientists have investigated mechanisms of switching between multiple tasks and, e.g., how task interferences result in costs and errors [1], fields like human factors, software ergonomics, and human-computer interaction strive to prevent multi-task situations or, since they are often unavoidable, to provide users with support mechanisms [2]. In this paper we study how a virtual agent can provide support in such multi-task scenarios by guiding attention with social signals, and what effects the presence of such an agent has on the user and the respective tasks.

Interface agents are far from being a novel idea (see e.g. [3]). However, although positive social effects of virtual agents on processes such as learning have been confirmed [4], only a little research has investigated the detailed cognitive effects of an agent's presence on users that have to fulfill given tasks. We focus on situations where users have to carry out multiple tasks simultaneously (see Fig. 1, A). In most cases, these tasks interfere in some way and automatic performance is hard to establish. This means users have to employ a task-switching strategy that requires to

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 295–302, 2013.

manage attention accordingly: they would focus on one task temporarily, but have to monitor the other to decide whether a switching is necessary (Fig. 1, B). This split attention as well as frequent task switching is likely to hamper performance in both tasks. We explore whether a virtual agent can assist in this task-switching by guiding attention to those tasks that need to be attended to (Fig. 1, C).

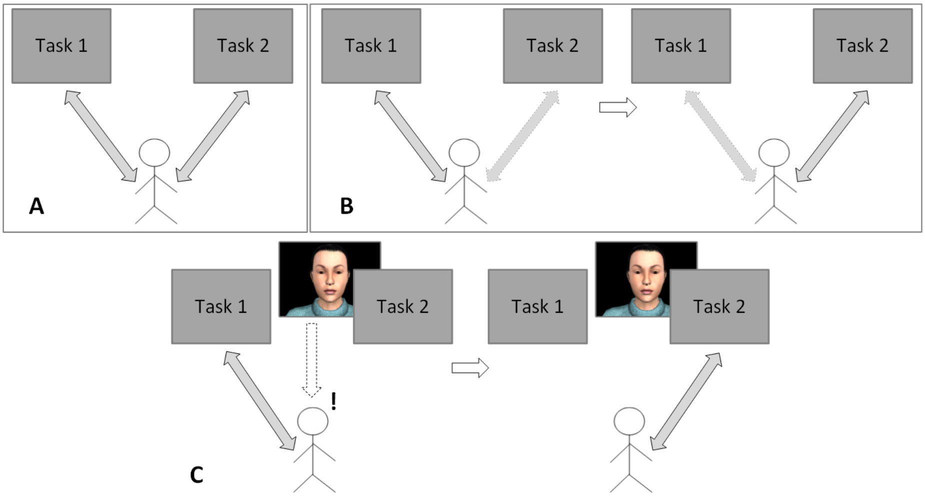


Fig. 1. Schematic rationale of our virtual assistant. In most situations, performing multiple tasks truly simultaneously is virtually impossible (A). Thus the user has to switch between the tasks, based on interlaced observation of both tasks which may hamper the performed task (B). We envision an agent to socially signal the need to switch tasks in resource-sparing ways (C).

Our approach is based on two assumptions: First, virtual human-like agents are social actors, i.e. they are able to issue social signals perceived as such, and a variety of research suggest that socially and emotionally salient cues can modulate selective attention, working memory, and motor responses (for an overview, see [5]). In particular, social stimuli appeal to human low-level perception and automatic, efficient bottom-up mechanisms of perception that amplify identification of environmental cues and release resources for executive processes, including planning and decision making [6]. For example, our brain is optimized to categorize facial expressions and efficiently de-correlates the basic expressions happiness, surprise, fear, anger, disgust, and sadness, such that overlapping signals are reduced [7]. The second assumption is that attention is controllable top-down, can be geared toward execution of specific behaviors, and is easier to divide in situations of low task load [8]. Such top-down attention mechanisms can facilitate visual attention for positive or arousing emotional stimuli [5]. Similar facilitation effects can be observed for the social information a face conveys, and executive functions were recently shown to be burdened less by social cues, pointing to a working memory unit for social information [9].

Clearly, gaze is another highly task-relevant cue. Friesen and Kingstone [10] have shown that people respond significantly faster to a target when gaze direction by a schematic face predicted the target location. This finding was confirmed for facial attention cues when additional head movements were presented to the observer which suggests that orienting toward another person's social attention is reflexive and stimulus-driven [11]. Within human-computer interaction, gaze-based attention regulation warrants interactive goal-directed communication [12, 13]. Naturally, from a conversational point of view, researchers are very much interested in cues that shape speaker/listener behaviors, one of which, of course, is gaze [e.g. 14, 15]. However, since gaze is closely tied to both sensing and communicating attention, it is in fact qualified for regulating information processing in a variety of scenarios [16].

Against this background, we want to explore whether a virtual agent that is additionally present in a multi-task scenario can assist users with managing attention and task-switching, such that less cognitive resources are expended on the switching and more are available for the tasks. To this end, the agent should facilitate task performance and display easy-to-perceive social behaviors that help to guide attention towards specific (sub-)tasks. Specifically, we are interested in the orienting and alerting dimension of attention [17] and whether an agent's social cues can provide better support for guiding attention during multi-tasking compared to non-social cues. We utilize emotional facial expressions and gaze as examples of such social cues that trigger robust, automatic bottom-up response mechanisms, bias attention, amplify processing, and can be facilitated top-down. In a related study on a virtual agent's ability to cue observer attention, subjects responded more quickly to fully animated gaze cues than to static or non-fully animated images [18]. For the virtual agent Greta, a facial expression taxonomy was created which lists "I warn" as one possible functional value, involving tense eyelids, a small frown, and eye-contact [19]. In the following we present an initial pilot study exploring the effects of a virtual agent with similar social cues vs. simple graphical symbols to guide the user's attention in a multi-task scenario. Such social cues will usually be very different from the task-related information such that an additional, non-interfering channel of easy-to-process information is opened. The results are discussed as first evidence and basis for future studies.

Obviously, a number of open questions arises. For one thing, pursuing a specific task-related goal can engage different cognitive processes that might interfere with the additional "task" of observing an agent, albeit the task-switching may be lightened. As a potential issue, distraction may arise if the agent's behavior cannot be decoded efficiently during task performance but instead leads to confusion or frustration.

2 Pilot Study

As a pilot study (N = 10), we conducted a dual-task experiment with a cognitively demanding primary task and a secondary task that does not require continuous but occasional attention. In the agent condition (see Fig. 2), a virtual agent was displayed that exhibits cues which relate to the status of the secondary task. In the control condition, basic graphical symbols fulfilled this function. All subjects were right-handed. They were randomly assigned to one of the conditions.

Procedure. Subjects were asked to perform both tasks at the same time and to consider any assisting cues for their actions. The primary task was a semantic task including word categorization (noun, verb, or adjective) by clicking the appropriate response button. The number of correct answers was displayed at the bottom of the window. The secondary task was to monitor a glass getting filled with water and try to prevent it from overflowing by pressing a 'Drain' button. The water level was not visible unless a 'Status' button was pressed and held. Since subjects were occupied with the primary task as well, they were required to find a strategy which enables them, instead of constantly switching between both tasks, to monitor the secondary task efficiently while performing the primary task. When the glass overflowed due to delayed draining, it gets replaced by an empty one, a warning light went on, and a score displayed how often this occurred. In order to ensure that the secondary task's status cannot be predicted, the rate of filling varied pseudo-randomly. Furthermore, to prevent subjects from blindly draining the water as often as possible, they were told that draining comes with a cost which, at some point, increases the difficulty. Subjects were told that the goal is to maximize the score of the primary task and to keep the negative score as well as the costs of the secondary task as low as possible.

Prior to the main run, subjects went through a short training phase. Shortly before the end of the experiment, subjects were asked to comment what they perceived on the screen and the way they perform their tasks, without letting their focus drift away from their actual performance. A brief acoustic signal was used as start signal. Comments were audio-taped for later analysis. The input and output channels were unimodal; users relied on visual perception and responded via mouse click. Each session lasted approximately 10 minutes.

Manipulations. The agent exhibited fully animated social cues. The status of the secondary task was linearly mapped onto the intensity of a surprised and thus alerting facial expression (eyebrows and eyelids raised, mouth open; see Fig. 2). That is, the more the glass is filled with water the more intense the expression. Additionally, a few moments before the glass is full, the agent gazes at its direction. Virtual agent behavior was generated with AsapRealizer [20]. In the control condition, a colored elliptic shape of the size of the agent's face replaced the facial expression. The coloring ranged from green (empty glass) to dark orange (full glass). Instead of gaze, an animated arrow of the size of the agent's head was used. Since these items usually are well-known, we expected them to be clear attentional cues.

Dependent Variables. In the dual-task study, primary task performance was assessed in terms of the number of correct and erroneous word categorizations. Secondary task performance was assessed using the number of overflowing glasses and how often subjects checked the status instead of relying on the assistant's cues.

Evaluation of Facial Expressions. An online study (N = 95) was conducted to evaluate the facial expressions of the agent. Above all, the study had the purpose of confirming that the agent's alerting face is reliably recognized. Subjects were presented two still pictures consecutively, showing the agent's neutral and alerting face, respectively. The pictures were rated on a 5-point Likert scale. A paired-samples t-test indicates that the alerting face (M = 4.28, SD = .93) was perceived as significantly more alerting than the neutral face (M = 2.08, SD = 1.24), $t(95) = 15.68$, $p < .001$. Moreover, the alerting face was attributed significantly stronger with conveying information about the environment, that is, giving a hint about something (alerting: M = 2.94, SD = 1.32, neutral: M = 2.06, SD = 1.12), $t(95) = 5.93$, $p < .001$, and carrying an important information (alerting: M = 2.79, SD = 1.35, neutral: M = 1.92, SD = 1.04), $t(95) = 5.93$, $p < .001$. Interestingly, the alerting face (M = 1.86, SD = 1.14) was perceived as carrying less worry/concern than the neutral face (M = 3.17, SD = 1.25), $t(95) = 8.16$, $p < .001$. No difference occurred for the attribution of a sense of danger to the faces. Note that the stimulus material only depicted the facial expressions and that rating instructions did not point to anything the agent actually or hypothetically observes.

Fig. 2. Snapshot of the system in the agent condition (left: primary task; right: secondary task). The agent gazes with an alerted facial expression toward the secondary task as the glass is almost full, demanding for user reaction.

3 Results

It is important to note that the assessed data merely serve as indicator of whether subjects are able to relate agent behavior to the task and whether a tendency to rely on it can be found. Obviously, the small dataset does not qualify for statistical inference and exhibits considerable standard deviations. The primary task error rate of one subject was excluded from the dataset due to performing worse than three times the standard deviation.

Results of the primary task indicate lower performance in the presence of the agent. On average, subjects categorized 303.20 words correctly (*SD* = 69.45) which is somewhat less compared to the control condition (*M* = 365.20, *SD* = 116.16). Error rates did not differ (agent: *M* = 20.00, *SD* = 12.10; control: *M* = 20.75, *SD* = 8.23). Performance in the secondary task too was somewhat worse under assistance of the agent as subjects let the glass overflow more often (agent: *M* = 18.40, *SD* = 6.91; control: *M* = 16.80, *SD* = 3.49). Finally, subjects checked the status more often when they were guided by the agent (*M* = 53.00, *SD* = 30.54) than by symbols (*M* = 24.60, *SD* = 15.85).

Analysis of the qualitative data suggests that subjects interpreted gaze of the agent as alerting signal and felt inclined to respond to it. The same holds for the arrow in the control condition, yet the combination of arrow and colored shape caused occasional confusion because some tried to connect the symbols' meanings with each other. All subjects reported relying on gaze of the agent entirely, ignoring or simply not perceiving its facial expressions.

4 Discussion

In both tasks the same pattern seemed to emerge: control condition subjects who were guided by symbols instead of social cues, performed somewhat better than agent condition subjects. They categorized more words correctly, did not make more categorizing errors, and managed the water level of the glass more successfully. These results do not seem in line with the fact that the agent‘s alerting facial expression was correctly identified in the online study and was attributed with a signaling function. Furthermore, subjects perceived the agent‘s gaze as an attentional cue and intuitively followed it attending the secondary task. So, why did participants in the agent conditions perform worse than in the control condition? A number of explanations seem possible.

The most likely explanation is that in the condition with human-like alerting cues, subjects allocated more attention to the hidden status of the secondary task. That is, instead of relying on the agent as a prompting device for the secondary task, subjects may have adopted more complex and time-consuming behavioral sequences consisting of checking and reacting operations. Since subjects apparently perceived gaze as an attentional cue, this means they went halfway toward our envisioned goal (Fig. 1). The downside is that, despite subjects' propensity to respond to the agent's cues and to use the agent as attention regulation device for the secondary task, they may have not relied on them completely. One possible reason could be that subjects did not yet "trust" it enough to switch tasks solely based on its behavior. It will be important to further explore this hypothesis against the background of potential issues mentioned in the introduction.

Another explanation is suggested by subjects' thoughts on their own proceeding throughout the tasks. They indicated that in a demanding situation, social cues of the agent have to exhibit a certain saliency to be recognized, despite the fact that they convey task-relevant information. It may thus be helpful to track user gaze in order to unravel the flow of information processing during the interaction and to check whether the cues themselves were attended to. Indeed, the results on interpretation of the facial expressions' emotional content point to the need for careful investigation of facial expressions, their saliency, and unequivocal interpretation in future studies. Furthermore, on the appearance level of the agent, even its gender may significantly alter attention regulation, at least during initial phases [21].

Finally, we cannot rule out the possibility that the mere social presence of the agent may have had a negative influence on primary and secondary task performance. That is, the presence of the agent with its human-like cues may have caused distraction or irritation, thus hampering task performance. Designing more natural interactions and inducing a stronger sense of a shared goal may at least minimize the social presence issue.

In sum, our study has revealed important and interesting findings on how social cues (here, gaze and facial expression in combination) are processed in cognitively demanding settings. The next steps will be to put increased emphasis on agent cues and users' perception of said cues, so as to create a sense of reliability and trust, arising either from task interaction or pre-task priming. While the fundamentals of information processing in cognitively demanding scenarios are laid out [22], we now seek to explore the context-dependent factors that influence the processing of virtual agent cues in such settings. This may lead to new insights in how human-like cues delivered by a virtual agent are processed by humans, and which of these cues can be exploited to support users in multi-task performance.

Acknowledgements. This research was supported by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster 'it's OWL', managed by the Project Management Agency Karlsruhe (PTKA), as well as by the Deutsche Forschungsgemeinschaft (DFG) within the Center of Excellence 277 'Cognitive Interaction Technology' (CITEC). The authors are responsible for the contents of this publication.

References

1. Monsell, S.: Task switching. Trends in Cognitive Sciences 7(3), 134–140 (2003)
2. McFarlane, D.: Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. Human-Computer Interaction 17, 63–139 (2002)
3. Lieberman, H.: Autonomous interface agents. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 67–74. ACM, New York (1997)
4. Mayer, R.E., DaPra, C.S.: An embodiment effect in computer-based learning with animated pedagogical agents. Journal of Experimental Psychology 18(3), 239–252 (2012)
5. Pourtois, G., Schettino, A., Vuilleumier, P.: Brain mechanisms for emotional influences on perception and attention: What is magic and what is not. Biological Psychology 92(3), 492–512 (2012)
6. Yantis, S.: Goal-directed and stimulus-driven determinants of attentional control. In: Monsell, S., Driver, J. (eds.) Attention and Performance XVIII, pp. 73–103. MIT Press, Cambridge (2000)

7. Smith, M.L., Cottrell, G.W., Gosselin, F., Schyns, P.G.: Transmitting and Decoding Facial Expressions. Psychological Science 16(3), 184–189 (2005)
8. Kahneman, D.: Attention and effort. Prentice Hall, Englewood Cliffs (1973)
9. Thornton, M.A., Conway, A.R.A.: Working memory for social information: Chunking or domain-specific buffer? NeuroImage 70, 233–239 (2013)
10. Friesen, C.K., Kingstone, A.: The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic Bulletin & Review 5(3), 490–495 (1998)
11. Langton, S.R., Bruce, V.: Reflexive visual orienting in response to the social attention of others. Visual Cognition 6(5), 541–567 (1999)
12. Chopra-Khullar, S., Badler, N.I.: Where to look? Automatic attending behaviors of virtual human characters. In: Etzioni, O., Müller, J.P., Bradshaw, J.M. (eds.) Proceedings of the Third International Conference on Autonomous Agents, pp. 16–23. ACM, New York (1999)
13. Peters, C., Asteriadis, S., Karpouzis, K.: Investigating shared attention with a virtual agent using a gaze-based interface. Journal on Multimodal User Interfaces 3(1–2), 119–130 (2010)
14. Lee, J., Marsella, S.C., Traum, D.R., Gratch, J., Lance, B.: The Rickel gaze model: A window on the mind of a virtual human. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 296–303. Springer, Heidelberg (2007)
15. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N.: Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In: HRI 2009 Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction, pp. 61–68. ACM, New York (2009)
16. Shell, J.S., Selker, T., Vertegaal, R.: Interacting with groups of computers. Communications of the ACM 46(3), 40–46 (2003)
17. Posner, M.I., Petersen, S.E.: The attention system of the human brain. Annual Review of Neuroscience 13(1), 25–42 (1990)
18. Martinez, S., Sloan, R., Szymkowiak, A., Scott-Brown, K.: Using virtual agents to cue observer attention. In: CONTENT 2010: The Second International Conference on Creative Content Technologies, pp. 7–12. International Academy, Research, and Industry Association (2010)
19. Pelachaud, C., Poggi, I.: Subtleties of facial expressions in embodied agents. The Journal of Visualization and Computer Animation 13(5), 301–312 (2002)
20. van Welbergen, H., Reidsma, D., Kopp, S.: An Incremental Multimodal Realizer for Behavior Co-Articulation and Coordination. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 175–188. Springer, Heidelberg (2012)
21. Mojzisch, A., Schilbach, L., Helmert, J.R., Pannasch, S., Velichkovsky, B.M., Vogeley, K.: The effects of self-involvement on attention, arousal, and facial expression during social interaction with virtual others: A psychophysiological study. Social Neuroscience 1(3–4), 184–195 (2006)
22. Wickens, C.D., Carswell, C.M.: Information processing. In: Salvendy, G. (ed.) Handbook of Human Factors and Ergonomics, pp. 111–149. John Wiley & Sons, Hoboken

Gesture with Meaning

Margaux Lhommet and Stacy C. Marsella

USC Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista, CA
{lhommet,marsella}@ict.usc.edu

Abstract. Embodied conversational agents (ECA) should exhibit nonverbal behaviors that are meaningfully related to their speech and mental state. This paper describes Cerebella, a system that automatically derives communicative functions from the text and audio of an utterance by combining lexical, acoustic, syntactic, semantic and rhetorical analyses. Communicative functions are then mapped to a multimodal behavior performance. Two studies demonstrate that the generated performances are meaningful and consistent with the speech.

Keywords: nonverbal behavior, embodied conversational agent.

1 Introduction

Although it may seem that minds somehow directly interact, human face-to-face interaction is realized through the body. Beyond the words uttered, nonverbal behavior such as the flip of a hand, a gaze aversion, or a slumped posture, can powerfully influence interaction. These behaviors are so pervasive in every moment of the dialog that their absence also signals information - that something is wrong, for example, about the physical health or mental state of the person.

Our interest in such behaviors lies in a desire to model and automate the generation of nonverbal behavior for convincing, life-like virtual character performances.

A key challenge to the automation is understanding the nature of this nonverbal channel. Nonverbal behaviors establish a pervasive flow of information between participants in a conversation, because there is a rich interconnection between a person's mental processes and their body. Communicative intentions are conveyed, providing information that embellishes, substitutes for and even contradicts the information provided verbally (e.g., [1,2]). Shifts in topic can be cued by shifts in posture or shifts in head pose. Comparison and contrasts between abstract ideas can be emphasized by abstract deictic (pointing) gestures that point at the opposing ideas as if they each had a distinct physical locus in space [3]. The form of these behaviors is often tied to physical metaphors thus underscoring the close connection between mental processes and the body. For example, the rejection of an idea can be illustrated by a sideways flip of the hand that suggests discarding an object as if an idea was a physical object [4]. Nonverbal behavior is also a reflection of the speaker's mental state.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 303–312, 2013.

Gaze reveals thought processes, blushing suggests shyness and facial expressions, unintentionally or intentionally, convey emotions and attitudes.

The focus of our work is on automatic approaches to generate expressive, life-like nonverbal behavior. We have developed a flexible technique that employs information about the character's mental state and communicative intent to generate nonverbal behavior when that information is available. Otherwise, it uses acoustic, syntactic, semantic, pragmatic and rhetorical analyses of the utterance text and audio to infer the communicative functions (CFs). This includes deriving both the communicative intent of the utterance as well as the underlying emotional and mental state of the speaker. In either case, the CFs are then mapped to nonverbal behaviors, including head movements, facial expressions, gaze and gestures, that are composed and co-articulated into a final performance by a character animation system. In this paper, we give a broad overview of the approach and detail the rhetorical and semantic analyses that detect the CFs. We then report on two evaluation studies using human subjects that assess the consistency of generated performances with the speech.

2 Related Work

Researchers have explored techniques to generate nonverbal behavior, differing in how the models were developed, the degree of automation in the generation process itself and the particular classes of nonverbal behaviors that are handled.

Utterances can be manually annotated to specify what information has to be conveyed nonverbally. Annotations are then automatically mapped to appropriate nonverbal behaviors (e.g. [5,6]). They are also used to communicate knowledge about the character's personality [7] or relationships [8].

Researchers have explored fully automatic generation of specific classes of nonverbal behaviors, using data-driven techniques. This includes models that generate gestures [9] or head movements [10] just by considering prosody, models that learn the mapping between speech text and head movements [11], and models of how speakers' gesture style differ [12,13].

Also, there is work on nonverbal behavior generation using manually constructed models. BEAT automatically generates the speech and associated nonverbal behavior given the text of the utterance and infers rheme and theme to determine intonation and emphasis [14]. The NonVerbal Behavior Generator (NVBG) [15] extends this analysis by inferring the CFs embedded in the surface text (e.g. affirmation, intensification, negation, disfluencies) by using a keywords mapping. When integrated into a larger virtual human architecture, NVBG automatically associates CFs to provided information (emotional state, coping strategy and dialog acts).

BEAT and NVBG can be viewed as the intellectual ancestors to Cerebella. However, the limited analyses that drive those systems also limit both the nature of CFs detected as well as the frequency of detection. The central contribution of this work is the integration of a wider range of analyses, such as acoustic, rhetorical and semantic analyses of the text and audio of the utterance.

This leads to a richer display of behaviors that are more meaningfully related to the utterance.

3 System Overview

Cerebella follows the SAIBA framework guidelines[1]. It takes as input communicative intents and generates a multimodal realization of this intent using the Behavior Markup Language (BML) [16], a high-level XML language that describes a behavior and an execution schedule and provides an abstraction to the animation system. Our system does not make strong assumptions about the provided inputs. If a complete Function Markup Language (FML) input containing detailed information about the mental state and communicative intents is provided, a direct mapping to nonverbal behaviors can be made. However, when only the utterance text and/or audio are given, the system tries to infer the communicative functions (CFs) through several analyses of the speech. We focus here on this last case.

Figure 1 presents an overview of our rule-based system. The central element is the Working Memory (WM) that stores the knowledge of the virtual human. The processing pipeline contains four sequential processes, detailed below.

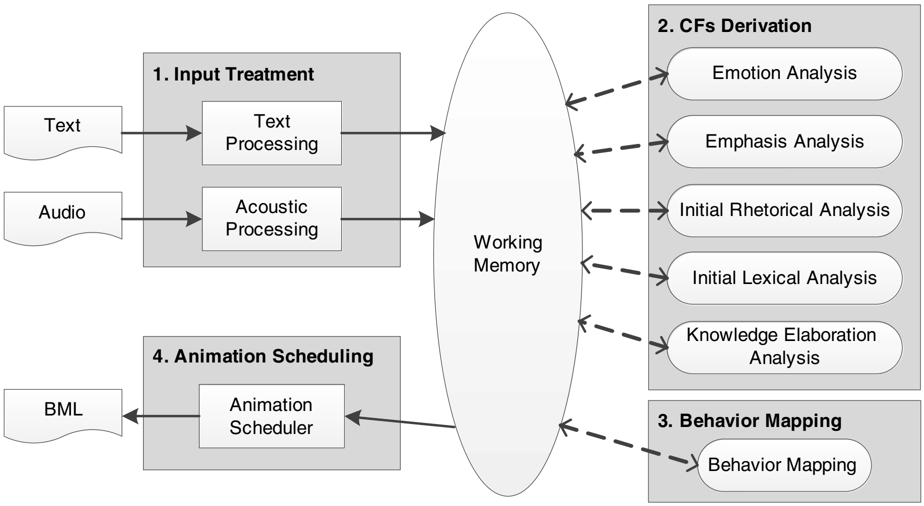


Fig. 1. Overview of Cerebella's processing pipeline when inferring communicative functions from text and audio

3.1 Input Treatment

The input text is tokenized and each token is added to the WM. A natural language parser derives the syntactic structure [17], and each element of the

[1] http://www.mindmakers.org/projects/saiba/wiki

resulting parse tree is added to the WM. A limitation encountered is that most parsers are destined to text and not to partial utterances with their disfluencies and non-grammatical constructions.

An acoustic pipeline processes the audio of the spoken utterance. Two elements are currently derived by our system, overall agitation and word stress, relying on [18].

3.2 Communicative Functions Derivation

This phase detects the CFs present in the utterance. Each analysis consists of rules that match the content of the WM to infer new knowledge. The rules run in parallel so a rule can exploit knowledge inferred by the others. First, some rules take care of the low-level analyses (Emphasis, Emotion, Initial Lexical and Initial Rhetorical Analyses). Then the Knowledge Elaboration Analysis combines the inferred knowledge, leveraging the CFs detected. When no more rules match the content of the WM, the derivation phase ends. Table 1 shows the CFs that Cerebella currently derives, grouped into categories.

Table 1. Communicative Functions

Communicative Function Group	Communicative Function
-	Interrogative, negation, affirmation, emphasis
Rhetorical	Contrast, enumeration, alternative, accumulation, comparison
Intensifier	Strong positive, weak positive, strong negative, weak negative
Quantifier	Nothing, few, many, all, over, approximation
Comparative	Positive, negative
Time	Now, before, after, period
Location	Here, away
Deixis	You, me, we, abstract left, abstract right
Mental state	Cognitive load, emotional states

Emphasis Analysis: uses the word stress knowledge inferred by the Acoustic processing module to detect which words of the sentence are emphasized.

Emotion Analysis: uses the overall agitation level detected by the Acoustic processing module to determine the emotional arousal of the virtual human. We associate tense speech to high arousal, modal speech to mid-level arousal, and lax speech to low arousal.

Initial Lexical Analysis: the initial lexical analysis serves two purposes:

Lexical classification: maps a list of words and phrases to CFs, such as a *deixis_you* with words like "you" or "yourself", or a *quantifier_nothing* with the words "nobody", "none", "nothing" or "never". In addition, the WordNet database [19] is used to expand this knowledge base. In this case, a *quantifier_nothing* is detected whenever the surface text word is hierarchically linked to one of the associated WordNet synsets ("nothing.n.01", "emptiness.n.01").

Abstract/concrete classification: metaphors in language are often linked to metaphorical gestures, for instance when referring to an idea as if it were a concrete object, allowing to discard it with the flip of a hand [3]. To assist in this analysis, the system uses WordNet to annotate whether the nouns of the sentence refer to concrete or abstract concepts.

Initial Rhetorical Analysis: this analysis detects the CFs related to the rhetorical structure of the sentence. We rely on two assumptions to perform this analysis. First, rhetorical relations can be computed without a complete semantic analysis of the sentences (e.g. [20]). Second, such structural analyses work well enough at the utterance level to support nonverbal generation. The rules only use the knowledge asserted by the Text Processing module to detect the rhetorical constructions and their associated CFs. We detail here some of the rules that detect contrast, comparative and comparison. Similar rules allow detecting other rhetorical constructions such as enumeration, addition and alternative.

Comparative and Comparison: a *comparative_positive* function is associated to the detection of "more" followed by a noun phrase, or to adjectives syntactically tagged as comparatives (e.g. "better", "smaller", "easier"). Moreover, whenever a *comparative* is followed by the word "than", a *rhetorical_comparison* is detected. For example, the detection of a *comparative_positive* on the beginning of the phrase "more interesting than before" will further lead to the association of a *rhetorical_comparison* to the whole phrase.

Contrast: a *rhetorical_contrast* function is detected, for example, by the following rule: word/expressions such as "but", "however", "unlike", ... surrounded by two part-of-speech that belong to the same syntactic category. When processing the sentence "This is really more interesting than before but I can only afford around 50 dollars", the previous rule will match and detect a *rhetorical_contrast* between the beginning of the sentence ("This is really more interesting than before") and the end ("I can only afford around 50 dollars").

Knowledge Elaboration Analysis: tests the knowledge previously asserted by the different analyzers and deepens, alters or removes the knowledge whenever required. We identify four purposes:

Combining CFs: together requires verifying that the global meaning is coherent, since the CF of separate elements may be completely different when taken together. For example, in the expression "absolutely not", "absolutely" is associated to an *intensifier_positive* function (by the lexical classification process), and "not" to a *negative* function. The same issue can also be seen in more elaborated inferred knowledge. For example, in the expression "absolutely not more interesting", a *comparative_positive* function ("more interesting") and an *intensifier_negative* function are detected. That would generate separate and inappropriate gestures. The benefit of using a rule-base system is that it simplifies the combinatorics, e.g. positive or negative valence can be associated to different type of knowledge elements, and there are multiple ways of combining them. We do not want to explicitly enumerate all combinations but rather rely on rule-based forward-chaining.

Solving Conflicts: it sometimes occurs that functions conflict with each other. For example, a *comparative* as well as an *emphasis* may be associated to the same word. So as part of the function derivation phase, each function inferred is assigned with a priority based on its CF and whether the words it spans are emphasized. Then these priorities are used to resolve conflicts between overlapping functions with lower priorities being dropped.

Semantic Disambiguation: the initial lexical analysis is sometimes not sufficient to determine the actual meaning of a word or phrase. For example, the *quantifier_approximation* is associated to the word "about" each times it is encountered. However, this association can be wrong, such as in "I was thinking about you". Some rules, by combining semantic and syntactic information, help distinguish those cases, for example by testing that the element associated to the supposed *quantifier* is a number.

Expanding the CF: when detected, a CF is associated to the words matching the detection pattern of the rule. In some cases, the matched words may not span the full phrase that realizes that CF. For example, a *time_before* function is associated to the word "ago" in the expression "two years ago", but the function should cover the whole expression to generate synchronized gestures. Therefore, Cerebella contains a set of generic rules that span the CF over groups of words.

3.3 Behavior Mapping

The CFs derived during the previous phase are mapped to a set of alternative sequences of behavioral types to generate a schedule of multimodal behavior. For example, a *rhetorical_contrast* function might be realized by a synchronized tilting of the head and appropriate gesture. The alternatives allow variability in the character's behavior from one utterance to the next, as well as specialization by character. For example, the agitation state derived from the audio affects this mapping. Characters in the low agitation state (sad or lethargic) are biased to move heads from side to side instead of front to back. Highly agitated characters (angry or energetic) emphasize points using behaviors that include a beat rather than just subtler eyebrow raises.

3.4 Animation Scheduling

The multimodal nonverbal behaviors are mapped to the BML language. Behaviors that can be specified include head movements, gazing, blinking, saccadic eye movements, gesturing, facial expressions and speech. Behaviors are specified with start and end times such that they correspond to word starts or endings, or other behaviors when they are part of a sequence. Finally, the Smart Body animation system [21] interprets these high-level instructions to synthesize the final motion.

3.5 Knowledge

The knowledge used in Cerebella comes from diverse sources. The CFs derivation phase is currently driven by handcrafted rules and associated internal and external (specifically WordNet) databases, but more automatic approaches are being explored, such as using a learning-based rhetorical parser (SPADE [22]).

The knowledge used in the system represents a multi-year effort. Initially, an extensive literature review of the research on nonverbal behavior was undertaken. This initiated the design of rules encoding the function derivation and behavior mapping rules. Also, videos of real human face-to-face interactions have been annotated and analyzed to verify the rule knowledge. This annotation and analysis was critical because existing literature says little about dynamics of behaviors. We characterize this approach as a *expert knowledge plus semi-automated analysis* approach. More recently, pure data-driven machine learning techniques have been used as a way to validate the features used in the rules and to learn the mapping between features of an utterance and nonverbal behaviors [11].

4 Studies

The baseline hypothesis behind the inferencing that takes place during the CFs derivation process is that it will lead to gestures that will convey the meaning of the functions that are inferred. Here we conducted two studies to test this hypothesis.

4.1 First Study

This study tests that the gestures generated by Cerebella convey the same meaning as the speech. We used 9 sentences containing CFs presented in this paper (time, comparative, location and quantity) to generate 9 virtual human video performances. The sound was removed from videos, leaving nonverbal behaviors as the sole indicator of meaning. 34 native English participants (14 female) completed the study via Amazon Mechanical Turk[2]. They could watch each video as many times as they wanted then had to select the sentence that matched the virtual human gestures in a set. Proposed choices included the original sentence as well as derivations created by reversing the original functional class(es). For example, the sentence "It is said that Spanish is much easier to learn than French" was derived into "It is said that French is less easy to learn than Spanish". This sentence tests the association of the gesture ("a two-hand gap that increases") to a *comparative_positive* function as opposed to a *comparative_negative* function. A choice of the original sentence implies a closer match between what the gesture and the original sentence convey, thereby helping to validate the CFs detection and behavior mapping used in our system.

Figure 2 shows the percentage of recognition of the functional classes. The red line marks the recognition rate that would be obtained by randomly selecting the answers. The participants were globally able to retrieve the original sentence by using the associate gesture (overall recognition percentage is above 50%). This

[2] `http://aws.amazon.com/mturk/`

is particularly true for most of the functional classes (with a recognition score between 67% and 85%), except when required to associate the classic "oscillating bowl" (described in [4]) to a *quantity_approximation* instead of a *quantity_few* (score=52.9%).

4.2 Second Study

This study evaluates the appropriateness of gestures regarding the content of the speech. We created 11 sets of 3 virtual human performances that were identical except for gestures. The first performance was accompanied by gestures generated by Cerebella (appropriate condition). The second one used gestures conveying the opposite meaning of the sentence and was generated by reversing the functional intents detected in the appropriate performance (opposite condition). The third one replaced the appropriate gestures by randomly selected gestures (random condition). 46 (26 female) native English speakers completed the study via Amazon Mechanical Turk. They had to watch the 3 videos and order them from 1 ("most consistent with the speech") to 3 ("less consistent with the speech").

Figure 3 shows the frequency with which each performance was ranked as the first choice. Performances generated by Cerebella are rated as the more consistent with the speech (f=0.57), followed by the random ones (f=0.24) and the opposite ones (f=0.19). A one-way ANOVA was conducted to compare the group effects between the different performances. Across the group a significant effect could be observed ($F(2,135) = 79.96$, $p < .0001$). Post-hoc comparisons using the Tukey HSD ($p < .01$) indicate that Cerebella's performance frequency is significantly higher than the two other ones, but no significant difference can be observed between the random and opposite conditions. However, the opposite performance is significantly the most frequently rated as second choice (f=0.43, $p < 0.01$).

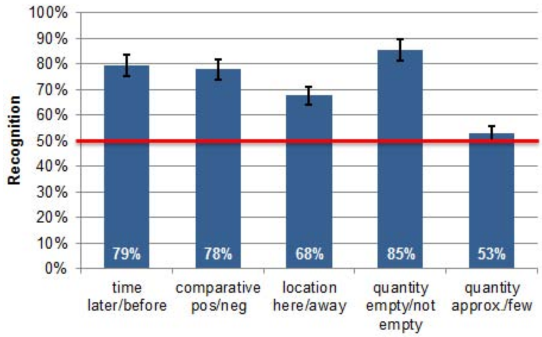


Fig. 2. First study results

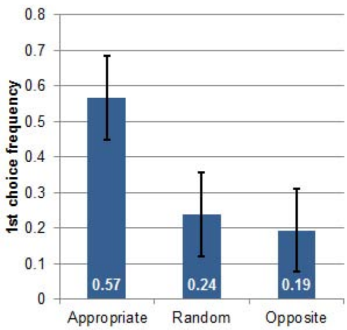


Fig. 3. Second study results

5 Discussion

Cerebella is an automatic approach to generate expressive, life-like nonverbal behavior. When available, it uses information about the character's CFs, including mental states and communicative intent, to generate behavior. Otherwise, it tries to infer CFs that underlie the input text and audio. Our system builds on previous works' approach [14,15]. Acoustic, syntactic, semantic and rhetorical analyses of the utterance are designed to expand the CFs that can be detected as well as improve the accuracy of this detection.

As noted before, nonverbal behaviors express meaning through their form and dynamics. By inferring and exploiting this CF, generated nonverbal behavior ideally reflects that CF and even convey it by themselves. This paper presented two studies to corroborate this statement. Beyond this baseline hypothesis tested in this work, we plan to go on to assess whether the virtual human's gestures influence relational and cognitive factors including attitudes about the speaker, persuasiveness and recall.

While deeper and more elaborate analyses allow inferring and conveying CFs present in the sentence text and audio, this method encounters a particular limitation. As noted previously, nonverbal behavior can stand in a range of relations to the dialog. Here, the automatic generation of nonverbal behavior is limited in the range of CFs that can be inferred from the speech utterance only. This limitation is shared by all techniques that aim at automatically generating nonverbal behavior using speech. This can be overcome whether by taking as input those additional CFs or whether by integrating complex cognitive processes to generate them.

Moving forward with the system itself, one of the key issues will be to maintain a model of the relation between the CFs detected over the course of an interaction. For example, a speaker's gestures can physically locate the elements of the discourse and use deictics to refer back to them [4]. Additionally, nonverbal behaviors are determined by culture, gender, personality, attitudes as well as the context in which the communication takes place [23]. Fortunately, the use of a rule-based system combined to the staged approach we have taken will allow us to easily integrate new sources of input and broaden the range of CFs inferred and conveyed.

References

1. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica 1, 49–98 (1969)
2. Kendon, A.: Language and gesture: Unity or duality. In: McNeill, D. (ed.) Language and Gesture. Language, culture & cognition, vol. 2, pp. 47–63. Cambridge University Press (2000)
3. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
4. Calbris, G.: Elements of Meaning in Gesture. John Benjamins Publishing (November 2011)
5. Kopp, S., Wachsmuth, I.: Model-based animation of co-verbal gesture. In: Proceedings of Computer Animation, pp. 252–257 (2002)
6. Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C.: Speaking with hands: creating animated conversational characters from recordings of human performance. In: ACM SIGGRAPH 2004 Papers, SIGGRAPH 2004, pp. 506–513. ACM, New York (2004)

7. Mancini, M., Pelachaud, C.: Generating distinctive behavior for embodied conversational agents. Journal on Multimodal User Interfaces 3(4), 249–261 (2009)
8. Bickmore, T.: Relational Agents: Effecting Change through Human-Computer Relationships. PhD thesis, Massachusetts Institute of Technology (2003)
9. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. In: ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010, pp. 124:1–124:11. ACM, New York (2010)
10. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1075–1086 (2007)
11. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, vol. 1, pp. 289–296 (2009)
12. Kopp, S., Bergmann, K.: Individualized gesture production in embodied conversational agents. In: Zacarias, M., de Oliveira, J.V. (eds.) Human-Computer Interaction. SCI, vol. 396, pp. 287–302. Springer, Heidelberg (2012)
13. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics (TOG) 27(1), 5 (2008)
14. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 477–486 (2001)
15. Lee, J., Marsella, S.C.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)
16. Kopp, S., Krenn, B., Marsella, S.C., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
17. Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, pp. 132–139 (2000)
18. Scherer, S., Kane, J., Gobl, C., Schwenker, F.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Computer Speech and Language 27(1), 263–287 (2013)
19. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
20. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press (2000)
21. Thiebaux, M., Marsella, S., Marshall, A.N., Kallmann, M.: SmartBody: behavior realization for embodied conversational agents. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2008, vol. 1, pp. 151–158. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2008)
22. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 149–156 (2003)
23. Burgoon, J.K., Guerrero, L.K., Floyd, K.: Nonverbal Communication. Allyn & Bacon (2009)

Conceptualizing Social Power for Agents

Gonçalo Pereira, Rui Prada, and Pedro A. Santos

INESC-ID and Instituto Superior Técnico and Technical University of Lisbon
Avenida Professor Cavaco Silva, Porto Salvo, Portugal
goncalo.pereira@gaips.inesc-id.pt, rui.prada@ist.utl.pt,
pasantos@math.ist.utl.pt
http://gaips.inesc-id.pt/

Abstract. One of the most pervasive concepts in human interactions is social power since many social situations entail disputes of social power. These disputes are power games and range from simple personal reasoning to the exercise of specific power strategies, which enhance or assert one's power. Recognizing the importance of such interactions and how they can enhance autonomous agents' socially intelligent behaviors, we present a formalization of the fundamental bases of power and conceptualize the diverse forces that should underlie an agent's deliberative decision process. Different bases of power stem from diverse factors and have particular dynamics and effects. The objective of this work is to establish a theoretical basis for social intelligent agents capable of both being aware of and manipulating social power.

Keywords: social power, autonomous agents, social intelligence.

1 Introduction

The motivation for studying social power, relates to its ability to act as a *social heuristic* [10] in many social situations such as friends' interactions [10], organizations [15] or even laboratory experiments [14] among others. Consequently, the behavior effects of power are broad, extending to social processes such as coordination, delegation, cooperation, hierarchy formation, alliance formation, resources allocation, conflict resolution and negotiation [5,10]. Given the widespread impact of power in people's attitude and behavior, it is fundamental to understand and emulate such power-based social dynamics in multi-agent systems in order to build socially intelligent agents.

Even though power has been previously explored as a social heuristic for agent behavior [1,3,19,18], these approaches have shortcomings when considered in the context of social intelligence for agent believability, namely inter-agent [13] and agent-human interactions [16]. Current approaches do not take into account the different bases of power as established by French and Raven [7] the particular dynamics and the contrasting effects associated with each one. Modeling the different bases of social power is crucial to simulate the complex processes behind social power and its ubiquitous influence in social interactions. For instance, even in a relatively simple situation where there are just two agents,

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 313–324, 2013.

playing the role of father and son, and a single decision, the son arriving home at the ordered time, many different types of power are at play. If one wants to take into consideration the possibility of power related to reward, punishment, legitimacy, rebel tendencies and love, the current models are too simplistic.

Our objective is to define a model of social power that can cover all these possibilities. To address this we propose a conceptual framework to support agent perception, reasoning and intelligent use of social power aimed at multi-agent and agent-human interactions. The framework was developed from fundamental concepts grounded in the seminal work from social psychology of French and Raven [7] addressing the cognitive ingredients of social power. By using those concepts, we argue that a broader range of social settings can be more accurately modeled in agent simulations through power-based reasoning and interaction.

This document is structured as follows. In Related Work we review previous contributions regarding power in multiagent systems. Next, we present our framework for social power aware agents, followed by an example scenario of its application. Finally we offer some conclusions and future directions.

2 Related Work

The subject of social power, namely the representation of power and the formalization of the associated dynamics have been previously researched from several approaches. The fundamental difference between them is the origin of the represented power, i.e. the main focus of the approach being described.

One approach to power is based on autonomy and is rooted in an agent's capability of pursuing its goals without the intervention of other agents. This approach's central work by Hexmoor [9] addresses "absolute autonomy" which is described as the measure of an agent's liberties (internal and external) over preferences. In this model an agent's power is modeled based on liberties and preferences. Liberties express freedom/inhibition forces regarding an agent decision and there are two types of liberties: endogenous (e.g. values and emotions) and exogenous (e.g. physical context limitations). The preferences affect the liberties forces according to an agent's characteristics: individual rationality (prefers individual welfare) or social rationality (prefers social welfare).

Many works are deeply rooted in the seminal work on Dependence Theory by Sichman and Conte [18] where a taxonomy of dependence and the fundamental concepts for agent reasoning in groups is supported through the formation of dependence networks. In [6] Castelfranchi relates the previous definitions of Dependence Theory to the concept of autonomy, and supports it on two subtypes, practical and deontic autonomy. Based on these two constituents of autonomy, in [3] personal power is grounded on an agent's individual capabilities and social power is formalized as originating from dependence relations of an individual regarding another's individual powers. In [2] dependence is related to utility based Decision Theory in order to reason about power in multi-agent plans. Power is conceptualized based on the possible costs and damages expected from a given dependence situation to the involved agents. These values are calculated as differences of utility between an intended joint plan and a possible alternative plan from one of the agents. In this work power is defined as the capacity to harm another agent, i.e. perform a plan that can lower the utility of the other.

Another approach is normative and emphasizes the part that norms and roles play into a group's power structures and individual powers. The focus is on the restrictions an individual experiences in a given group, and also the benefits it collects from it. The work of López [12] is central to this approach and addresses the impact that power can have on agent behavior regarding the social processes of agent membership to a normative society, individual norm adoption and inter-agent goal delegation. At the core of these processes are the powers agents have due to their roles and personal capabilities. In [12] a taxonomy of powers is presented and agent powers are categorized under "Circumstancial Powers" (facilitation, illegal coercive, exchange, reciprocation, support) and "Institutional Powers" (legal, legal benefit, legal preventive, legal punishment, legal reward).

In [8] power is regarded as an exchangeable natural resource in the agents' environment where it replenishes/decays automatically. For an agent to influence another, it must transfer a given amount of power to it. In such a situation the influenced agent can resist to the influencer by transferring a higher amount of power back.

2.1 Discussion

The autonomy perspective heavily constraints the behavioral diversity of agents by abstracting all the influencing factors of an agent's power to four values. Especially the liberties value abstract components of agent's decisions which should be taken into account separately, for example different norms. The perspective of power as an exchangeable resource also largely disregards the behavioral aspects of power in which we are interested in. For example the perspective of power as resource that is spent or depleted upon usage when in fact a power (e.g. expert knowledge) might be used repeatedly without depletion.

The Dependence Theory approach relates more to agent reasoning about its individual ability or capacity of performing actions independently. However, social power is based on more than dependence. There is power grounded on relational factors such as attraction or credibility. Consider a situation where a music band fan dresses like its members because it wants to be like them. This situation cannot be represented in a context of pure dependence.

Finally, the normative approach connects power with norms and agent membership to a normative society so tightly that it is hard to model some situations. For example in many cases an agent society is defined a priory, e.g. the culture where we are born, no deliberation needed. Another difficulty relates to the taxonomy of powers which seems to limit power to specifically two types of context, circumstantial and institutional. An indication of this strict separation are the powers "illegal coercive power" and "legal punishment power". We argue that social powers are present in all contexts and they can be identified, used or manipulated in a given situation.

3 Power Based Agents

Our main goal is to create an agent architecture with social intelligence capabilities regarding social power. In this work our focus is on the core concepts which make such an architecture possible. Our contribution is two-fold, first a psychology inspired conceptualization of the different bases of power to enhance agents with the capability of identifying powers and perform situational analysis regarding power. Second, the description of the fundamental mechanisms of a power-based decision and the factors which underlie each base of power in that decision.

3.1 Elements of Social Power

But what is social power? The subject has been researched from different perspectives and there are many definitions. However, the one we follow is an adapted version from Lewin's [11] and Cartwright's [4] definitions of social power: "**Social Power** of A over T regarding a possible change in T is the resultant potential force that A can induce on T towards that change.". This definition captures the essence of power as a potential force that results from the accumulation of a variety of social power components with different sources.

Before specifying the different types of social power and their characteristics we must first introduce the three central elements to such a situation:

Actor (A) the agent which exerts power over the actions of another;
Target (T) the agent whose actions are affected by the Actor's power;
Action (C) the action evaluated by the Target in a given interaction.

3.2 Fundamental Notions

Our *MultiagentSystem* is formalized as the 5-tuple $\langle$*Agents, Actions, Roles, Rights, Relationships*$\rangle$. *Agents* is the set of agents in the multi-agent system, where each contains its own set of beliefs and goals. *Actions* is the set of actions available for the Agents to perform. *Roles* is the set of roles available in the society of Agents. The *Rights* is the set of right / obligation contract relationships. The predicate $Right\_to\_Influence(r_1, r_2, a)$ will be used to specify a mapping, belonging to *Rights*, between the right of role r_1 to influence role r_2 regarding action a. Finally the *Relationships* is the set of social relationships between the agents which will be used to represent information regarding how an agent regards another and can be of different types in the set $RelTypes = \{liking, friendliness, attraction\}$. The predicate $Relationship(a_1, a_2, t)$ will be used to represent that agent a_1 has a relationship of type t with a_2 (the unidirectionality is intentional in order to mimic human relations).

An important notion for the conceptualization of power is context, since every power has an associated context. For example, an university teacher might easily prescribe actions to his master or doctoral students in an academic context due to his superior skill difference, but in a context of personal relationships he cannot use the same power to direct behaviors of others in the same way. The context has, however, different meanings depending on the power base under consideration. For example, for a coercive power, the context is the resources and skills needed to perform that coercive action. When making an implementation of the defined predicates and functions, those contexts must be also implemented, as they are necessary for the predicates and functions to be well-defined.

3.3 Identifying Different Bases of Power

Many social power studies propose a set of bases of power. However, most can be represented by one of the first sets introduced by French and Raven [7]. Their work introduced a differentiation and dynamics of social power grounded on five bases of power: reward, coercive, legitimate, referent and expert. Our work is inspired on these bases of power due to their simplicity, behavior expressive potential and repeated validation over the years. We will formalize the five bases of power with four categories.

Welfare Power. This power is based on the ability of the Actor to mediate some welfare (reward or coercion) to the Target. It is formalized in definition (1): if there are two agents A and T, where A can do an action a which T values (positively or negatively), then we are in a situation where A has *Welfare_Power* over T. In this definition the predicate $\mathit{Values}(T, a)$ represents that action a brings about some benefit or harm to T. Additionally, the predicate $\mathit{Can\_Do}(A, a)$ represents the ability of agent A to perform action a. The force of an instance from this power increases with the value of a and also with the T's believed probability that it will perform a.

$$\exists a \in \mathit{Actions} \;\wedge\; a \neq C \;\wedge\; \{A, T\} \subseteq \mathit{Agents}$$

$$\mathit{Welfare\_Power}(A, T, a) := \mathit{Values}(T, a) \wedge \mathit{Can\_Do}(A, a) \tag{1}$$

This power results from the abstraction of French and Raven's reward and coercive power bases under the *Values* predicate. Regarding the identification of the power an agent has over another the two underlying bases of power simply have symmetric valuations regarding a for T. In the reward case a is desired and in the coercive case a is undesired or avoided. However, even though these power bases have symmetric dynamics regarding their identification they have different effects when the interplay between power bases is considered [7].

A crucial example is that in a case of illegitimate coercion (e.g. bullying) the negative effects of using coercion are exacerbated, while in the illegitimate reward case (e.g. a bribe) it depends upon the personality characteristics (e.g. the relation of the conscientiousness trait with dutifulness) of the Target.

Legitimate Power. The power based on internalized beliefs in the Target regarding the right for the Actor to influence the Target and its obligation to accept that influence. It is formalized in definition (2): if there are two agents A and T with a *Rights* relationship confering role r_1 the right to influence role r_2

towards the compliance of C, where A plays role r_1 and T plays role r_2, then A has *Legitimate_Power* over T regarding C. In this definition the predicate $Plays(X, r)$ represents that the agent X assumes the role r. The force of an instance from this power increases with T's degree of adherence to the role he plays and also with T's believed probability that A will legitimately enforce C.

$$\exists r_1, r_2 \in Roles \ \wedge \ \{A, T\} \subseteq Agents$$

$$\begin{aligned} Legitimate\_Power(A, T, C) := Plays(A, r_1) \wedge Plays(T, r_2) \wedge \\ \wedge \ Right\_to\_Influence(r_1, r_2, C) \end{aligned} \tag{2}$$

In multi-agent systems there are several representations which fit this generic description of internalized values. As presented in the work of Carabella *et al.*[3] there are norms (formal or informal), contracts and commitments.

Referent Power. The power based on the identification of the Actor with the Target. It is formalized in definition (3): if there are two agents A and T, where T acknowledges a relationship of type t with A, then A has referent power over T. In this definition the predicate $Relationship(T, A, t)$ acknowledges the relationship of type $t \in RelTypes$ and from T with A. The predicate $Identifies(T, A, t)$ represents the recognition of the relationship factors of type t between T and A. The valuation is based on factors underlying each type of relationship and it can be positive (e.g. attract, like) or negative (e.g. repulse, dislike). Additionally, notice that the relationship does not need to be bidirectional, a person can acknowledge and establish a "relationship" with another and the later not even know of the first's existence. The force of an instance from this power increases with T's magnitude of identification with A (based to the value of the liking/friendliness/attraction) and also with T's probability of identification with A.

$$\begin{aligned} \{A, T\} \subseteq Agents \ \wedge \ Relationship(T, A, t) \\ Referent\_Power(A, T) := Identifies(T, A, t) \end{aligned} \tag{3}$$

Expert Power. The power based on the perceived skill difference between the Actor and the Target. It is formalized in definition (4): if there are two agents A and T, where A has a higher skill than T regarding the knowledge domain of the interaction C between the agents, then A has *Expert_Power* over T. In this definition the predicate $Higher(X, Y)$ represents that X is higher than Y. The function $topic(X)$ identifies the knowledge domain of an interaction. Additionally, function $skill(X, Y)$ quantifies the skill of X regarding the knowledge domain Y. The force of an instance from this power increases with T's believed skill difference to A regarding $topic(C)$ and also with the T's believed credibility (a probability value) of A regarding $topic(C)$.

$$\{A, T\} \subseteq Agents$$

$$Expert\_Power(A, T, C) := Higher(skill(A, topic(C)), skill(T, topic(C))) \tag{4}$$

3.4 Social Power Decision Mechanism

The social power bases identified in the previous section impact intelligent agents' decisions in the environment. As such we must be able to operationalize them and to do so we consider an environment representation as presented in 3.2 and the basic elements of a social power interaction: Actor(A), Target(T), Action(C). To operationalize an agent decision in an influence attempt situation by A over T regarding C, we define the possible environment outcome for the case in which it decides to perform C as $S_{f,C}$ in (5) and for the case it decides not to perform it as $S_{f,\neg C}$ in (6). Notice that S_i represents the environment state before the decision.

$$S_{f,C} = Do(T, C, S_i) \tag{5}$$

$$S_{f,\neg C} = \neg Do(T, C, S_i) \tag{6}$$

In a given social power interaction, an agent may identify several powers according to the different power bases. This set of powers may contain powers from only one base (e.g. have several punishments) or be composed by powers of several distinct bases. We define the set of identified power bases as *IdentifiedPowers* (IP). Each of these powers is a *Force* (F) exerted by the Actor on the Target, influencing its decision. To operationalize each force we quantify its strength based on a probability and a magnitude according to equation (7). The probability represents the perception of mediation capability (e.g. from history of interaction) and the magnitude captures the strength of the power underlying factor (e.g. value of a coercion or degree of liking).

$$p \in \mathit{IdentifiedPowers}$$

$$\mathit{Force}_p = \mathit{probability}_p * \mathit{magnitude}_p \tag{7}$$

The probability component of the *Force* assumes values between $[0, 1]$. The higher the probability value the stronger the force. The magnitude component of the *Force* assumes values in $\mathbb{R}$ in which positive values represent a positive influence towards the decision $S_{f,C}$ and negative ones towards $S_{f,\neg C}$ (conceptualizing the concept of negative power[17], e.g. from disliking). The total social power force exerted by the Actor over the Target in a given situation is the sum of all the individual forces in the IP, as presented in equation (8).

$$\mathit{social\_power\_force}(T, A, C) = \sum_{p \in IP} \mathit{Force}_p \tag{8}$$

Besides social power there is another indisputable force in agent decisions: utility. We can better expose this parallel force by comparing two situations. First, if a person asks a friend to lend him a cellphone temporarily. It is reasonable that the friend does so given the friendship relation and the low loss of utility he experiences by lending the cellphone temporarily. In a second situation the person now asks the friend to buy him a cellphone. In this case he is also reasonable that the friend refuses given the high loss of utility he would experience by buying the cellphone. Generically utility measures the usefulness of world states regarding the agent's goals. Notice that this can incorporate many other social concepts such as emotions.

In our mechanism we model the utility force according to this comparison of the world states as presented in equation (9). The $utility\_force$ assumes values in $\mathbb{R}$ and increases with the increase of $utility(T, S_{f,C})$ or decrease of $utility(T, S_{f,\neg C})$ and vice versa. Positive values represent beneficial situations (favoring final state $S_{f,C}$), negative values harmful ones, and when 0 indicates indifference.

$$utility\_force(T, C) = utility(T, S_{f,C}) - utility(T, S_{f,\neg C}) \tag{9}$$

Finally we model the agent's decision to either perform C or not ($\neg C$) as a combination of the two major forces identified: social power and utility. To do so we assume a simple resultant force approach in accordance with our definition of power (see 3.1) and represented in equation (10).

$$\begin{aligned} res\_force(T, A, C) &= utility\_force(T, C) \ + social\_power\_force(T, A, C) \\ Decision &= \begin{cases} Do(T, C, S_i), & \text{if } res\_force(T, A, C) > 0 \\ \neg Do(T, C, S_i), & \text{if } res\_force(T, A, C) \leq 0 \end{cases} \end{aligned} \tag{10}$$

If the value is positive then the agent choses C, if not then it choses $\neg C$. Notice that this decision formalization takes into account the resistance that an agent can offer to a given prescription of behavior from another agent [7]. This is present at two distinct levels. First, if utility is negative it represents an opposing force to the social power being exerted. Second, for each $Force$ (from any base of power) the magnitude can reflect negative power, which is also another form of opposing force to the influence attempt. For instance when a person dislikes another this will be represented as a negative force (in case of any influence attempt) from the referent power base.

4 Example Scenario

Consider a situation where a boy is going out with his friends, but before leaving home he is instructed by his father to be at home before midnight. At a certain point after leaving home, and before midnight, he will be faced with the decision to either do as told by his father or defy his wishes. What are the forces at play for the boy's decision? How do we model it under our framework? In this situation the boy is the Target (T), his father the Actor (A) and the Action (C) upon which he must decide is "be at home before midnight".

4.1 Initial Situation

The agent set for this scenario is $Agents = \{boy, father\}$ where $A = father$ and $T = son$. The $Roles$ and $Rights$ for the situation depend on the existing norms for a specific family which can vary a lot. However, for this example we can at least consider one (informal) norm which is frequently adopted in families: children should always obey to their parents. Based on this we have $Roles = \{parent, children\}$ and the rights relationship for the situation $Rights = \{\{parent, children\}\}$ meaning that agents in the role of parent have the $Right\_to\_Influence$ agents in the role children. As for the relationships they are represented by the following set $Relationships = \{\{boy, father, like\}, \{father, boy, like\}\}$ meaning that there is a bi-directional relationship of type "liking" between the boy and father. Finally as for actions we consider the following set of possible actions $Actions = \{WithdrawAllowance, GroundChild, AllowReturnLate, ReturnOnTime\}$ where the action $C = ReturnOnTime$. Finally, the moment before the boy makes his decision is S_i and the moment after he decides is either $S_{f,ReturnOnTime}$ or $S_{f,\neg ReturnOnTime}$.

4.2 Modeling Boy's Decision

Following our conceptual framework, the *boy* starts by analyzing what power categories are at play by using the definitions (1) to (4):

1. $Welfare\_Power(father, boy, WithdrawAllowance)$
2. $Welfare\_Power(father, boy, GroundChild)$
3. $Welfare\_Power(father, boy, AllowReturnLate)$
4. $Legitimate\_Power(father, boy, ReturnOnTime)$
5. $Referent\_Power(father, boy)$

The welfare powers represent the $father$'s ability to punish the *boy* by withdrawing his allowance (1) or ground him (2) on future opportunities to go out with its friends. As for benefits, the $father$ can allow the *boy* to stay out longer (3) next time if it behaves properly this time. The legitimate power (4) represents the $father$'s right to influence the *boy* regarding $ReturnOnTime$, since *boy* plays the role of child and $father$ that of parent. The referent power (5) is based on the son-father relationship between the two agents. Expert Power does not exert any force because there is no relevant skill in this situation.

The *boy*'s decision between $S_{f,ReturnOnTime} = Do(boy, ReturnOnTime, S_i)$ or $S_{f,\neg ReturnOnTime} = \neg Do(boy, ReturnOnTime, S_i)$ then only depends on the actual values of each of these forces. From the perspective of what happens in real life both cases are believable given the appropriate personal characteristics, relationships and beliefs. We will present both cases accordingly.

4.3 Case 1: A Well-behaved Child

If we assume a well-behaved son all the social power forces work towards C, and even though the utility offers resistance, he will probably return home before midnight. In order to illustrate this case let's consider the values in table 1 and the utility values $Utility_{S_{f,C}} = -100$ and $Utility_{S_{f,\neg C}} = 20$. Notice that for space reasons we abbreviated the names of the components. For example, for the power situation $Referent\_Power(father, boy)$, in the table its probability component and value component are represented by P_{RP1} and V_{RP1} respectively.

Table 1. Example values for the identified social powers

Property	P_{WP1}	V_{WP1}	P_{WP2}	V_{WP2}	P_{WP3}	V_{WP3}	P_{LP1}	V_{LP1}	P_{RP1}	V_{RP1}
Value	0.2	100	0.7	20	0.15	50	0.95	100	0.6	40

Based on table 1 we can then calculate the *utility_force* using definition (9):

$$utility\_force(boy, ReturnOnTime) = -100 - 20 = -120$$

From the previous subsection (see 4.2) we can also determine the set $IdentifiedPowers = \{WP1, WP2, WP3, LP1, RP1\}$. Now according to definitions (7) and (8) it is possible to calculate the social power force:

$$\begin{aligned} &social\_power\_force(father, boy, ReturnOnTime) = \\ &\quad = F_{WP1} + F_{WP2} + F_{WP3} + F_{LP1} + F_{RP1} = \\ &\quad = 0.2 * 100 + 0.7 * 20 + 0.15 * 50 + 0.95 * 100 + 0.6 * 40 = 160.5 \end{aligned}$$

Once we have both the $social\_power\_force$ and the $utility\_force$ we can then calculate the res_force and know the decision of the agent *boy* according to (10). Therefore the resultant force value is the following:

$$res\_force(boy, father, ReturnOnTime) = -120 + 160.5 = 40.5$$

Based on the value of the res_force and the condition $res\_force(boy, father, ReturnOnTime) > 0$ from definition (10) the *boy* agent decides for $Do(boy, ReturnOnTime, S_i)$ meaning that it will return home on time, before midnight.

4.4 Case 2: Disobedience in a Well-behaved Child

Now imagine the *boy* is having more fun than he has ever had? In this case it might occur the situation where a well behaved son actually disobeys to the father's command since the utility force can surpass the social power force exerted by his father. This means the utility evaluation of the situation changes drastically. For example, consider the values $Utility_{S_{f,C}} = -100$ and $Utility_{S_{f,\neg C}} = 100$ and table 1 with the same values. Since the values associated with the $social\_power\_force$ did not change, its value remains the same. However, the new utility force is the following:

$$utility\_force(boy, ReturnOnTime) = -100 - 100 = -200$$

The impact of this change in the res_force is given by:

$$res\_force(boy, father, ReturnOnTime) = -200 + 160.5 = -39.5$$

Based on the value of the res_force and the condition $res\_force(boy, father, ReturnOnTime) \leq 0$ from definition (10) the *boy* agent decides for $Do(boy, \neg ReturnOnTime, S_i)$ meaning that it will not return home on time and will stay out late.

4.5 Other Complex Cases

An additional interesting case would be that of the rebellious son, where the legitimate component of the social power force now exerts negative influence since the command from the father actually works as a resistance force to $ReturnOnTime$. In this case, the rebellious son is much more probable to stay out with its friends (do $\neg C$) even in the first case. Notice that this does not mean that he "does not like" his father, a son might like his father and simultaneously exhibit rebellious behavior. Our model enables the simulation of these complex situations.

5 Conclusions and Future Work

In this work we introduced a conceptual framework for agent social intelligence regarding social power awareness. The identification of situations where different sources of power emerge enable agents to take many social influences that an individual has to deal with when making decisions in a social context. The definitions presented are founded in a well established social psychology study and build upon a small but behaviorally expressive set of social powers that enable agents to participate in a wide range of power games as found in human societies. These are the fundamentals of an agent's power assessment. We also introduced the basic mechanisms for an agent's decision making processes including social power influences aiming for behavioral believability which can be employed in diverse agent-based applications.

In future work we will first conceptualize the different effects on power dynamics associated with different power bases, the mechanism used to replicate power strategies utilization and its integration in a theory of mind reasoning. Finally this will be applied to a test scenario where agents can reason and influence other agents or humans in a believable way, considering the distinct social power forces and its effects.

Acknowledgments. Work supported by FCT(INESC-ID multi annual funding) through the PIDDAC Program PEst-OE/EEI/LA0021/2013 and FCT scholarship SFRH/BD/66663/2009.

References

1. Boella, G., Sauro, L., van der Torre, L.: Power and dependence relations in groups of agents. In: Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2004, pp. 246–252 (2004)
2. Brainov, S., Sandholm, T.: Power, dependence and stability in multiagent plans. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence, AAAI 1999/IAAI 1999, pp. 11–16. American Association for Artificial Intelligence, Menlo Park (1999)
3. Carabelea, C., Boissier, O., Castelfranchi, C.: Using social power to enable agents to reason about being part of a group. In: Gleizes, M.-P., Omicini, A., Zambonelli, F. (eds.) ESAW 2004. LNCS (LNAI), vol. 3451, pp. 166–177. Springer, Heidelberg (2005)

4. Cartwright, D.: Power: A neglected variable in social psychology. Studies in Social Power 6, 1–14 (1959)
5. Castelfranchi, C.: Modelling social action for AI agents. Artificial Intelligence 103(1-2), 157–182 (1998)
6. Castelfranchi, C.: Founding agents' "autonomy" on dependence theory. In: ECAI, vol. 1, pp. 353–357 (2000)
7. French Jr., J., Raven, B.: The bases of social power. Studies in Social Power, 150–167 (1959)
8. Hayes, D., Hexmoor, H.: Social power as an exchangeable resource for distributed multi-agent systems. In: Proceedings of the International Symposium on Collaborative Technologies and Systems, pp. 278–281. IEEE Computer Society, Washington, DC (2006)
9. Hexmoor, H.: Absolute model of autonomy and power: Toward group effects. Connection Science 14(4), 323–333 (2002)
10. Keltner, D., Van Kleef, G., Chen, S., Kraus, M.: A reciprocal influence model of social power: Emerging principles and lines of inquiry. Advances in Experimental Social Psychology 40, 151–192 (2008)
11. Lewin, K.: Field theory in social science: selected theoretical papers (Cartwright, D. (ed.)). Harpers (1951)
12. López, F.: Social Power and Norms: Impact on agent behaviour. PhD thesis, University of Southampton (2003)
13. Marsella, S., Pynadath, D.: Modeling influence and theory of mind. In: Artificial Intelligence and the Simulation of Behavior. Joint Symposium on Virtual Social Agents, pp. 199–206 (2005)
14. Milgram, S.: Obedience to authority: an experimental view. Harper & Row (1974)
15. Pfeffer, J.: Power in organizations. Pitman, Marshfield (1981)
16. Prada, R., Paiva, A.: Teaming up humans with autonomous synthetic characters. Artificial Intelligence 173(1), 80–103 (2009)
17. Raven, B.H.: The bases of power: Origins and recent developments. In: Annual Meeting of the American Psychological Association (100th), Washington, DC (August 1992); Raven, B.H., Department of Psychology, UCLA, Los Angeles, CA 90024-1563
18. Sichman, J.S., Conte, R., Demazeau, Y., Castelfranchi, C.: A social reasoning mechanism based on dependence networks. In: Proceedings of 11th European Conference on Artificial Intelligence, pp. 188–192 (1994)
19. Ward, D., Hexmoor, H.: Deception as a means for power among collaborative agents. In: Int. WS on Collaborative Agents: Autonomous Agents for Collaborative Environments, pp. 61–66 (2003)

Social Importance Dynamics: A Model for Culturally-Adaptive Agents

Samuel Mascarenhas[1], Rui Prada[1], Ana Paiva[1], and Gert Jan Hofstede[2]

[1] INESC-ID, Instituto Superior Técnico, Technical University of Lisbon, Av. Prof. Cavaco Silva, 2780-990 Porto Salvo TagusPark, Portugal
{samuel.mascarenhas,rui.prada}@gaips.inesc-id.pt, ana.paiva@inesc-id.pt
[2] Wageningen University, Hollandseweg 1, 6706 KN Wageningen, The Netherlands
gertjan.hofstede@wur.nl

Abstract. The unwritten rules of human cultures greatly affect social behaviour and as such should be considered in the development of socially intelligent agents. So far, there has been a large focus on modeling cultural aspects related to non-verbal behaviour such as gaze or body posture. However, culture also dictates how we perceive and treat others from a relational perspective. Namely, what do we expect from others in different social situations and how much are we willing to do for others as well. In this article we present a culturally configurable model of such social dynamics. The aim is to facilitate the creation of agents with distinct cultural behaviour, which emerges from different parametrisations of the proposed model. The practical application of the model was tested in the development of an agent-based application for intercultural training, in which the model is responsible for driving the socio-cultural behaviour of the virtual agents.

Keywords: Socially and Culturally Competent Virtual Agents, Cognitive Model.

1 Introduction

In this article we describe the SID Model (Social Importance Dynamics Model), which addresses the problem of building autonomous agents that are culturally influenced in the way they perceive and socially interact with others. The proposed model is based on the status-power theory by Kemper [15]. The reason why this theory was chosen was because it considers that the cultural rituals we participate in our everyday lives, from greetings to weddings, are ultimately driven by two behavioural dimensions, which Kemper denominates as status and power. The first refers to the voluntary compliance with the interest of others whereas the latter concerns the involuntary compliance caused by coercion.

The proposed model partly operationalises the behavioural dynamics of Kemper's status, defined as "the acts or means by which the scalar standing, worth, prestige, honor of a person or social position is conveyed in interaction". To avoid confusion with other possible definitions of the word status, we opted to refer to

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 325–338, 2013.

this construct simply as social importance (SI). The model endows agents with a general desire to confer social importance to others when it is duly deserved. This desire motivates the agent to perform, for instance, appropriate greetings or give a direction when asked. Finally, the model affects the process of planning by filtering plans that involve the agent claiming more SI then it has.

To test its expressiveness in creating groups of agents with distinct cultures, the model was first implemented in an agent architecture for embodied agents. Then, the resulting computational architecture was used to create different synthetic cultures of autonomous characters. Interaction with such cultures can then be experienced in a interactive-storytelling application that aims to teach cultural differences on a generic level, taking inspiration from the work done before in real-life role-playing simulations with synthetic cultures [9].

The outline of this article is described as follows. In the next section, some background on culture theory is presented. In section 3 we discuss related work, focusing on models for simulating socio-cultural behaviour. Afterwards, in section 4, the proposed model is presented. In section 5, we illustrate how the model was applied in the development of an intercultural training application. Finally we draw some conclusions and present some future work.

2 Background on Culture

What is culture? Many different attempts have been made to answer this question [16] and as of yet, no consensual definition has been agreed upon. One of the difficulties in defining culture is due to the fact that the concept is used to refer to both concrete aspects of a particular society, namely their artifacts (tools, architecture), as well as more abstract aspects such as shared beliefs about what is right and wrong or what is desirable or undesirable.

Focusing on the abstract aspects of culture, Geert Hosfede defined it as "the collective programming of the mind that distinguishes the members of one group or category of people from another" [8]. To better understand how this "programming" differed across nations, he conducted a large survey on values across several countries. From that study four cultural dimensions were derived: (1) individualism vs collectivism, (2) power distance, (3) uncertainty avoidance and (4) masculinity vs femininity. Later, two additional dimensions were found and added to the theory [10], namely, (5) long-term orientation vs short-term orientation and (6) indulgence vs restraint.

All of these dimensions indicate a set of core differences between national cultures. For instance, if a culture scores high on individualism, it means that its members are more inclined to believe that everyone should be independent and have the same rights. Conversely, members of collectivistic cultures tend to view themselves as part of strongly interdependent groups and are more receptive to the idea that rights should differ across groups. Differently, the power distance dimension reflects how people deal with the distribution of power amongst its members. In cultures with a small power distance, people tend to view others as equal, regardless of their formal status. Conversely, members of cultures with a large power distance treat people with a higher status in a privileged manner.

As described in [8,12] these dimensions are manifested in several aspects of behaviour. As argued in [11], one type of cultural manifestation that is largely relevant for modelling social interaction in agents, is the notion of ritual. Although it has no consensual definition, researchers agree with the general idea that a ritual is a set of actions that are performed mostly for their symbolic value in a manner that is prescribed by the members of the culture. Yet, what exactly is the motivational force behind our participation in cultural rituals? In his status-power theory [15], Kemper argues that rituals are the means by which we signify our relations with others. Ultimately, ritual interaction is driven by the wish to convey the right amount of respect to those who we believe to deserve it, with the right amount being prescribed by shared cultural assumptions. The goal of the model proposed in this paper is to allow the encoding of such assumptions in a manner that agents can more easily adapt their relational behaviour to different cultures.

3 Related Work

There is an increasing interest on representing cultural influences on virtual agents given their importance in human social interaction. Culturally-adaptable agents can be used to facilitate the interaction with users from different cultures as people prefer interacting with an agent when it has a similar cultural background [17]. The development of cultural agents is also an essential effort in the development of agent-based applications for intercultural training such as TLTS [14], ELECT BiLAT [7], or ORIENT [1]. TLTS is a system designed to teach U.S. Soldiers how to speak Arabic dialects and use proper gestures. Similarly, ELECT BiLAT is also designed for U.S. Soldiers, but focuses on teaching them negotiation skills. ORIENT takes a different approach as it simulates a fictional culture of an alien species named Sprytes. By having users trying to save a strange culture in an interactive narrative, the system aimed to promote cross-cultural empathy on a more general level, without focusing on any particular country.

So far, there has been a large focus on addressing cultural differences on directly observable aspects of behaviour. For instance, in the CUBE-G project, a culturally-adaptable model [6,5] was developed that affects the agent's gesture expressivity, usage of pauses, overlapping speech and posture. The developed model is based on Hofstede's dimensional theory [10] and on a large video corpus analysis of conversations held between Japanese and German people. Jan et al. [13] also proposed a model of directly observable behaviour that models aspects such as proxemics, gaze and turn taking.

The work presented in this paper differs from the aforementioned models in the sense that it focus on the representation of cultural influences in the way agents internally construct a social reality by which they determine if a certain behaviour is appropriate or not. In this regard, the Culturally Affected Behaviour (CAB) model [22] allows the representation of explicit links between certain actions and one or more norms from a specific culture. A limitation of the model is that the association between an action and a norm remains the same regardless of the agent who performs it. Differently, in our model the same action can be perceived as appropriate if it is performed by some agents or it can be inappropriate if done by others.

Another model which represents cultural biases in the agent's decision making is the model proposed in [19], in which two of Hofstede's dimensions, namely individualism and power distance are directly used as factors in the agent's goal utility function. In comparison, our model allows a more flexible parametrization of cultural influences that affects not only the deliberation process of the agent but also its perception and planning processes.

Also related to our work is the agent architecture named Thespian [21], an architecture for simulating social behaviour that was used to drive the behaviour of the virtual agents in the Tactical Language Training System [14]. It is able to embed norms in the agent's conversational behaviour through the concept of obligations. These are created when an agent performs a certain action on another agent such as greeting him or asking a question. The other agent becomes aware that there is a social expectation and decides whether to satisfy it by performing an appropriate response or not. As described later, our model is also capable of a similar dynamic in a sense that a greeting or a question are both claims on the agent's social importance which evoke an act of conferral from the other agent. Compared to our model, one limitation of Thespian's obligations is that they require an explicit action in order to evoke a response from the other agent. Sometimes it is the situation itself that implicitly creates an obligation. For instance, a friend's birthday is a situation that automatically creates an obligation for saying happy birthday to her.

Finally, in [18] the notion of ritual was formalised and implemented in an existing agent architecture [4]. In this work, rituals were modelled as a particular type of shared goal, which requires a specific sequence of symbolic actions in order to be achieved. One limitation of the model is that, being a goal on their own, agents are motivated to participate on rituals for the sake of participation and not to signify their level of relationship with others. While not modelling rituals explicitly, the model presented on this paper can be parametrised to have agents engage in ritualistic activities with one another, such as greeting or having a toast. But more importantly, their decision to participate or not in such interactions is based and on how agents perceive each other, from a relational perspective.

4 The Social Importance Dynamics Model

As previously mentioned, the SID Model is strongly based on the status-power theory by Kemper [15]. More specifically, the model aims to operationalize Kemper's notion of status, which he argues to be, together with power, the ultimate motivational forces in relational activity. In his theory, status, which we will refer to as social importance (SI), represents how much are we willing to act in the interest of another social entity, taking into account their needs and wishes above our own. Power, on the other hand, represents the negative side of relational behaviour as it refers to our ability to coerce others to act in our favour. For instance, it is possible to drastically increase our power by holding a loaded weapon against a person. Given that our current goal is to model culturally-appropriate behaviour, the proposed SID model focuses only on status, assuming that neither agents nor users will attempt to coerce or manipulate others.

There are several factors that will influence how much SI we attribute to others, such as:

– **Interpersonal Relation** - How much a person likes or dislikes another greatly affects SI. Best friends will usually attribute a high SI to each other. On the other hand, disliking someone lowers their SI.

– **Group Membership and Role** - In-group favoritism is a well researched phenomenon in social psychology. Humans have a strong need to form cohesive groups in which they trust so being part of the same group increases one's SI from the perspective of the other members. Moreover, the importance of the role taken in the group is also directly correlated with SI.

– **Task Interdependence** - To require the help of someone to achieve an important goal is also a factor that raises SI. For instance, if a group of people gets stranded in a deserted island and only one of them knows how to hunt wild animals, then his SI will be significantly raised because of it.

– **Personal Attributes** - Societies regard certain attributes as a sign of SI. These can be physical such as height and weight or non-physical such as richness or intelligence.

– **Conformity to Standards** - When someone acts against our standards of conduct, it is normal to lower their SI in our mind. The amount lowered naturally depends on the gravity of the misbehavior. Oppositely, when others match our standards we automatically confer SI to them.

All of the above factors greatly affect our willingness to act in the interest of another. Moreover, our cultural background plays a major role in determining which factors are more important than others. For instance, in cultures that are more collectivistic, group membership will have a higher weight than it does in individualistic cultures. However, it is important to note that this effect is not entirely deterministic in the sense that there will always be deviation amongst the individuals from a certain culture due to factors such as personality and different life experiences.

In terms of how it affects our behaviour, SI works both as a restraining factor and as a motivational source. The restraining aspect takes place when considering how much it is possible to have others acting in our interest, as that will largely depend on the amount of SI they attribute to us. If our action claims more SI than what we have, the other person will likely not comply the way we would like and it is possible that our SI becomes lower in their mind.

Social importance is also a motivational factor, as when someone performs a claim to another, it creates a desire on that other person to do a conferral in response. Such desire comes from the need to reinforce or improve the relation between the two, with different acts conferring different amounts of SI. For instance, consider the difference between explaining directions to someone who is lost and accompanying the person to the desired destination.

Asides from the conferrals that are done in response to explicit claims, it is also possible that the situation itself implicitly evokes a conferral. For instance, the situation of meeting a friend implicitly evokes a greeting action as a conferral act, with different types of greetings conferring different amounts of importance.

The aim of the SID model is to increase the social intelligence of regular BDI agents by integrating the aforementioned notions in their reasoning and behavior. As shown in Figure 1, the model is based on the following three elements, which can have different cultural parametrisations: (1) SI Attribution Rules, (2) SI Conferrals, and (3) SI Claims. Each of these elements will influence a different process of the agent.

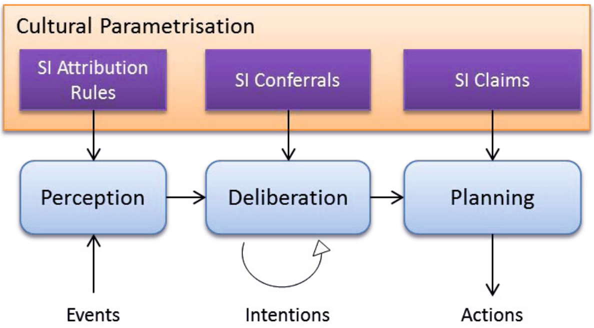


Fig. 1. General Diagram of the SID Model

4.1 Impact on Perception

When modelling a social interaction scenario, agents must determine how much social importance they should attribute to one another. In the case of humans, this knowledge is progressively ingrained into our minds, first from our parents then from the other members of our culture. The purpose of the SI Atrribution Rules is to encode such knowledge. Formally, a SI Attribution Rule is defined as a tuple $< T, A, V >$ where:

- T - Specifies the target of the rule.
- A - Corresponds to a list of conditions that specify when the rule is activated.
- V - The amount of SI the target of the rule gains/loses.

Table 1. SI Attribution Rules - Examples

T	**A**	**V**
x	isPerson(x)	+20
x	isCoWorker(x)	+10
x	isCloseFriend(x)	+20

For illustration purposes, consider a group of agents that share the simple set of attribution rules described in Table 1 for the target 'x' (the values chosen are merely illustrative). In this scenario, agents attribute the most importance to other agents that are their friends, as the combined result of simultaneously activating the first and third rule. Note that the first rule applies to every agent. Its purpose is to encode a default value of SI to others.

When another agent is encountered, his initial SI will be determined by the sum of all SI Attribution Rules that are activated when considering that agent as the rule's target. Then, each time the agent updates its beliefs, the SI of all other agents is updated by checking if the belief change results in the activation or deactivation of any existing rules.

Finally, agents are also able to infer how much SI do they have in the perspective of others. This is done by using a Theory of Mind mechanism, described in [2], which creates a mental model of the other agents' beliefs, including the amount of SI they attribute to others. Then, the same process that is used for updating the agent's own SI values is applied in these mental models of others. The only difference is that the agent switches to the perspective of the other agents when doing so, assuming they have an identical set of SI Attribution Rules. When other agents do have the same cultural profile, then the inference performed will be accurate. However, this will not be the case if other agents have a different cultural configuration. This resembles what happens in intercultural communication, where a significant mismatch between SI attribution rules can potentially harm the success of the social interaction between agents.

4.2 Impact on Deliberation

The deliberation cycle of a typical BDI agent starts with the generation of possible goals to pursue, followed by the selection of the goal with the highest utility and the creation of an intention to achieve such goal. In his theory, Kemper argues that there are two main motivations concerning status, namely: (1) obtaining it from others and (2) conferring it to others when it is appropriate. Our model focuses on the latter, by endowing agents with a general desire to perform acts to signify the amount of SI they have ascribed to others. As stated by Kemper, "Culture specifies what concrete acts and to what degree they signify status-conferral." [15] The aim of the SI Conferrals of our model is precisely to encode such knowledge. Formally, a SI Conferral is defined as a tuple $< C, A, T, V >$ where:

- C - Is a set of preconditions that dictate the context in which the conferral is expected.
- A - Is the name of the action that is perceived as a social importance conferral.
- T - Corresponds to the target agent to which the conferral applies, which is usually the same target of the action but not always. For instance, consider a person who asks you to close a door. The conferral act would be to close the door but the conferral's target would be the person who made the request.
- V - Specifies the amount of social importance conferred by the action.

Table 2. SI Conferrals - Examples

A	**V**
offer-surprise-dinner	40
say-happy-birthday	20
explain-direction	10

Some examples of SI Conferrals are described in Table 2. In these examples C was not represented for simplicity reasons and T corresponds to the same target of the action. The first two examples correspond to two different conferrals that are usually given when it is someone's birthday. While to some it is enough to just say a congratulation message there are others to whom we want to do more such as organizing a surprise dinner party. The third conferral exemplifies a response to a person that asked for a direction (a very low SI claim).

SI Conferrals affect the deliberative process of the agent in the following manner. Firstly, for each SI Conferral a corresponding goal to perform the conferral act is automatically added to the agent. Each of these goals will become active when all the conditions specified in C are true and if T has an equal or superior SI than V. When a conferral goal becomes active, its utility is determined in a straightforward manner: it is linearly proportional to the amount of SI it confers. The rationale is that agents want to confer as much as they think the other agent deserves but not more. Keep in mind that the agent will still choose regular non-conferral goals provided they have a higher utility. For instance, consider a situation where a person invites a close friend to a party. The friend might decline the invitation because he needs to work late on a project for his company and not because the host has not enough SI.

4.3 Impact on Planning

After committing to an intention, agents must search for a valid plan of actions in order to achieve it. When the aim is to simulate social scenarios, it is often the case that agents need or can greatly benefit from the help of others, similar to what happens with humans which are constantly interacting with one another.

Cultural conventions establish what seems reasonable to ask of another and what is not. The purpose of the SI Claims in our proposed model is to endow the agent with knowledge about such conventions, so he can plan more successfully in a particular socio-cultural context. Formally, a SI Claim is defined as a tuple $< A, T, V >$ where:

- A - Is the name of the action that is perceived as a claim for social importance.
- T - Is the target of the claim. Usually it is the same target of the action but not always. For instance, consider the claim of entering a house that is not your own. The target of the claim would be the owner of the house, not the house itself.
- V - Is the amount of social importance the action is claiming.

Table 3. SI Claims - Examples

A	V
ask-direction	10
ask-for-ride	40
offer-surprise-dinner	40

Table 3 provides some examples of possible SI claims, in which T is the same as the corresponding action's target. The first two examples are possible actions an agent might consider when building a plan to go to an unknown destination. Considering the attribution rules specified in Table 1, the agent would have enough SI to ask a direction to any other agent that is a person. However, the same does not apply in the case of asking for a ride. An agent who would perform these actions to a stranger would be claiming more SI than it has and most likely the stranger would not be willing to abide by the request.

Not only agents need to be concerned about their SI in the perspective of others when performing requests, they also should be concerned when conferring SI to others. The last claim example from Table 3 exemplifies this with an action that is simultaneously a SI conferral and a SI claim. This allows us to model situations in which people would like to perform an action that would confer more or less SI but choose not to because they themselves lack SI in the perspective of the other person.

The agent's planning process is affected by the SI Claims in the following manner. After a valid plan to achieve the agent's current intention is created, the planner will determine if any of the actions corresponds to an SI-Claim. For each of these actions, the agent will determine if the value of the claim is superior to the inferred amount of SI ascribed by the target agent. If so, the action is removed from the plan and an alternative is searched.

4.4 Modelling Cultural Influences

After establishing how the SID Model affects the main cognitive processes of a general BDI agent, we now return to the challenge of how it is possible to use the model to create agents with different cultural profiles. A possible approach is to manually configure the values assigned to all the different elements of the model (attribution rules, claims, and conferrals) in a way that the resulting behaviour of a group of agents reflects the behaviour found in a particular culture. The main disadvantage in this approach is that it cannot be easily adapted to model several distinct cultures. As such, we propose a more flexible approach that is based on the association between a SI component and a *Cultural Influence*. The latter is formally defined as a pair $< D, M >$ where: D corresponds to the name of a cultural dimension (e.g Individualism), and M is a multiplier, either positive or negative, that is applied to modify the value V of the associated SI component. This is done by using the following equation, in which $Score(D)$ corresponds to the score associated to the dimension D in the agent's cultural profile, ranging from 0 to 100:

$$V_{modified} = V_{initial} + |V_{initial}| * M * \frac{Score(D)}{100} \tag{1}$$

The aim of a *Cultural Influence* is to be able to represent general tendencies of behaviour indicated by different scores on the cultural dimensions model. For instance, a defining characteristic of cultures that score high on Power Distance is that people who are older are treated in a more respectful and privileged manner [10]. Such tendency can be modelled in the SID Model by the following SI Attribution Rule, $< T = x, A = isElder(x), V = \alpha >$, associated with the following Cultural Influence, $< C = PowerDistance, M = \beta >$, with $\beta > 0$. The effect of this association is that the amount of SI an agent A will attribute to an agent B that is an elder will linearly increase with the score attributed to the Power Distance dimension.

The advantage of using this approach is that it becomes possible to adapt the agent's cultural behaviour just by changing the scores associated to their cultural dimensions. Still, the values of α and β have to be fine-tuned in relation to the other SI elements defined. However, note that this type of cultural influence, unlike a specific norm, is not situation-specific i.e. it will potentially affect every decision of an agent in which an elder person is involved.

5 Creating Cultures with the SID Model

In order to test the applicability of the SID model, it was implemented in an existent architecture for virtual agents [3] that follows the BDI paradigm at its core. Afterwards, the resulting architecture has been applied in the development of an intercultural training application [20]. The application focuses on training generic aspects of cultural behaviour that can distinguish a broad set of cultures, taking inspiration from the work in conducting role-playing simulations for generic intercultural training [9].

To promote engagement, the application follows an interactive storytelling approach, where the user plays an active role on a story that takes place in different fictional countries. In each of these countries, the user must solve practical problems such as finding directions to a hotel. Solving these problems requires the user to engage in social interaction with small groups of autonomous characters that will behave in a culturally-distinct manner, particularly in the way they treat users and respond to their actions.

The SID model drives the different cultural behaviour of the characters. To give an example, there is a situation that takes place in a museum in which the user is looking to find the supervisor of a wild park, to ask his permission for a visit. The scene starts with the user encountering the supervisor's assistant who he has met before. After greeting each other, the assistant indicates to the user who the supervisor is. The user can then decide to directly approach the supervisor who is looking at one of the museum exhibitions and ask his permission.

Fig. 2. Example of a cultural difference in the museum scene

Figure 2 shows how the response of the supervisor is culturally different when the user requests his help directly. On the left side, the supervisor was specified with an extremely small power distance, i.e. $PDI = 0$. On the right side, the other extreme $PDI = 100$ was applied. Table 4 shows the SI Attribution rules applied in this situation. Because of the second rule, the SI of the elderly supervisor is much higher in the large power distance culture. Conversely, the SI that the supervisor attributes to the user is much lower because of the third attribution rule. Consequently, in a large power distance culture the supervisor will choose to perform the following conferral, $< A =$ ask-to-wait$, V = 10 >$ which confers less to the user than the alternative, $< A =$ give-permission$, V = 18 >$, causing the observable difference shown in Figure 2.

In this example, as well as in the other situations the user ecounters troughout the game, the behavioural differences obtained by using the SID model with different cultural configurations have been validated by an expert on intercultural training. Although real cultures are infinitely more complex and subtle than the ones defined with the proposed approach, our goal is to make certain key aspects of real cultures highly salient so that it is easier for people to notice and understand them.

Table 4. Museum Scene - SI Attribution Rules

A	V	D	M
isPerson(x) = True	20	-	-
isElder(x) = True	1	PDI	5
isElder(x) = False	-1	PDI	-5

6 Conclusion

In this paper we have argued about the importance of considering cultural aspects of behaviour when developing socially-intelligent agents. Particularly, we focused on the problem of being able to express cultural differences in the way agents relationally perceive and interact with others.

In order to address this problem, we described a culturally-adaptable model of relational behaviour that is based on a particular view of status proposed in [15]. The proposed model endows BDI agents with a set of specific social interaction dynamics. These dynamics impact how agents perceive others, how much they are willing to act for others, and how much they feel entitled to have others acting in their favour. In humans, these dynamics are greatly affected by cultural conventions. Our model enables the encoding of such conventions as a set of parametrisable beliefs.

The model has been applied to develop an application for inter-cultural training in which the goal is for the user to learn cultural differences on a generic level by interacting with synthetic cultures. The model facilitates the creation of agents capable of simulating these cultures, through an explicit and flexible parametrisation of cultural influences. An example of the model being applied to generate different cultural behaviour was provided. The example is taken from one of the situations that users encounter in the training application. The example reflects two different extremes of the power distance dimension from Hofstede's theory [10].

As future work, we want to conduct a user evaluation to determine how users from different cultures perceive and react to distinct cultural configurations of the agents. More precisely, we plan to have participants from two countries that differ on a particular cultural dimension interacting with agents that are configured with both extremes of that dimension to see the impact of the model. Moreover, we plan to extend the current model to also address the link proposed by Kemper between emotional appraisal and his status-power theory [15]. Finally, it would also be interesting to integrate the other behavioural dimension of Kemper's theory, which is power, in order to be able to model situations in which agents unwillingly comply with others.

Acknowledgments. This work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013. It was also partially supported by a FCT scholarship (SFRH BD /62174/2009), by the European Community (EC), and was partially funded by the ECUTE project (ICT-5-4.2 257666). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC or the FCT, which are not responsible for any use that might be made of data appearing therein.

References

1. Aylett, R., Paiva, A., Vannini, N., Enz, S., Andre, E., Hall, L.: But that was in another country: agents and intercultural empathy. In: Proceedings of AAMAS 2009, Budapest, Hungary. IFAMAAS/ACM DL (May 2009)
2. Dias, J.: Lie to me: Virtual agents that lie. In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS), St. Paul, USA. IFAAMAS/ACM DL (May 2013)
3. Dias, J., Mascarenhas, S., Paiva, A.: Fatima modular: Towards an agent architecture with a generic appraisal framework. In: Workshop on Standards in Emotion Modeling, Leiden (2011)
4. Dias, J., Paiva, A.: Feeling and reasoning: A computational model for emotional characters. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 127–140. Springer, Heidelberg (2005)
5. Endrass, B., André, E., Rehm, M., Lipi, A., Nakano, Y.: Culture-related differences in aspects of behavior for virtual characters across Germany and Japan. In: Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), pp. 441–448 (2011)
6. Endrass, B., Nakano, Y., Lipi, A.A., Rehm, M., André, E.: Culture-related topic selection in small talk conversations across Germany and Japan. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 1–13. Springer, Heidelberg (2011)
7. Hill, R., Belanich, J., Lane, H.C., Core, M.: Pedagogically structured game-based training: Development of the elect bilat simulation. In: Proceedings of the 25th Army Science Conference (2006)
8. Hofstede, G.: Culture Consequences: Comparing Values, Behaviors, Intitutions, and Organizations Across Nations. Sage Publications, Thousand Oaks (2001)
9. Hofstede, G.: Role playing with synthetic cultures: the evasive rules of the game. In: Proceedings of the 9th International Workshop of the IFIP, Helsinky University of Technology, SimLab Report no 10, pp. 49–56 (2005)
10. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organizations: Software of the Mind, 3rd edn. McGraw-Hill, New York (2010)
11. Hofstede, G.J., Mascarenhas, S.F., Paiva, A.: Modelling rituals for Homo Biologicus. In: Proceedings of the Seventh Conference of the European Social Simulation (2011)
12. Hofstede, G.J., Pedersen, P.B., Hofstede, G.: Exploring Culture - Exercises, Stories and Synthetic Cultures. Intercultural Press (2002)
13. Jan, D., Herrera, D., Martinovski, B., Novick, D., Traum, D.: A computational model of culture-specific conversational behavior. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 45–56. Springer, Heidelberg (2007)
14. Johnson, W.L., Vilhjálmsson, H.H., Marsella, S.: Serious games for language learning: How much game, how much ai? In: Looi, C.-K., McCalla, G.I., Bredeweg, B., Breuker, J. (eds.) AIED. Frontiers in Artificial Intelligence and Applications, vol. 125, pp. 306–313. IOS Press (2005)

15. Kemper, T.: Status, power and ritual interaction: a relational reading of Durkheim, Goffman, and Collins. Ashgate Publishing Limited, England (2011)
16. Kroeber, A., Kluckhohn, C.: Culture: A Critical Review of Concepts and Definitions. Peabody Museum, Cambridge (1952)
17. Lee, E., Nass, C.: Does the ethnicity of a computer agent matter? an experimental comparison of human-computer interaction and computer-mediated communication. In: Emobied Conversational Agents (1998)
18. Mascarenhas, S., Dias, J., Afonso, N., Enz, S., Paiva, A.: Using rituals to express cultural differences in synthetic characters. In: Proceedings of AAMAS 2009, Budapest, Hungary. IFAMAAS/ACM DL (May 2009)
19. Mascarenhas, S., Dias, J., Prada, R., Paiva, A.: A dimensional model for cultural behaviour in virtual agents. International Journal of Applied Artificial Intelligence: Special Issue on Virtual Agents (2010)
20. Mascarenhas, S., Silva, A., Paiva, A., Aylett, R., Kistler, F., Andr, E., Degens, N., Hofstede, G.J.: Traveller: An intercultural training system with intelligent agents (demonstration). In: Proceedings of the International Conference on Autonomous Agents and Multiagent Systems. IFAAMAS/ACM DL (2013)
21. Si, M., Marsella, S., Pynadath, D.V.: Thespian: Modeling socially normative behavior in a decision-theoretic framework. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 369–382. Springer, Heidelberg (2006)
22. Solomon, S., van Lent, M., Core, M., Carpenter, P., Rosenberg, M.: A language for modeling cultural norms, biases and stereotypes for human behavior models. In: BRIMS (2008)

Looking Real and Making Mistakes

Yuqiong Wang[1], Peter Khooshabeh[2,⋆], and Jonathan Gratch[2,⋆⋆]

[1] Rochester Institute of Technology, Rochester NY, 14623, USA
wang@mail.rit.edu
[2] University of Southern California, Institute for Creative Technologies
Playa Vista CA, 90094, USA
{khooshabeh,gratch}@ict.usc.edu

Abstract. What happens when a Virtual Human makes a mistake? In this study we investigate the impact of a VH's conversational mistakes in the context of persuasion. Users interacted with a VH that told persuasive information, and they were given the option to use the information to complete a problem-solving task. The VH occasionally made mistakes such as not responding, repeating the same answer, or giving an irrelevant reply. Results indicated that a VH who makes conversational mistakes is capable of social influence. Individual differences also shed light on the cognitive processes of users who interacted with error-prone VHs. We discuss the implications of these results with regard to VH design.

Keywords: Virtual Humans, Human-Virtual Human Interaction, Virtual Reality, Photorealism, Conversational Mistakes, Need For Cognition.

1 Introduction

To err is human, to forgive, divine.
– Alexander Pope

Virtual Humans (VHs) are representations of humans in virtual environments, in the form of online puppetry of actual humans (i.e., avatars), or computer algorithms simulating people (i.e., agents). The technology has made possible Human-Virtual Human (H-VH) interaction in various applications such as medicine [1], education [2], and therapy [3]. Limitations of this technology, on the other hand, have left VHs prone to errors. For instance, a VH empowered by natural language processing may fail to answer a user's question, or give a nonsensical response [4]. Based on theories that suggest people treat computers as social actors [5], one would expect a social response to a VH is similar to that

⋆ Peter Khooshabeh was supported by a fellowship from the Cognitive Sciences Branch of the Army Research Lab Human Research and Engineering Directorate.
⋆⋆ This research was sponsored by the National Science Foundation (grant # IIS-1211064) and the US Army RDECOM. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 339–348, 2013.

elicited by a real human. In the real world, if an individual makes a mistake, others are apt to criticize him or her. How do humans respond to a VH's mistake? In this study we observe how interactants respond to VHs' mistakes in a problem-solving scenario.

In the broader view, our study seeks to inform theories of VH design. The Threshold Model of Social Influence (TMSI) [6,7] is one theoretical model that aims to explain how aspects of a VH's design impact its effectiveness. According to this model, the extent to which people are socially impacted by a VH depends on three variables. *Agency* is the extent to which the participant believes he or she is interacting with another sentient human being. *Communicative realism* (formerly *behavioral realism* [6]) is the degree to which human representations and objects behave as they would in the physical world, which consists of *movement realism*, *anthropometric realism*, and *photographic realism*. Finally, *response system level*, which varies from automatic to deliberative, characterizes the extent to which the interaction with a VH involves shallow automatic cognitions, or deeper deliberate ones. With these three factors, the model argues that the influence of a VH is heavily impacted by its communicative realism, unless interactants believe it is controlled by another sentient human, or if interactants are not thinking deeply about the interaction. Our study, manipulating communicative realism and observing individual differences, serves as a test on the validity of the TMSI.

Communicative realism includes not only the extent to which it looks human (i.e., anthropometric realism and photographic realism) but also how it acts (i.e., movement realism) and speaks [8]. From this perspective, the errors produced by current state-of-the-art natural language processing serve to reduce the realism of the VH's behavior. Previous works studying VHs' social influence have typically focused on errorless VHs [9,10]. In this study we independently manipulate VHs' conversational errors. Based on the TMSI, we predict that erroneous VHs' social influence, if any, will be less than that of errorless ones.

Regarding the *photographic realism* component, the TMSI suggests that visual fidelity should have modest contribution when compared with behavior [6]. Indeed, a later study [9] demonstrated that a non-photorealistic VH can be as persuasive as a photorealistic one. However, there is disagreement on this point: some other studies have found that *photographic realism* enhances V-VH interaction [11]. To look into the debate, the present experiment manipulates *photographic realism* independently.

While communicative realism refers to aspects of the VH, the third factor of the TMSI - response system level - refers to aspects of the human interactant. Response system level could be impacted by the nature of the task in which the human and VH are engaged (e.g., a thought-provoking puzzle vs. an improvised conversation). The factor can also be influenced by the mind-set of the human interactant. For example, some people are inherently deep-thinkers, whereas some others tend to act intuitively and automatically. Such distinction is addressed by social psychologists as the Need For Cognition (NFC) [12]. More recently and consistently with the TMSI, NFC has been proposed as one of the factors explaining when people will anthropomorphize non-human entities [13]. Specifically, people low in NFC are more likely to engage with VHs, leading to the speculation that such people are more influenced by VHs.

The literature on NFC suggests to further unpack the effect of NFC on social influence. It has been reported that individuals low in NFC (i.e., intuitive thinkers) tend to focus on peripheral cues, instead of content of the information [14]. In a H-VH persuasion dialog, presumably the content of the dialog is central, whereas aspects of communicative realism (such as photorealism and behavioral realism) are peripheral, at least to the extent that they do not modify the content of the persuasive dialog. Therefore we can expect an interaction between communicative realism and NFC: when presented with departures from communicative realism, individuals low in NFC are likely to overlook the underlying persuasive information, and become less persuaded.

Other than factors presented by the TMSI, other variables may play an important role as well. For instance, a VH's and interactant's gender have been found to interact during social scenarios. Zanbaka and colleagues reported that people are more affected by VHs of the *opposite* gender [9], whereas Guadagno found an opposite pattern of results [10]. This study manipulates VH gender to revisit the discrepancy.

To summarize, the purpose of this study is to determine:

1. Whether an error-prone VH is capable of social influence during a persuasion task.
2. How do potentially interfering variables, including VH's photographic realism, VH's gender, interactant's NFC, and interactant's gender, interact and moderate VH's social influence.

2 Method

2.1 Participants

Three-hundred and twenty-six workers were recruited from Amazon Mechanical Turk and participated in our study for monetary reward. All workers resided in the United States. There were 48% females and 52% males, with an average age of 31.5 (SD=10.8). Caucasians comprised 71% of the sample, whereas the remainder consisted of Asians (13%), Hispanic Americans (8%), and African Americans (6%). One participant was excluded from analysis because the measured social influence was more than three SD's away from the mean.

2.2 Material

We assessed VHs' social influence by adapting a standard laboratory task that was designed to measure persuasion. In the Lunar Survival Scenario [15,16], the participant is asked to imagine he or she is stranded on the moon and need to prioritize a set of items necessary for survival. The participant can discuss the priorities with a teammate, and possibly make changes as a result of the conversation. Persuasiveness is measured by the extent to which the participant shifts his or her priorities in the direction of the teammate's advice.

The VHs in our study acted as the teammate and provided advice consistent with NASA's expert opinions. The conversation was text-based: the participant asked a question by selecting from a pre-determined set of questions, and the VH responded via printing text on the screen. A screenshot of the interface is shown in Fig. 1. To make the VH appear less mechanic, a delay of 2.5 seconds was introduced before the VH answered each question.

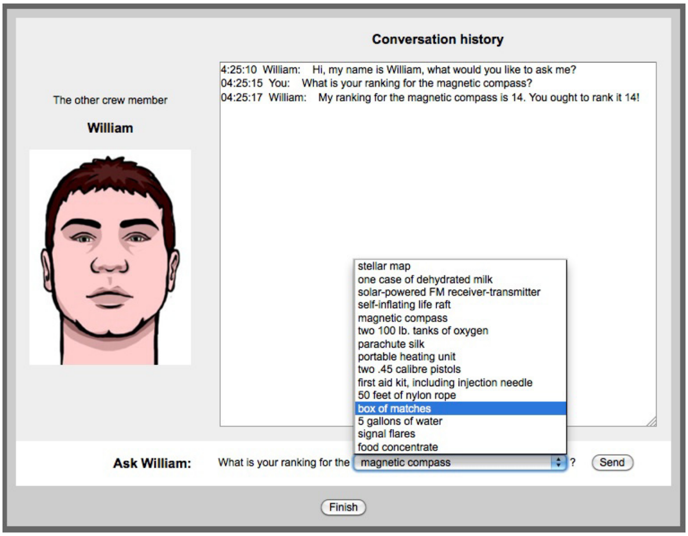


Fig. 1. The H-VH interaction interface

About half the participants were assigned an error-prone VH as the teammate. An error-prone VH makes a conversational mistake 33.33% of the time when he or she answered a question. The type of mistakes is randomized among three patterns [4]:

No response. The VH does not reply to the participant's question. An empty line is printed in place of un expected answer.

Wrong answer. The VH gives information about an irrelevant item, e.g.:
H: What is your ranking for the ***box of matches****?*
VH: I ranked ***oxygen*** *1. You should rank it 1 if you haven't done so!*

Repeat. The VH answers the same question twice, for instance:
H: What is your ranking for the stellar map?
VH: My ranking for the stellar map is 3. Rank it 3!
VH: My ranking for the stellar map is 3. Rank it 3!

Despite the introduction of conversational errors, the information provided by the VH is always accurate, i.e. the VH always gives correct rankings. In situations where the VH is talking about an irrelevant item (such as *oxygen* in the example above), the ranking of that item (i.e., oxygen) is accurately given.

2.3 Design

This study used a between-subject design with three independent variables: Error {Errorless, Erroneous} X Photographic realism {Photorealistic, Non-photorealistic} X VH's Gender {Female, Male}. Social influence is measured via three dependent variables:

Correctness of Final Rankings. How well a participant's *final rankings* match the VH's rankings (i.e., NASA's expert rankings), with 100 being a perfect match, and lower scores indicating further deviation from the VH's rankings, in terms of Euclidean Distance:
$Correctness = 100 - \sqrt{\sum_{i=1}^{15}(H'sRanking_{item_i} - VH'sRanking_{item_i})^2}$

Improvement Score. How much a participant's rankings have shifted toward the VH's rankings, calculated as the subtraction of the Correctness of initial rankings by the Correctness of final rankings:
$ImprovementScore = Correctness_{InitialRankings} - Correctness_{FinalRankings}$

Change Score. How much a participant's rankings have changed, defined as the difference between the initial and final rankings, in terms of Euclidean Distance. (The difference between *Improvement Score* and *Change Score* is that *Improvement Score* is a signed measure, with a positive value indicating the amount of changes made in the direction of the VH's suggestions, whereas *Change Score* is an unsigned variable reflects the amount of total changes, both toward and against the VH's advice.)

$$ChangeScore = \sqrt{\sum_{i=1}^{15}(InitialRanking_{item_i} - FinalRanking_{item_i})^2}$$

2.4 VH Faces

The purpose of this study is to carefully manipulate communicative realism while holding other factors - the ones that could impact persuasiveness - constant. Previous work has emphasized that people judge by facial appearance and act differently [17,18]. To control for appearance, we used an independent set of workers to rate the perceived personality of a number of male and female VH faces. We selected pairs of faces that were perceived to have the same personality. The procedure is described below.

Three male and three female VHs were chosen from an institutional database. For each VH, one photorealistic and one non-photorealistic facial image were created, in a way that outlines of the two faces overlapped. One hundred workers, recruited through Amazon Mechanical Turk, rated each face (in randomized order) on twenty-seven traits. Twenty-four traits are personality traits from [19], whereas the remaining three are: *Honest*, *Familiar*, and *Photorealistic*. Factor analysis on the twenty-seven traits yielded three main factors:

Trustworthy/Honest, which accounts for 33.00% of the variance; *Clever/Wise*, which explains 15.84% variance; and *Introvert/Unsociable*, contributing to 8.25% of the variance. All three factors have a Chronbach's alpha greater than 0.8.

Ratings in terms of the three main factors revealed that, one male VH's photorealistic and non-photorealistic faces appeared to convey the same personality. One female VH's photorealistic and non-photorealistic faces were considered similar: the first two factors achieved consistent ratings, and the third factor resulted in slightly different ratings ($M = 0.07$ vs. $M = -0.10$; $t(195) = 2.1, p = 0.034$; Cohen's $d = 0.3$, effect size= 0.15). These two VHs (shown in Fig. 2) were selected for later study.

A manipulation check on photorealism was performed. The female VH's photorealistic and non-photorealistic representations were rated 3.0 (SD=1.1) and 2.1 (SD=1.1) respectively, on a scale of 1 to 5, with 5 meaning *Very photorealistic*. The ratings are significantly different with $t(198) = 5.12, p < 0.001$. The male VH's photorealistic version had a mean rating of 3.2 (SD=1.0), which differs dramatically from his non-photorealistic representation ($M = 2.2, SD = 1.2; t(198) = 6.40, p < 0.001$).

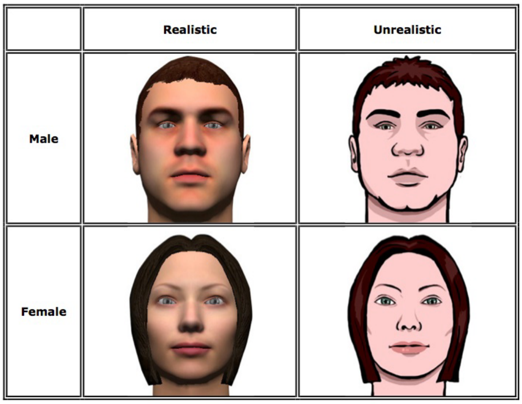


Fig. 2. VHs used in the study

2.5 Procedure

The participants were directed to our online questionnaire, which started with basic demographic questions including age, gender, ethnicity, and the Need for Cognition Scale (NFCS).

Next, participants were informed that they had crash landed on the moon and needed to choose items in order to trek 200km back to a life-saving rendezvous point. Additionally, they were told that another crew member was also present. Due to the crash landing, the captain had been incapacitated and the participant was now the officer in charge. Participants were told that the the other crew member, who knew the ship's inventory well, would be able to answer questions about the item rankings.

Although the other crew member knew the inventory well, the instructions stressed that the participant was the officer in charge. As the captain, the participant had to make the final decision on how to rank the items. We intended the participants to have the decision power and responsibility in the scenario so that they would not blindly take or reject the virtual agent's advice.

Participants initially ranked items prior to chatting with the virtual agent by dragging and dropping the task items on a graphical user interface. After the participants completed their pre-chat rankings, they were reminded that they were in charge and now had an opportunity to chat with the other crew member, who was a VH. Participants had to decide how they would use the information that the VH told them in order to make a set of post-chat rankings.

3 Results

Social influence. Participants' rankings changed significantly after receiving persuasive information from the VHs (within-subjects effect $F(1, 323) = 37.6, p < 0.001$). Participants' rankings shifted toward VH's rankings in both the errorless condition ($F(1, 184) = 113.0, p < 0.001$), and erroneous condition ($F(1, 139) = 37.6, p < 0.001$). These results imply that VHs, even erroneous ones, are capable of social influence.

Error. Error has an effect on *Improvement Score* ($F(1, 323) = 4.4, p = 0.037$) and *Correctness of Final Rankings* ($F(1, 323) = 10.9, p = 0.001$). As predicted, erroneous VHs are less influential (Improvement Score: $M = 2.7, SD = 5.2$ vs. $M = 3.9, SD = 5.0$; Correctness of Final Rankings: $M = 84.9, SD = 5.9$ vs. $M = 87.0, SD = 5.5$).

NFC. We performed a median split on participants' NFCS scores ($Median = 21.0$). The means of the two groups are significantly different ($M = -3.5, SD = 19.0$ vs. $M = 40.0, SD = 12.2$; $t(323) = 24.3, p < 0.001$).

We found an effect of NFC on *Change Score* ($F(1, 323) = 8.1, p = 0.005$), but not on *Improvement Score* or *Correctness of Final Rankings*. Participants low in NFC made more changes ($M = 13.0, SD = 6.7$) than those high in NFC ($M = 11.0, SD = 5.6$).

A three-way interaction between $ErrorXPhotorealismXNFC$ is found (Improvement Score: $F(1, 317) = 5.9, p = 0.010$; Correctness of Final Rankings: $F(1, 317) = 5.9, p = 0.016$). In the context of low-NFC individuals interacting with non-photorealistic VHs, conversational errors have an effect on persuasiveness (Improvement Score: $F(1, 317) = 12.0, p < 0.001$; Correctness of Final Rankings: $F(1, 317) = 12.1, p = 0.001$). Individuals who experienced errors were less persuaded. This observation is consistent with the error effect found on the whole sample. However, this is the only cell that shows an error effect, implying that it is driving the overall error effect.

A second cell in the three-way interaction reveals a NFC effect: under the condition that the VHs are non-photorealistic and errorless, NFC has an effect on persuasiveness (Improvement Score: $F(1, 317) = 12.9, p < 0.001$). Low-NFC individuals are more persuaded than high-NFC individuals under this condition.

Photographic realism. Photographic realism alone does *not* have any effect on VHs' social influence.

Gender. The expected two-way interaction between *VH's Gender* and *Participant's Gender* was *not* found. Instead, we observed an effect of *Participant's Gender* on *Change Score* ($F(1, 323) = 7.6, p = 0.006$), but not on *Improvement Score* or *Correctness of Final Rankings*. In short, female participants made more changes ($M = 13.0, SD = 6.0$) than male participants ($M = 11.1, SD = 6.4$).

4 Discussion

In a problem solving scenario, we measured VHs' social influence by observing interactants' attitude change after receiving persuasive information. To study the effect of design characteristics of VHs, we manipulated communicative realism (via introducing conversational errors), photographic realism, and VH's gender. Our study suggested:

1. VHs who make conversational mistakes are capable of social influence.
2. Individual differences in participants' response system level (indexed by NFC), together with communicative realism (both photographic and linguistic realism), play a role on VH's social influence.

The TMSI suggests that individuals who use deliberative cognitive processing when responding in socially motivated scenarios have relatively higher thresholds of social influence. The observations in this study in the errorless, non-photorealistic condition confirmed this prediction: participants high in NFC (i.e. deep-thinkers) were *less* influenced, whereas users low in NFC were *more* influenced.

For individuals low in NFC (i.e., intuitive thinkers), their decision differed significantly when facing errorless vs. erroneous VHs. As noted above, studies [20,14] have reported that individuals low in NFC tend to focus on peripheral cues. The conversional errors in our study are peripheral to the content of the conversation, suggesting that individuals low in NFC were distracted by errors and overlooked the underlying persuasive information; consequently, they were not persuaded. However, as demonstrated by our data, error alone was not sufficient to produce this effect. Rather, conversational errors combined with non-photorealistic appearance drove the observed effect. These results provide support for the TMSI with regard to low-NFC participants whose response-system level is more automatic.

In comparison, the high NFC group were impervious, both to the manipulation of conversational errors and VH's photorealism. This is possibly due to the fact that individuals high in NFC tend to focus on the content of the information rather than peripheral cues [14]; they may, hypothetically, have filtered out conversational errors and photographic differences from underlying persuasive arguments. As a result, they processed the arguments as if peripheral changes were nonexistent. These observations on high NFC participants align with the TMSI. Together with the findings for low-NFC participants, the results stress the importance of the response system level factor in the TMSI.

Our results may shed some insight into the mixed findings concerning photo-realistic vs. non-photorealistic VHs. Observations from this study is consistent with the interpretation that visual fidelity may be insufficient in and of itself to alter the effectiveness of VHs. Rather, it may interact with other factors in determining social influence. In our experiment, conversational errors combined with non-photorealistic appearance demonstrated an effect on individuals low in NFC. One possible explanation is that the error plus non-photorealism combination produced a significant effect that would have been less apparent if the factors were to be introduced in isolation. Alternatively, the impact of photo-realism might have overridden the effect of errors for individuals low in NFC. Either way, our findings emphasize the importance of separately controlling for different aspects of communicative realism in future studies.

Another characteristic of the interactants we investigated was gender. The expected interaction between VH's gender and participant's gender was absent in our study. More research is in demand to justify whether people are more affected by VHs of the *opposite* gender [9], or the *same* gender [10].

There are several future directions that we are interested in pursuing. In the present study, we controlled for communicative realism and response system level, but not *agency*. It is possible that conversational errors not only impacted realism, but also participants' willingness to attribute human sentience to the agent. To determine the impact of errors on agency, we need to independently manipulate *agency*.

The concept of NFC also calls for further exploration. In this work we analyzed between-subject differences in NFC in a repeated task. Additionally, we can manipulate within-subject variations in NFC by having the same subject performing different tasks that vary in cognitive needs.

While the present study examined factors from one theoretical model regarding VH influence (i.e., TMSI), other models exist in the literature. For example, the Ethopoeia concept of Nass and colleagues [21] has been raised as an alternative model for guiding VH design. Although our results lend support to the TMSI, we did not directly contrast TMSI with other models. More research is in demand to guide the theoretical foundations of VH design.

References

1. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer–patient working alliance in automated health behavior change interventions. Patient Education and Counseling 59(1), 21–30 (2005)
2. Woolf, B.P.: Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning. Morgan Kaufmann (2010)
3. Rizzo, A., Buckwalter, J.G., Forbell, E., Reist, C., Difede, J., Rothbaum, B.O., Lange, B., Koenig, S., Talbot, T.: Virtual reality applications to address the wounds of war. Psychiatric Annals 43(3), 123–138 (2013)
4. Skarbez, R., Kotranza, A., Brooks, F., Lok, B., Whitton, M.C.: An initial exploration of conversational errors as a novel method for evaluating virtual human experiences. In: 2011 IEEE Virtual Reality Conference (VR), pp. 243–244. IEEE (2011)

5. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 72–78. ACM (1994)
6. Blascovich, J.: Social influence within immersive virtual environments. In: The Social Life of Avatars, pp. 127–145. Springer (2002)
7. Blascovich, J., McCall, C.: Social influence in virtual. In: The Oxford Handbook of Media Psychology, p. 305 (2013)
8. Khooshabeh, P., Dehghani, M., Nazarian, A., Gratch, J.: The cultural influence model: When accented natural language spoken by virtual characters matters. Journal of Artificial Intelligence and Society
9. Zanbaka, C., Goolkasian, P., Hodges, L.: Can a virtual cat persuade you?: the role of gender and realism in speaker persuasiveness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1153–1162. ACM (2006)
10. Guadagno, R.E., Blascovich, J., Bailenson, J.N., McCall, C.: Virtual humans and persuasion: The effects of agency and behavioral realism. Media Psychology 10(1), 1–22 (2007)
11. Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J.K., Robles, E.: Evaluating the effects of behavioral realism in embodied agents. International Journal of Human-Computer Studies 67(10), 842–849 (2009)
12. Cacioppo, J.T., Petty, R.E.: The need for cognition. Journal of Personality and Social Psychology 42(1), 116–131 (1982)
13. Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: a three-factor theory of anthropomorphism. Psychological Review 114(4), 864 (2007)
14. Haugtvedt, C., Petty, R.E., Cacioppo, J.T., Steidley, T.: Personality and ad effectiveness: Exploring the utility of need for cognition. Advances in Consumer Research 15(1), 209–212 (1988)
15. Johnsen, K., Raij, A., Stevens, A., Lind, D.S., Lok, B.: The validity of a virtual human experience for interpersonal skills education. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1049–1058. ACM (2007)
16. Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J., Blascovich, J.: Does it matter if a computer jokes. In: PART 1-Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, pp. 77–86. ACM (2011)
17. Wilson, R.K., Eckel, C.C.: Judging a book by its cover: Beauty and expectations in the trust game. Political Research Quarterly 59(2), 189–202 (2006)
18. Stirrat, M., Perrett, D.I.: Valid facial cues to cooperation and trust male facial width and trustworthiness. Psychological Science 21(3), 349–354 (2010)
19. Hassin, R., Trope, Y., et al.: Facing faces: Studies on the cognitive aspects of physiognomy. Journal of Personality and Social Psychology 78(5), 837–852 (2000)
20. Petty, R.E., Cacioppo, J.T.: The elaboration likelihood model of persuasion. In: Communication and Persuasion, pp. 1–24. Springer (1986)
21. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of Social Issues 56(1), 81–103 (2000)

What's Going on? Multi-sense Attention for Virtual Agents

Tim Balint and Jan M. Allbeck

Laboratory for Games and Intelligent Agents
George Mason University
4400 University Drive, MSN 4A5
Fairfax, VA 22030
{jbalint2,jallbeck}@gmu.edu

Abstract. When designing virtual humans, it is imperative that the virtual agent behave in a human-like fashion. In certain circumstances, this requires agents to be bounded in their ability to sense and understand the environment. To this end, we create a methodology that provides perceptual attention based on a linear combination of senses. We use different heuristics on the object and events that can be sensed by a virtual agent, and combine these scores to create an overall score for a given object or event. This allows the virtual agent to perceive interesting or unexpected items. To demonstrate our technique, we give an example showing the ability of using a linear combination of senses.

1 Introduction

Virtual humans play a growing role in movies, games, and simulations. The primary role of these agents is to behave like physical humans, requiring the agent to make human-like decisions. As physical humans interact within the bounds and understanding of their environment, which can change unexpectedly, virtual humans should be able to as well. To simulate this, an agent must be able to know and understand what is going on in their environment, as can be seen from Figure 1.

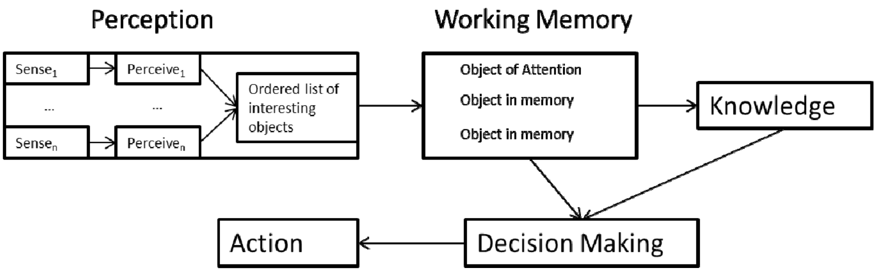


Fig. 1. A sample agent framework, with the perception and working memory portions expanded. In this figure, the perception module is broken down to display a multitude of sensing abilities.

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 349–357, 2013.

An agent interacting within a semantic environment, also known as a smart object environment [1], must be able to use and understand the semantics found within that environment. Complex computer vision systems have been designed to imbue an agent with synthetic vision, allowing the agent to sense their environment [2]. However, data in a virtual semantic environment is attached to the objects; this can make understanding the virtual world as simple as looking up values in a table. If the agent's model of perception consists solely of looking up values, then the agent exists inside a fully known environment. This may not be plausible for certain cognitive processes (such as agent memory models and decision-making processes), but can be addressed by modeling various aspects of agent perception. Also, when simulating virtual humans, it is natural to believe that the five human senses will suffice. However, humans augment their senses every day, with tools as simple as thermometers or as complex as radar and night vision systems. Instead of designing all possible senses, another alternative is to generalize what a percept is, and from there, add specifics based on the requirements of a simulation. Determining the best solution for such a problem is generally context dependent, as the information present within the environment plays a necessary role in how the agent can process its environment. A simulation author facilitates this process by creating the environment and populating it with information. Therefore, the simulation author is able to choose how an agent understands its world.

An agent can sense the environment, but if the environment is sufficiently complex, there could be too much information for the agent to realistically process. Physical humans do not process all objects within an environment at once, and generally only concentrate on a few at a time [2]. For example, a person who glances at a cluttered coffee table may not notice their keys amidst the remotes and books. In order to create more believable virtual humans, an agent may need to prune away at its sensed search space, keeping only the interesting or stimulating information- thus adding the realistic effect of being unobservant or overlooking.

To allow virtual humans to understand their environment, we propose and implement a generalized framework for agent sensing and perception. This framework includes:

- A generalized agent sensing system that allows simulation authors to create specific senses for a given simulation.
- A perception system that uses a heuristics to create sense attention through a bottom-up and top-down process.
- A linear combination of forms of perception that allow for agents to combine perception scores across multiple senses.

2 Related Work

Perception for virtual agents is a diverse topic, as it is a vital part of the sense-think-act cycle found in both physical and virtual modern agents. A large sector of the computer vision community has focused on agent perception, and a survey can be found in [3]. Many of these systems create false color images or saliency maps, the latter being especially used in the computer vision community [4]. Specifically within the virtual agents community, work has been made in top-down perception [5], bottom-up perception [6], and a combination of the two [7]. All of these systems provide agents with a sense of vision while [6] and [7] also provide an auditory sense. They forgo much of the processing that is done in the vision community, as all the information available to them can be accomplished in a search of the object. We have adopted this format, and much like these systems, use information inherent and available to the agents. We chose to generalize the sensing ability to allow for the definition of multiple senses, instead of specific formats for each individual sense.

Many systems couple visual attention with perception, which allows an observer to determine what the agent actually perceives. The agent could give attention to events in their environment [7], or other agents [8]. Visual attention has also been used to facilitate memory systems [9] and general agent decision-making [5]. While some of these systems provide multiple senses for the agent to receive information though, they handle information from each distinct sense separately, that is, there is no attempt to combine senses. We have chosen to allow agents to combine different senses within their perception system when computing costs of visual attention, making our system determine sensory attention, rather than just visual attention.

Perception and visual attention have been added to many different systems for the purpose of creating more intelligent virtual agents. Embodied conversational agents in particular have used perception to allow them to understand the physical or virtual human they are conversing with [10]. This has been researched enough to warrant study into a markup language for perception in conversational agents [11]. There has also been an emphasis on creating perception for embodied agents interacting with a virtual environment, and many of the systems listed above do so including [5] [6] [7]. Reactive agents, such as those found in [12] also use perception and visual attention to understand concepts of a virtual environment. [13] and [14] take this one step further by using visual attention as a pre-cursor to other motor activities, such as reaching, searching, and catching. We focus on perception for agents within a virtual environment. Our agents use an understanding generated through a combination of multiple senses to create an attention span and populate the working memory component seen in Figure 1. This means that the agent focuses mostly on objects within the environment, and so our work is not necessarily suited for use in agent conversation.

While there has been a great deal of work on creating multiple sense attention for virtual agents, many of the solutions focus only on information from one sense at a time. These models process information as it is received, and most perform (visual) attention on a set of objects once it is seen in the environment, making them an event driven model of perception. However, when attempting to combine information from multiple senses and reason over these senses, processing objects one at a time is no longer manageable, especially when the reasoning for one object requires understanding the surrounding environment. In an attempt to combine multiple senses, we provide a method of grouping, reasoning, and combining information about objects, in order to create a sense attention.

3 Methodology

3.1 Object and Action Representations

The representation of objects in our environment are inspired by Smart Objects [1], which contain sets of properties representing semantic information. Semantic properties, hereby denoted p, are categorized into sets, S, and these sets make up the group of semantic information. Objects may contain one property per set, and may be marked with many different sets. For example, properties p_i and p_j are members of set S_m, and property p_k is a member of set S_n. An object may have either property p_i or p_j, without an effect on its ability to also have p_k. These semantic sets are generally contained globally, and individual objects may inherit a single property from these sets, although the inherited property may change during the simulation. Before the simulation, the sets of semantic properties can be authored by a user, or possibly determined from a common-sense database [15]. A subset of the property sets we are currently using is found in Table 1.

Table 1. A subset of the property sets understood by our perception system

Property	Type
Olfactory type	String set
Visual Hue/saturation/brightness	Integer set
Visual Luminance	Integer set
Auditory frequency/Intensity	Integer set

3.2 Sense Preprocessing

In order to endow virtual agents with sense attention, the agent must be able to determine if it can sense an object. We define sensing as the ability an agent uses to determine the presence of a semantically labeled object through some semantic information attached to that object. We represent a change in semantic information upon an object as an event. For example, an agent can sense a pizza through seeing its shape and color or through a pizza's distinct smell. Different senses should discern different properties of a semantic object, although some properties can manifest themselves in many ways. A semantic object that is hot could have a *heat_signature* semantic information type, detectable by an infrared sense or touch sense. An object that is hot could also glow red and therefore be detected by a visual sense. We consider these two semantic properties distinct, and each one is added to the virtual environment separately.

We provide our agents a method to determine which objects are in their general sensing area. Much like [6], our agents have an area they can sense in, which is determined by the simulation author, and generally varies by sense and by agent. This is determined by two variables, a subtended angle α and a sensing distance d. α and d form polar coordinates that allow agents to determine what objects are within their sensing area, as seen in Figure 2. Unlike techniques such as ray-tracing, our technique does not inform an agent if an object is blocked by another object. While this removes some realism from certain sensing abilities, we believe that it does not negatively impact the agent. Certain senses, such as auditory and olfactory senses, do not rely on objects being in the agent's line of sight, but being within a certain range of the agent, thus providing them with back up receptors to visibility.

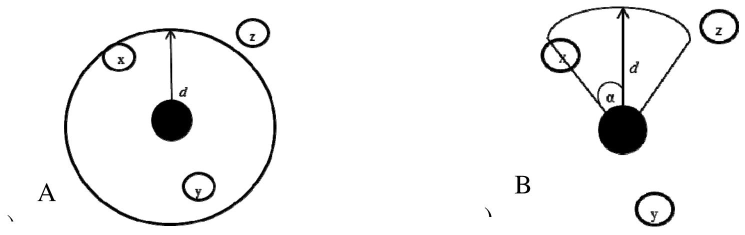


Fig. 2. Sense pre-processing areas A) α is 360 degrees around the agent. In this example, x and y are sensed by the agent, but z is not. B) α is less than 360 degrees. In this example, only object x is sensed.

To determine if an object is in the sensing area, we keep a list of all sorted objects that are then examined by the agent. Objects that exist within this sensing area are fed into the sense where they are pruned from the agent's consideration based upon the semantic properties they contain. At this stage of processing, it is only important for the agent to determine if an object has one semantic property useful to the sense. For example, certain properties, such as a frequency and decibel level, are generally associated with an auditory sense. If an agent is examining an object that has semantic information from one of these two properties, the object is still passed on for further processing. This allows for some generalization between sets and between objects.

3.3 Perception Attention

Physical humans can only process a limited number of items at a time, which is simulated by providing the agent with an upper bound on the working memory portion of the agent system (See Figure 1). Only unusual or important information is generally retained, thus rejecting much of what a physical human senses. However, what is generally considered interesting by one sense, when experienced by multiple senses, may not be considered as interesting. In order to create more plausible virtual humans, we believe that this ability should also be mimicked. To accomplish this, we model virtual human perception over a multitude of senses.

In order to create specificity when designing perception for multiple senses, we employ a series of heuristics, similar to [12], a subset of which is seen in Table 2. However, unlike [12], which determines common attributes from sensed objects, we attempt to rank objects based on pre-authored information readily available from either the agent or the environment (such as the distance between objects or an object's hue). We design heuristics based on two forms of comparison: agent-object interaction, which we label as top-down interactions, and object-object, or bottom up interactions.

Table 2. A sample of the heuristics used in our perception system. *Sense* is the sense type we designed the heuristic for. *Form* is whether the heuristic compares objects to other objects (Bottom-up) or objects to an agent (Top-down). The *type* of heuristic shows the basic way in which it is calculated.

Name	Sense	Form	Type
Auditory Saliency	Auditory	Bottom-up	Comparative
Olfactory Saliency	Olfactory	Bottom-up	Comparative
Velocity Saliency	Multiple	Bottom-up	Comparative
Interacting	Multiple	Top-down	Selective
Using	Multiple	Top-down	Selective
Useful Object	Multiple	Top-down	Selective

Top-down interaction heuristics allow an agent to create a personal score with the object, therefore yielding different results for different agents. For example, an *interacting* heuristic, determines if an object (such as another agent) is interacting with the agent. From Table 2, it can be seen that we regard all top-down heuristics as selective heuristics, which follow the basic form found in Equation 1.

$$h(object, agent) = \begin{cases} score & if\ object\ passes\ selection\ test \\ 0 & otherwise \end{cases} \tag{1}$$

Other heuristics, which employ object-object interactions, are considered bottom-up heuristics. As can be seen from Table 2, many of these heuristics use saliency and are comparative. Unlike most systems, we do not use computer vision techniques to create a saliency image, but instead perform comparisons based on the objects through the sense pre-processing phase. While in some cases this will remove information that is generally used in saliency maps, it maintains object-object interaction while removing certain forms of pre-processing (such as background subtraction). As can be seen from Table 2, many bottom-up heuristics are comparative, and so the object must be compared to all other objects within the environment. These heuristics, being comparative in nature, also create relative perception within the heuristic. A less intense smell, when compared to a stronger one, will be scored much lower using these types of heuristics.

Many heuristics used by the agent require statistical processing. Heuristics such as saliency require not only comparisons to each object in the area, but also to the average object score. We implement comparisons to the average, maximum, and minimum score for a set of objects for a given heuristic. Since heuristics can be re-used over multiple senses, it is not efficient to embed these statistics within the heuristics, but better to process them after the heuristic has been run on all objects.

After a heuristic, H, is processed for all objects, it is normalized with a weight, ws, over the total of all weights, wn, and added to a hash table object->score. After all heuristics are processed for a given sense, that sense is normalized with a weight w and total of all weights n as well. The total score for a given object over all senses is then given as a summation, seen in equation 2. By using linear weights on both heuristic scores and senses, a simulation author can control whether one heuristic or sense should dominate the others. Finally, the hash table is pruned to only the objects that score highest from senses, creating a bounded memory in complex scenes, such as scenes found in [16].

$$u = \sum_{i=1}^{t} \frac{w_i}{n} \sum_{j=1}^{s} \frac{ws_j}{wn} H_j(o) \tag{2}$$

4 Analysis and Results

4.1 An Example

In order to highlight the difference between using an agent-based linear combination perception system and one using simple selection, a sample scenario has been created. We have modeled a complex diner environment, which contains several objects. Situated within this diner is a kitchen, complete with a stove, microwave, pots, pans, cups, food, and other objects typically found in a kitchen. Each of these objects has been labeled with different semantic properties, a subset of which is found in Table 1. Certain properties from these sets (such as sound properties) are also added to objects through events. Actuators for the virtual agents are provided through Smartbody [17]. Additional examples can be found with the accompanying video.

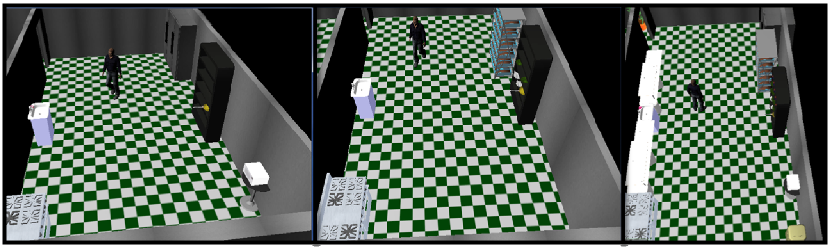


Fig. 3. The agent observing its environment. Objects seen in the environment are the objects perceived by the agent. Left: The agent observes its environment using an attention selection method. Middle: The agent observes its environment using our linear combination method. Right: The full kitchen environment.

Our scenario has an agent enter the kitchen, examining its surroundings and reporting on the objects it perceives, using visual, auditory, and olfactory senses. All items have at least visual semantic information, and many items, such as the refrigerator or oven, have a sound or smell associated with them. Figure 3 displays the agent's observation using a selection method on the left and a linear combination method in the middle. As can be seen from the images, several of the objects are the same. However, some items, such as the refrigerator, are ignored by the linear combination method. As the refrigerator is only recognized by the auditory sense with both methods, the uninteresting sound that it makes goes unnoticed when more interesting objects, such as the glasses and cups, are in range. The refrigerator's sound is noticed with the selection method due to the lack of objects that have auditory semantics, and that the auditory sense is the last sense to be checked.

5 Conclusions and Future Work

We have provided a method to create perceptual attention based on a combination of different senses. This is accomplished using a preprocessing step to determine objects and events capable of being considered by a given sense, and a ranking step to determine objects that are useful or interesting to the agent. This provides the agent with an ability to differentiate objects in its environment through the use of all available information.

Future research will examine the heuristics used in this work. Certain heuristics may be applicable only for certain situations, and so the ability for the agent to adapt and control its heuristics may provide faster and more interesting results. Optimizations on these heuristics, especially the comparative ones, would also prove useful.

References

1. Kallmann, M., Thalmann, D.: Modeling Objects for Interaction Tasks. In: Eurographics Workshop on Animation and Simulation, Lisbon, pp. 76–86 (1998)
2. Hill, R.W.: Perception Attention in Virtual Humans: Towards Realistic and Believable Gaze Behaviors. In: AAAI Fall Symp. Simulating Human Agents, pp. 46–52 (2000)
3. Peters, C., Castellano, G., Rehm, M., et al.: Fundimentals of Agent Perception and Attention Modeling. In: Emotion-Oriented Systems. Springer, Heidelberg (2011)
4. Canosa, R.: Modeling Selective Perception of Complex, Natural Scenes. International Journal on Artificial Intelligence 14, 233–260 (2005)
5. Joost van, O., Frank, D.: Scalable Perception for BDI-Agents Embodied in Virtual Environments. In: International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 46–53 (2011)
6. Herrero, P., Greenhalgh, C., De Antonio, A.: Modeling the Sensory Abilities of Intelligent Virtual Agents. In: Autonomous Agents and Multi-Agent Systems, pp. 361–385 (2005)
7. Kim, Y.-J., van Velsen, M., Hill Jr., R.W.: Modeling Dynamic Perception Attention in Complex Virtual Environments. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 266–277. Springer, Heidelberg (2005)
8. Rymill, S.J., Dodgson, N.A.: Psychologically-Based Vision and Attention for the Simulation of Human Behavior. In: Proceedings of the 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia (GRAPHITE 2005), pp. 229–236 (2005)
9. Cha, M., Cho, K., Um, K.: Design of Memory Architecture for Autonomous Virtual Characters using Visual Attention and Quad-Graph. In: ICIS, pp. 691–696 (2009)
10. Peters, C.: Direction of Attention Perception for Conversation Initiation in Virtual Environments. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 215–228. Springer, Heidelberg (2005)
11. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A(S.), Morency, L.-P.: Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 455–463. Springer, Heidelberg (2012)

12. Steel, T., Kuiper, D., Wenkstern, R.Z.: Context- Aware Virtual Agents in Open Environments. In: Sixth International Conference on Autonomic and Autonomous Systems, pp. 90–96 (2010)
13. Chopra-Khullar, S., Badler, N.I.: Where to Look? Automating Attending Behaviors of Virtual Human Characters. In: Autonomous Agents, Seattle, pp. 16–23 (1999)
14. Yeo, S.H., Lesmana, M., Neog, D.R., Pai, D.K.: EyeCatch: Simulating Visuomotor Coordination for Object Interception. In: SIGGRAPH, p. 4 (2012)
15. Li, W., Allbeck, J.M.: Virtual humans: Evolving with common sense. In: Kallmann, M., Bekris, K. (eds.) MIG 2012. LNCS, vol. 7660, pp. 182–193. Springer, Heidelberg (2012)
16. Hill, R.W., Kim, Y., Gratch, J.: Anticipating Where to Look: Predicting the Movements of Mobile Agents in Complex Terrain. In: Autonomous Agents and Multi Agent Systems, Bologna, pp. 821–827. ACM (2002)
17. Feng, A., Huang, Y., Xu, Y., Shapiro, A.: Automating the Transfer of a Generic Set of Behaviors onto a Virtual Character. In: Kallmann, M., Bekris, K. (eds.) MIG 2012. LNCS, vol. 7660, pp. 134–145. Springer, Heidelberg (2012)

A Qualitative Evaluation of Social Support by an Empathic Agent

Janneke M. van der Zwaan, Virginia Dignum, and Catholijn M. Jonker

Delft University of Technology

Abstract. There is a growing interest in employing embodied agents as companions and coaches to achieve beneficial outcomes for users, such as adherence to diets. The goal of our research is to explore how and to what extend embodied agents can provide social support to victims of cyberbullying. To this end, we implemented a prototype of an empathic virtual buddy that uses verbal and nonverbal behavior to comfort users. In order to understand what aspects of interaction with the virtual buddy are perceived as being supportive, we organized a focus group discussion among pedagogical experts. The results indicate that the experts felt supported by interaction with the virtual buddy. In addition, our results demonstrate a method that can be utilized to evaluate embodied agents for vulnerable target audiences (e.g., children), and in sensitive domains (e.g., cyberbullying).

1 Introduction

There is a growing interest in employing embodied agents as companions and coaches. Increasingly, this type of agents is equipped with the ability to understand and express emotions. It is generally assumed that emotional agents provide for an improved interaction [1]. Exactly what 'improved interaction' means depends on the application in question; for companion and coaching agents this often has to do with achieving beneficial effects, such as improving adherence to diets [9].

To what extent the behavior of emotional agents actually affects users remains unclear. In their overview of research into the impact of emotional agents on user attitudes, perceptions, and behavior, Beale and Creed report inconclusive and contradictory results [1]. Following Dehn and Van Mulken [7], Beale and Creed argue that a fine-grained approach is required to better understand the impact of emotional agents on users:

> [I]t is not sufficient to simply ask whether emotional agents are better or worse than unemotional agents. Instead, a better question is that of which kind of emotional expression, expressed in which way, influences which elements of a user's perceptions and behaviour [1].

Quantitative studies are the norm in embodied agents research. All of the 20 papers reviewed by Beale and Creed present quantitative results [1]. However,

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 358–367, 2013.

quantitative methods provide little insight into how and why certain interaction strategies affect user attitudes and perceptions, and are therefore less suitable to provide for the fine-grained analysis that Dehn and Van Mulken, and Beale and Creed stand for. Qualitative research methods seem a better fit to explore what aspects of an agent's behavior contribute to achieving certain effects in users.

In our research, we explore how and to what extent an embodied agent provides social support to children aged 10–14 years old that are victims of cyberbullying. We implemented a prototype that uses strategies employed by human counselors to comfort children in chat conversations. In this paper, we present a qualitative study that aims to understand what aspects of the interaction are perceived as being supportive, what aspects disrupt the perception of support, and how the overall experience of social support can be improved. Since we are dealing with a vulnerable target audience and a sensitive application domain, the evaluation was conducted by pedagogical experts. The experts were invited to try the prototype and give their opinions in a focus group discussion.

In the next section, we review qualitative research on embodied agents. Section 3 describes the application that was evaluated. In section 4, we detail how the focus group was conducted. The results are discussed in section 5. Finally, we present our conclusions.

2 Related Work

Although qualitative research methods are mentioned in literature on the evaluation of affective interfaces and embodied conversational agents (see [6,10,11]), qualitative studies are underrepresented in the field of embodied agents. An exception is Bickmore, who evaluated virtual exercise coach Laura both quantitatively and qualitatively [3]. 28 of 82 participants were subjected to interviews after the experiment was completed. The results of the interviews provide insight into what aspects of the agent's behavior were considered important by the participants, including the repetitiveness of the dialogue and how the agent compares to a human trainer.

To evaluate the appearance of virtual characters in FearNot, a virtual learning environment for the exploration of bullying and coping strategies, Hall et al. analyzed what was discussed during a number of Classroom Discussion Forums (CDFs) [8]. The CDFs were conducted after a day during which 345 children aged 9-11 years participated in a number of sessions related to interaction with robots and virtual agents. The results of the CDFs indicated that the participants liked the visual style of FearNot, but that the animations needed to be improved.

Leite et al. interviewed school children after they played chess against an empathic chess playing robot [12]. To get insight into the role of empathic emotions during the game, the participants were asked to explain how they felt during the game and how they thought the robot felt. The results show that the children understood the robot's empathic behavior.

To investigate what older adults would want to talk about with a virtual companion, Pfeifer Vardoulakis et al. conducted a wizard of Oz study in which 12 participants had daily chats with a virtual agent for a week [14]. During this study, quantitative measures were used together with semi-structured interviews. During the interviews participants gave their opinions on different topics, including their experience with the agent and privacy concerns. Most participants liked interacting with the agent, but were concerned the webcam was recording them when they were not using the system.

Related work shows that qualitative research yields valuable insights. In all studies discussed above, useful details have been discovered. Data generated by qualitative research can be used to construct informed hypotheses that require further empirical evaluation. The related work also shows qualitative methods are often used together with quantitative methods. In these cases, qualitative data are used to illustrate or explain quantitative results.

3 The Application

Our research concerns an embodied conversational agent (ECA) that provides social support to cyberbullying victims[1]. Social support can be defined as communicative attempts to alleviate the emotional distress of another person [5]. Figure 1 shows a screen shot of Robin, the empathic virtual buddy prototype we implemented. The prototype is available online[2].

The user communicates with the buddy by selecting pre-defined response options. To get an impression of the user's emotional state, the AffectButton [4], a tool for explicit emotion input, is used (not depicted in figure 1). The AffectButton shows a rudimentary and gender-neutral face that changes its expression based on the position of the mouse cursor. By clicking the button when it shows the emotional expression the user wants to communicate, the emotion is send to the buddy.

In order to understand, comfort and suggest actions to the user, the buddy combines a conversation and an emotional model. The conversation model specifies the structure and contents of the conversation. The conversation is scripted and proceeds according to the 5-phase model [2]:

1. Welcome: the buddy greets the user
2. Gather information: the buddy asks questions about the cyberbullying incidents
3. Determine conversation objective: the buddy asks what the user wants to achieve with the conversation (options are *to get tips on how to deal with cyberbullying* and *to tell my story*)
4. Work out objective: if the conversation objective is *to get tips on how to deal with cyberbullying* the buddy asks the user his plan for dealing with the bullying and gives tips afterwards. If the conversation objective is *to tell my story*, the buddy asks to whom the user is going to talk. If necessary, the buddy suggests the user to contact an online help line. Afterwards, the buddy offers the user tips
5. Round off: the buddy says goodbye to the user

More details about the conversation model can be found in [15].

[1] We would like to emphasize that our research is focussed on creating the experience of feeling supported, rather than reducing or 'solving' the problem of cyberbullying.
[2] `http://ict1.tbm.tudelft.nl/empathicbuddy/`

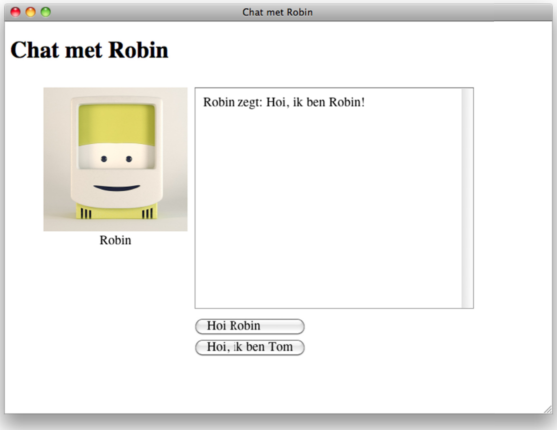


Fig. 1. Screen shot of Robin, the empathic virtual buddy

The emotional model determines when the virtual buddy expresses sympathy, compliments and encourages the user. The buddy executes these support strategies by changing its facial expression and uttering a verbal message. For example, if the user selects the option *I talked to someone about being bullied* to answer the question *How do you respond to bullying?*, happiness is triggered, because in the domain of cyberbullying this is a good response. Subsequently, the buddy's facial expression changes to happy and the message *'Good of you to talk to some one!'* appears in the chat window. During conversation phase 4, the buddy also gives advice and explains how to execute that advice (teaching).

4 Evaluation Method

In order to explore different aspects of the virtual buddy's behavior, we organized a focus group. A focus group is a group discussion in which individual's views and experiences are explored through group interaction [13]. Six pedagogical experts were invited to try the prototype of the virtual buddy and discuss their opinions. The discussion was led by an experienced moderator (11 years' experience in conducting (group)interviews). One of the authors was present as an observant.

Because we were interested in exploring different views on the virtual buddy, experts with different backgrounds were selected to participate. Four of the six participants were experts from practice (i.e., two anti-bullying trainers, an online counselor, and an expert on online counseling and therapy) and the other two had an academic background (i.e., a PhD student working on an online training for cyberbullying victims and a master student interested in cyberbullying). Additionally, two experts from practice and one from academia could be characterized as having experience with Internet technology and online interventions aimed at (cyber)bullying victims, while the other experts were less experienced with technology.

The session was conducted according to the following plan:

1. Welcome and introduction of the participants
2. Introduction of the virtual buddy
3. Associations triggered by the idea of a virtual buddy
4. Interaction with the virtual buddy
5. The virtual buddy in general
6. The virtual buddy's supportive behavior
7. Final remarks

After the moderator welcomed everybody and the participants had introduced themselves, the observant briefly introduced the virtual buddy. To make sure the participants focussed on the buddy's supportive behavior, the observant also explained that the goal of the discussion was to determine to what extend the buddy is able to give social support rather than finding out whether the buddy is able to reduce cyberbullying.

After that, the moderator introduced the first topic 'associations triggered by the idea of a virtual buddy'. This topic was included because, in our experience, people have a lot of ideas and assumptions about the virtual buddy and cyberbullying that might get in the way of an open-minded evaluation of the buddy's behavior. So, in order to allow the experts to focus on the buddy's actual behavior instead of their opinions and views, they were asked to make all ideas and assumptions explicit before interacting with the system. Participants were given a sheet of paper and ten minutes to write down their thoughts.

When the discussion on associations was finished, the experts tried the virtual buddy individually, each on their own laptop. For the interaction, the participants received a scenario on paper. The experts also received two sheets of paper, one to write down their responses to the virtual buddy in general and one to write down their responses to the buddy's supportive behavior.

To make sure feedback was collected on all important components of the virtual buddy, the moderator prompted the participants to give feedback on the impression of the program as a whole, the virtual character's appearance, the quality of the advice, the conversation structure, and the use of predefined response options; and for the discussion on the supportive behavior the participants were prompted on the advice and teaching, and the buddy's encouragements, compliments and sympathetic remarks.

The focus group was concluded by one final round of remarks. In total, the focus group took approximately three hours. The whole session was audiotaped.

5 Results

During the focus group, the experts discussed many different aspects of the prototype. This section is focussed on the experts' evaluations of the virtual buddy's supportive behavior, and their recommendations to improve the experience of social support.

5.1 Emotional Support

Multiple participants expressed they felt supported by interacting with the virtual buddy.

> *I think the conversation is very comforting, it is a safe way to get support, even though it is just a computer that says these nice things, like 'I'm sorry for you' and things like that. I think it is very nice to hear these things.*[3] (E3; practice)

> *The emotional expressions were very understanding, I just felt I was being listened to, even if it was by a computer.* (E5; academia)

The participants substantiated their claims about feeling supported by referring to specific behavior of the virtual buddy. Three different aspects were mentioned: the verbal expressions of emotions, Robin's facial expressions, and the combination of verbal and facial expressions. The verbal expression of emotions was commended:

> *I really liked that Robin said: 'It was brave of you to stand up to the bully', because things like that are very supportive of course.* (E3; practice)

However, the participants agreed the number of verbal empathic responses should be increased:

> *It has already been said a few times that there should be more empathic responses, for example Robin asks, 'do you know the bully?' and after the response immediately there is another factual question, 'how long have you been bullied?' It is better to first give a response to every answer, such as 'that's annoying' or 'that's awful'.* (E3; practice)

Additionally, the buddy's facial expressions were mentioned in connection to feeling supported; see for example expert 5's statement that was quoted at the beginning of this section. In general, the emotions selected for expression by the buddy's emotional model seem to be appropriate:

> *I noticed that Robin is already quite adaptive, even in this version, he is flexible. He quickly adapts to the emotion content of what is being said. And I think that that's going very well.* (E2; practice)

Nevertheless, the participants also reported a number of emotion mismatches. Most importantly, one participant felt the facial expressions change too quickly back to happy:

> *And after you answered a question, he responds to that and then he starts laughing again. But that's very strange! Because you've just told him that you are very sad and then he says, 'Oh, I don't like that you're sad' and then he laughs again. I think he should remain sad or neutral a little longer.* (E1; practice)

[3] All quotes have been translated from Dutch by the authors.

Also, a sad face was considered an inappropriate response to being threatened:

> *I entered I was being threatened, then the corners of Robin's mouth went down, and I was wondering whether that is the right response, because when someone tells you he is being threatened you should show concern, but now it seems it makes him sad.* (E3; practice)

At some point in the conversation, the buddy asks the user what he did to try to stop the bullying. One of the response options to this question is *retaliate*. If this option is selected, Robin responds by saying that you should not retaliate. Robin's next question is about whether the user's actions helped to stop the bullying and if the user says the bullying stopped, he is happy for the user. One of the participants remarked this response is inconsistent:

> *First Robin says that you are not allowed to retaliate, but then he is glad it helped!* (E1; practice)

To substantiate the social support felt during the interaction, the participants also referred to the added value of combining verbal utterances with facial expressions:

> *I find Robin left a very sympathetic impression, precisely because of the combination of language and those different expressions. [...] I say it works better than text-only; it really struck me that that did something more to me than just reading.* (E2; practice)

In summary, the participants recognized the Robin's attempts to provide emotional support. They also felt supported by the buddy's behavior, in particular by the supportive verbal utterances, the facial expressions, and the combination of verbal utterances and facial expressions. To improve the virtual buddy, we will add more supportive verbal utterances, and update the emotional model to remove emotion mismatches. Additionally, we will add a neutral facial expression, and slow down the decay rate of emotions.

5.2 Information Support

On the whole, the participants approved of the advice given by Robin. Furthermore, the explanation of how to block a contact on MSN was '*crystal clear*' (E1; practice). Additionally, this expert suggested screen shots might be used to further clarify technical know-how when needed.

One participant suggested the contents of the advice should depend on the severity of the user's situation:

> *It can be very serious if you are being threatened. In this case ignoring it might not be a sensible solution. So, maybe you should allow the user to indicate the severity of the situation.* (E5; academia)

The participants were also missing some domain-specific information in the conversation. An important way to comfort a bullying victim is to make clear he is not the only one who is being bullied. This '*subjective norm*' (Expert 4; academia) should be communicated to the victim. They also suggested Robin should explain that bullying is a complicated problem and that no perfect solution exists. Another comforting strategy is to explain to bullying victims that they can learn to deal with bullying (e.g., by giving assertive responses). Victims often feel relieved when this is made clear to them. Finally, advice should be presented as suggestions that might not work for every person.

Both the domain-independent and domain-specific suggestions will be incorporated in the next version of the prototype.

5.3 Other Ways to Communicate Social Support

During the discussion, the participants came up with additional ways to express support. The first suggestion was to use humor to put things into perspective. When making this suggestion, the participants realized very well using humor is tricky:

> *I think you are right when you say humor is important, but it is so hard to do it right, it's easy to mess things up.* (E3; practice)

Also, it was recognized that it depends on the specific situation whether humor or other ways to put things into perspective are appropriate. To express humor and other '*dangerous emotions*' (E2; practice) the participants suggested several visual means, including winks and other smileys, and holding up signs.

According to the experts, an important threat to the user experience is the certainty with which Robin assumes he is correct, when he responds to user input, for example when interpreting the emotional state that is communicated with the AffectButton. One expert suggested Robin should check its interpretation:

> *When I had to use the face to choose an emotion, I thought I selected a frightened expression and then Robin said: 'It doesn't seem to bother you that much'. [...] And I thought, to solve it you could ask, 'Is that correct or am I wrong?'* (E3; practice)

Also, Robin should check whether the advice he gives is acceptable to the user.

The experts made several other suggestions to communicate social support, including expressing other, more 'dangerous' emotions, using more visual means to express emotions, and requesting feedback on certain assertions made by the buddy. In order to improve the virtual buddy, we will incorporate requests for feedback at the suggested moments in the conversation. Currently, we do not have the intention of including humor and other 'dangerous' emotions or other means of expressing emotions, because we prefer to refine the original design before adding new ideas.

6 Conclusion

The aim of the study presented in this paper was to assess what aspects of interaction with the virtual buddy prototype were perceived as being supportive, what aspects disrupt the perception of support, and how the overall experience of social support can be improved according to a panel of pedagogical experts. However exploratory, this study offers insight into how embodied agents can comfort users. Additionally, we gathered detailed feedback on how to improve the prototype. The results demonstrate that the experts recognized the different social support strategies employed by the virtual buddy, including the supportive statements (sympathy, compliments, and encouragement), different facial expressions, and advice and teaching. The experience of being supported was contributed to Robin's supportive verbal utterances, the different facial expressions, and the combination of verbal utterances and facial expressions. The embodiment clearly has added value for the application.

The second contribution of this paper is the demonstration of a method that can be used to evaluate embodied agents for vulnerable target audiences (e.g., children), and in sensitive domains (e.g., cyberbullying). In particular, we invited adult experts to try the system. The detailed feedback we gathered shows this qualitative method yields valuable insights on how a certain emotional experience –in our case the experience of social support– can be provided for. We believe that explicitly discussing associations with the buddy before trying it facilitated the process of collecting constructive feedback.

The approach of having adults evaluate the buddy instead of the intended target audience (children aged 10-14) is justified given the vulnerability of the target audience, and the sensitivity of the domain of cyberbullying. Before children can be involved in the evaluation, we need to be certain the system does what it is supposed to do. Because the results indicate that the prototype works and that social support is communicated, we are currently gathering feedback from the target audience in one-on-one interviews.

Acknowledgements. This work is funded by NWO under the Responsible Innovation (RI) program via the project 'Empowering and Protecting Children and Adolescents Against Cyberbullying'.

References

1. Beale, R., Creed, C.: Affective interaction: How emotional agents affect users. International Journal of Human-Computer Studies 67(9), 755–776 (2009)
2. de Beyn, A.: In gesprek met kinderen:de methodiek van de kindertelefoon. In: SWP (2003)
3. Bickmore, T.: Relational Agents: Effecting Change through Human-Computer Relationships. PhD thesis, Massachusetts Institute of Technology (2003)
4. Broekens, J., Brinkman, W.P.: Affectbutton: Towards a standard for dynamic affective user feedback. In: Affective Computing and Intelligent Interaction, ACII (2009)
5. Burleson, B.R., Goldsmith, D.J.: How the Comforting Process Works: Alleviating Emotional Distress through Conversationally Induced Reappraisals. In: Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts, pp. 245–280. Academic Press (1998)

6. Christoph, N.: Empirical Evaluation Methodology for Embodied Conversational Agents. In: From Brows to Trust: Evaluating Embodied Conversational Agents, pp. 67–90. Kluwer Academic Publishers (2005)
7. Dehn, D.M., Van Mulken, S.: The impact of animated interface agents: a review of empirical research. International Journal of Human-Computer Studies 52(1), 1–22 (2000)
8. Hall, L., Vala, M., Hall, M., Webster, M., Woods, S., Gordon, A., Aylett, R.: Fearnot's appearance: Reflecting children's expectations and perspectives. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 407–419. Springer, Heidelberg (2006)
9. Blanson Henkemans, O.A., van der Boog, P.J.M., Lindenberg, J., van der Mast, C., Neerincx, M., Zwetsloot-Schonk, B.J.H.M.: An online lifestyle diary with a persuasive computer assistant providing feedback on self-management. Technology & Health Care 17, 253–257 (2009)
10. Höök, K.: User-Centred Design and Evaluation of Affective Interfaces. In: From Brows to Trust: Evaluating Embodied Conversational Agents, pp. 127–160. Kluwer Academic Publishers (2005)
11. Höök, K., Isbister, K., Westerman, S., Gardner, P., Sutherland, E., Vasalou, A., Sundström, P., Kaye, J.J., Laaksolahti, J.: Evaluation of affective interactive applications. In: Cowie, R., Pelachaud, C., Petta, P. (eds.) Emotion-Oriented Systems, Cognitive Technologies, pp. 687–703. Springer, Heidelberg (2011)
12. Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2012, pp. 367–374. ACM, New York (2012)
13. Litosseliti, L.: Using Focus Groups in Research. Continuum Research Methods. Continuum (2003)
14. Vardoulakis, L.P., Ring, L., Barry, B., Sidner, C.L., Bickmore, T.: Designing relational agents as long term social companions for older adults. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 289–302. Springer, Heidelberg (2012)
15. van der Zwaan, J.M., Dignum, V., Jonker, C.M.: A conversation model enabling intelligent agents to give emotional support. In: Ding, W., Jiang, H., Ali, M., Li, M. (eds.) Modern Advances in Intelligent Systems and Tools. SCI, vol. 431, pp. 47–52. Springer, Heidelberg (2012)

All Together Now

Introducing the Virtual Human Toolkit

Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch

Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Playa Vista, CA 90094, USA
{hartholt,traum,marsella,shapiro,
stratou,leuski,morency,gratch}@ict.usc.edu

Abstract. While virtual humans are proven tools for training, education and research, they are far from realizing their full potential. Advances are needed in individual capabilities, such as character animation and speech synthesis, but perhaps more importantly, fundamental questions remain as to how best to integrate these capabilities into a single framework that allows us to efficiently create characters that can engage users in meaningful and realistic social interactions. This integration requires in-depth, inter-disciplinary understanding few individuals, or even teams of individuals, possess. We help address this challenge by introducing the ICT Virtual Human Toolkit[1], which offers a flexible framework for exploring a variety of different types of virtual human systems, from virtual listeners and question-answering characters to virtual role-players. We show that due to its modularity, the Toolkit allows researchers to mix and match provided capabilities with their own, lowering the barrier of entry to this multi-disciplinary research challenge.

Keywords: Virtual Humans, Embodied Conversational Agents, Architectures, Standards, Real-Time Integrated Systems, Toolkits, Audio-Visual Sensing, Natural Language Processing, Nonverbal Behavior.

1 Introduction

Virtual humans, autonomous digital characters who interact verbally and nonverbally with users, can be powerful tools in a wide range of areas, including the teaching of interpersonal skills, cognitive science studies, military training, medical education, and entertainment. In fact, virtual humans have moved from the lab to actual deployment in a variety of fields in recent years, from virtual patients [1, 2] to pedagogical agents [3, 4] and military training [5, 6].

[1] https://vhtoolkit.ict.usc.edu

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 368–381, 2013.

However, while virtual humans have advanced in both capability as well as applicability, they are still in their infancy. Realizing their full potential requires 1) compelling characters that can engage users in meaningful and realistic social interactions, and 2) an ability to develop these characters effectively and efficiently. There are several challenges associated with these goals.

First, in order for virtual humans to be effective, they need to exhibit a range of capabilities, simulating those of real humans. This includes the ability to perceive human behavior (particularly communicative behavior aimed at the virtual humans), to process and understand this behavior, and to reason and produce appropriate (verbal and nonverbal) behaviors. Specialized knowledge and considerable resources may be needed to significantly advance any of these virtual human abilities.

Second, it is important that these abilities not work only in isolation, they also need to be integrated into a larger system (and further, into systems of systems), where they can inform, influence and strengthen each other. For example, the meaning of a spoken word like "yeah" might have a different meaning if it is accompanied by a head nod and a smile versus a head shake and frowning. This interdependence of functionalities can create obstacles for research, since individual researchers seeking to advance a specific topic (e.g., natural language understanding) might need to work in a large team with expertise in all other relevant aspects of the system, in order to solve the full problem. On the other hand, the integrated nature of virtual humans also creates novel research opportunities. Researchers can explore which integrated abilities are essential and desired for which type of interaction, what minimal and preferred dependencies exist between these abilities, how they affect each other, and how they can best leverage each other in order to achieve greater effectiveness.

Third, even with the appropriate knowledge and resources available, virtual humans can still be costly to develop. They are large, complex systems, often lacking specific frameworks or solid standards. Furthermore, certain principles may only be understood in a narrow context and can be difficult to generalize across multiple domains. This limits the ability to re-use knowledge and assets, often resulting in the need to start new characters or systems from scratch.

To address these challenges, we introduce the *ICT Virtual Human Toolkit,* which is designed to aid researchers with the creation of embodied conversational agents. The Virtual Human Toolkit enables further research by offering a collection of modules, tools, and libraries, as well as a framework and open architecture that integrates these components. The Toolkit provides a solid basis for the rapid development of new virtual humans, but also serves as an integrated research platform to enable context-sensitive research in any of the Virtual Human subfields, taking advantage of and examining the impact on other system modules. Rather than focusing on a specific type of agent, the Toolkit offers a flexible framework for exploring the vast space of different types of agent systems.

The Toolkit is based on over a decade of multi-disciplinary science and contains a mix of research and commercial technologies to offer full coverage of subareas including speech recognition, audio-visual sensing, natural language processing, dialogue management, nonverbal behavior generation & realization, text-to-speech and rendering. Example virtual humans are included to illustrate how components can work together and to enable the sharing of these assembled systems to the research community.

In this paper, we will show how the Virtual Human Toolkit provides an integrated platform for a variety of different virtual human architectures and how these architectures can be applied to a wide range of research efforts. Section 2 discusses the background of inter-disciplinary virtual human research. Section 3 explains the overall architecture and APIs on which the Toolkit is based, while Section 4 discusses the specific modules that are included in the Toolkit. In section 5 we explore how the Toolkit and related technologies have been used to create a range of different virtual human systems. Section 6 ends with a conclusion and future work.

2 Background

Reducing barriers to virtual human research requires shared tools and architectures and the Virtual Human Toolkit builds on several attempts to address this need. Efforts to develop and share individual capabilities include automated speech recognition [7], perception [8], task modeling [9], natural language generation [10], animation [9, 11, 12, 13, 14], and text-to-speech systems [15].

Some standardization efforts have attempted to consolidate these. SAIBA [16] specifies on a high level the generation of multimodal behavior. It aims to define an interface between the intent planning and behavior planning phases, called the Function Markup Language [17], and in between the behavior planning and behavior realization phases, called the Behavior Markup Language (BML) [18]. Several BML realizers are publically available, including LiteBody [9], GRETA [11], Elckerlyc [12], SmartBody [13], and EMBR [14]. While these realizers often depend on custom extensions to the BML standard, they have been shown to be compatible in real-time larger systems [19].

Further work has focused on including these capabilities and standards in larger frameworks. LiteBody and DTask form the basis for “Relational Agents” [9]. These agents aim to form long-term, social-emotional relationships with users. While these agents exhibit social verbal and nonverbal behavior, they are usually limited in their natural language processing and are often confined to 2D representations.

The Virtual People Factory [1] allows for the creation of virtual humans by domain experts themselves, rather than system experts. It is mostly used for creating virtual patients for medical and pharmacy education. Using a crowd sourcing approach focused on natural language interaction, new characters can be created more rapidly than through more traditional authoring methods.

GRETA [11] is a SAIBA compliant embodied conversational agent that focuses less on natural language interaction and more on affective nonverbal behavior generation and realization. In particular, it employs a model of complex facial generation.

The SEMAINE project [20] aims to integrate various research technologies, including some of the above, into creating a virtual listener. The emphasis is on perception and back-channeling rather than deep representations of dialogue.

Virtual Humans are not the only area of research in which these kinds of frameworks are desired. For example, the Robotic Operating System [21] defines structures that facilitate distributed development and sharing of robotic capabilities.

While these systems are both capable and successful, they typically focus on a limited subset of capabilities. In addition, few standards exist across capabilities, making efforts for further interoperability between them cumbersome.

3 Architecture, APIs and Virtual Human Capabilities

The Virtual Human Toolkit is the next step towards more integrated frameworks. It aims to offer a more comprehensive set of integrated capabilities than has been previously attempted, within a framework that allows for the rapid creation of new characters and systems. The Virtual Human Toolkit is an instantiation of a more general, modular Virtual Human Architecture, see Figure 1. This architecture defines at an abstract level the capabilities of a virtual human and how these capabilities interact. Not every virtual human system will include all the capabilities, and some will implement capabilities to a greater or lesser extent. The architecture also allows for multiple implementations of a certain capability and simple substitution of one implementation for another during run-time, facilitating the exploration of alternative models for realizing individual capabilities. Thus the general architecture can be specialized in many different ways, as we'll discuss in section 5. As presented in section 4, the Toolkit contains a set of modules providing at least one possible realization of each the capabilities.

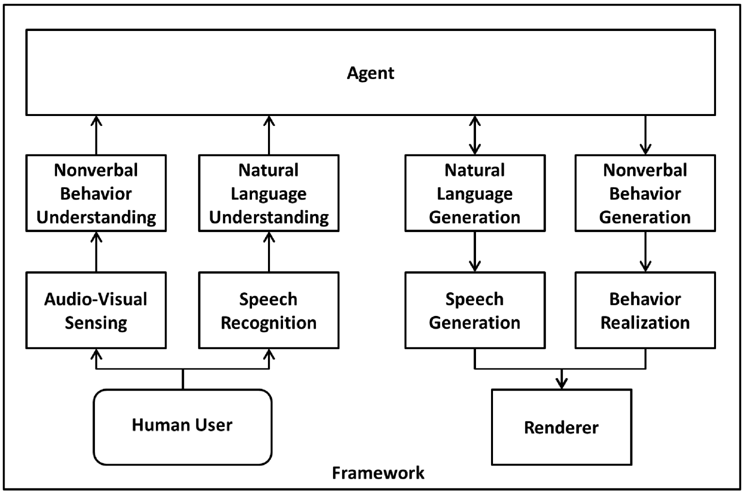


Fig. 1. The Virtual Human Architecture

A virtual human system typically contains a subset of the capabilities depicted in Figure 1. A human user interacts with the system, which transforms the user's speech to a textual representation using speech recognition. This text is translated to a semantic representation through natural language understanding, which a dialogue manager within the agent can reason with. Audio-visual sensing relies on sensory input and has the ability to localize features and recognize specific expressions of nonverbal communication. Nonverbal behavior understanding combines information from different modalities in order to link certain observations through tracking and recognition to higher-level nonverbal communicative behaviors. Based on these inputs and internal state, the agent will create a communicative intent. This intent is further fleshed out both verbally and nonverbally through natural language generation and nonverbal behavior generation processes. Speech can be generated on the fly (e.g. text-to-speech) or be pre-recorded audio. Behavior realization will synchronize all behaviors (speech, gestures, lip synching, facial expressions, etc.) for a renderer to show.

Table 1. Main elements of the Virtual Human API

API	Producer	Consumer	Description
vrSpeech	Speech Recognition	Natural Language Understanding	Text of partial or finalized user speech; also prosody, emotion.
vrNLU	Natural Language Understanding	Agent	Semantic representation of user's verbal input.
vrPerception	Nonverbal Behavior Understanding	Agent	Representation of user's nonverbal behavior in PML [34].
vrGenerate	Agent	Natural Language Generation	Communicative intent.
vrGeneration	Natural Language Generation	Agent	Surface text.
vrExpress	Agent	Nonverbal Behavior Generation	Communicative and functional intent, in FML and BML.
vrSpeak	Nonverbal Behavior Generation	Behavior Realization	Instructions of desired behavior in BML.

Capabilities are realized through specific modules, most of which communicate with each other through a custom messaging system called VHMsg, build on top of ActiveMQ [39]. Messages are typically broadcasted, although there are intended consumers. Libraries have been built for several languages, including Java, C++, C#, Lisp and TCL, so that developers have a wide latitude in developing new modules that can communicate with the rest of the system. There is a standard set of message types used by existing modules (see Table 1), and it is very easy to create new message types, as needed for new kinds of communications. Systems typically include a customizable launcher application, which allows launching, monitoring, and quitting of individual modules. The launcher can be configured with different sets of modules (to support different systems), or with different versions, to support individual customization and experimentation. There is also a logger that keeps track of all VHMsg traffic, and allows customized reporting. These three components together allow for easy reconfiguration and experimentation with specific system architectures.

4 Toolkit Provided Modules

The Toolkit contains a variety of specific modules that collectively cover the areas of speech recognition, audio-visual sensing, natural language processing, dialogue management, and nonverbal behavior generation & realization. See Figure 2 for how these are assembled within the context of the Toolkit. The modules are discussed in more detail below. Together with available authoring and debugging tools they can be used immediately to create new so-called question-answering characters (section 5.1). The Toolkit contains two examples of such characters, called Brad and Rachel (see Figure 3), as well as several other characters. The Toolkit's main target platform is Windows, with limited support for MacOS and Linux. Further support for these platforms, as well as Android and iOS is in development.

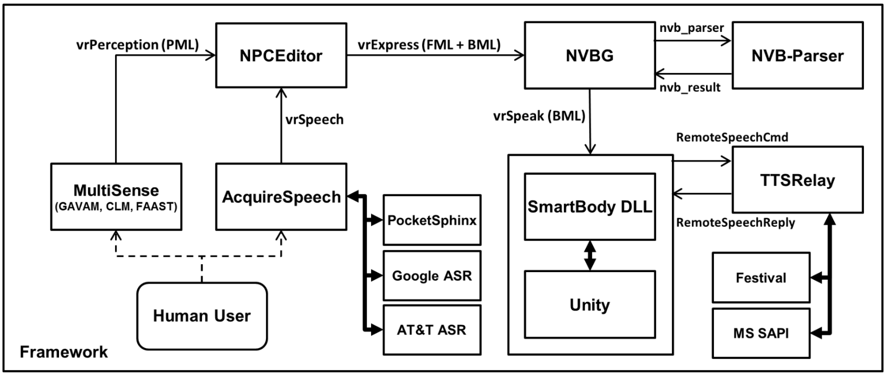


Fig. 2. Virtual Human Toolkit Architecture; regular lines are messages, bold lines direct links

Fig. 3. Rachel (left) and Brad (right)

4.1 MultiSense

MultiSense provides the capabilities of both audio-visual sensing and nonverbal behavior understanding, as illustrated in Figure 1. The output messages are broadcast using the Perception Markup Language (PML) [34]. MultiSense is a multimodal sensing framework which is created as a platform to integrate and fuse sensor technologies and develop probabilistic models for human behavior recognition. MultiSense tracks and analyzes facial expressions, body posture, acoustic features, linguistic patterns and higher-level behavior descriptors (e.g. attention, fidgeting). MultiSense can provide quantitative information about the human's behavior that enhances training and diagnostic scenarios (e.g., public speaking [51] or psychological distress assessment [34] and section 5.3, below). MultiSense is designed with a modular approach, including synchronized capture of multiple modalities such as audio and depth image via Microsoft Kinect sensor and RGB video via webcam device. MultiSense includes updated components from the Social Signal Interpretation framework (SSI) [30] but also integrates multiple tracking technologies including CLM-Z FaceTracker [31] for facial tracking (66 facial feature points), GAVAM HeadTracker [32] for 3D head position and orientation, skeleton tracking by Microsoft Kinect SDK [52] and FAAST [33] for skeleton action coding. MultiSense utilizes a multithreading architecture enabling these different technologies to run in parallel and in real-time. Moreover, the framework's synchronization schemes allow for inter-module cooperation and information fusion. By fusing the different tracker results one can create a multimodal feature set that can be used to infer higher level information on perceived human behavioral state such as attentiveness, agitation and agreement.

4.2 NPCEditor

In the current distribution of the Toolkit, natural language understanding and agent functions such as dialogue management and output decisions are handled by the *NPCEditor* [22]. At the core of the NPCEditor is a statistical text classification algorithm that selects the character's responses based on the user's utterances. A character designer specifies a set of responses and a set of sample utterances that should produce each response. The algorithm analyzes the text of the sample utterances and of the responses and creates a mathematical description of the "translation relationship" that defines how the content of an utterance is mapped to the content of a response. When NPCEditor receives a new (possibly unseen) utterance, it uses this translation information to build a statistical language model of what it believes to be the best response. It then compares this representation to every stored response and returns the best match. The details of the algorithm can be found in [23]. We have shown the algorithm to be effective for constructing limited domain conversational virtual characters [22]. The NPCEditor also contains a dialogue manager that specifies how to use the classifier results, as well as tracking user-defined aspects of dialogue context. There are several provided dialogue managers, as well as a scripting ability for users to create their own dialogue managers.

4.3 NVBG

The Toolkit uses the *NonVerbal Behavior Generator* (NVBG) [24] to plan the character's nonverbal responses. The NPCEditor invokes character speech by sending the surface text to NVBG. In addition, it can receive optional functional markup that indicates the communicative function of the dialogue and the emotions/attitudes of the speaker. NVBG analyzes the text and functional markup to propose nonverbal behaviors. The resulting behavior set is communicated as BML [18]. Note that while conceptually the NVBG receives pure FML [17], in practice, the need for the surface text requires the addition of BML. NVBG is a rule-based behavior planner designed to operate flexibly with the information it is provided: it generates behaviors given the information about the agent's mental processes and communicative intent, but in the absence of such information infers communicative functions from a surface text analysis, using a parser. The rules within NVBG determine which nonverbal behaviors should be generated in a given context and were crafted using psychological research on nonverbal behaviors as well as analyses of annotated corpora of human nonverbal behaviors.

4.4 SmartBody

Behavior realization is achieved by *SmartBody*, [13]. It realizes behavior requests as a suitable animation sequence that conforms to the constraints of a BML request, such as synchronizing the emphasis phase of a gesture with a particular word of speech. The BML specification is heavily focused on synchronizing behaviors for a conversational agent: speech, gaze, gesturing and head movements. SmartBody uses additional BML parameters to control other aspects of character motion, such as locomotion and object manipulation [25], as well as parameters to tune the performance of other behaviors, such as the speed of the joints that a character will use to affix their gaze on another object. The generation of animation for a virtual character requires complex interaction between various body parts in order to adhere to the animation constraints and achieve a satisfactory performance. SmartBody uses various methods to handle animation constraints, both using procedurally-generated control mechanisms [26], as well as by using motion data either captured or hand-generated by digital artists. For example, gazing uses an inverse-kinematics based approach on which additional motion is then layered [27]. Locomotion and reaching use large amounts of motion data to handle full-body control that is difficult to do procedurally [28]. In addition, SmartBody uses a retargeting system that allows the transfer of groups of data designed for one character onto another [29].

4.5 Other Modules and Tools

Users can interact with characters by typing in text and by using Automated Speech Recognition (ASR). There is a module called *AcquireSpeech* (see Figure 2), which connects microphones and ASRs to the Toolkit with a common API, so other modules do not have to be adapted individually. There are interfaces for a number of 3rd party

ASRs (including PocketSphinx [7], Google ASR and AT&T Watson). The main rendering platform of the Toolkit is Unity [35], a proprietary game engine which offers a free version. The Toolkit contains a variety of character art assets and backgrounds. Ogre [36] is included as an open source game engine example. Text-to-speech is provided by a single *TTSRelay* module that can interface to several text-to-speech engines, including Festival [15], CereVoice [37] and MS SAPI [38]. A variety of tools help with authoring, developing, configuring and debugging virtual human characters and systems. In particular, the *Machinima Maker* allows authors to create cut-scenes in Unity. An integrated authoring tool, called *VHBuilder*, abstracts some of the more advanced capabilities of the NPCEditor and NVBG, offering novices a simplified way to quickly author basic virtual humans.

5 Creating New Virtual Human Systems

Due to its modularity, the Virtual Human Toolkit is both configurable and extendable. Large scale permutations of module combinations and their individual configurations can be stored as system profiles, with smaller variations saved as user profiles. Researchers can mix and match existing Toolkit modules with their own, when adhering to the defined messaging API's. Given this flexibility, the Virtual Human Toolkit offers broad support for a range of virtual human systems. This enables researchers to explore and develop best practices for the research, design and development of virtual humans. In the remainder of this section we will explore several types of virtual humans and how the Toolkit can support some of these systems.

5.1 Question-Answering Characters

A common type of agent is the *question-answering character* [22]. Such agents offer a user-driven, interview-type style of conversation where a single user question is answered with a single character response. There is little to no maintenance of dialogue state nor are there advanced turn-taking strategies. These systems focus on providing the user with particular information within a given domain, often leading to a fixed set of pre-defined responses from one or more characters. The Toolkit supports these types of characters through the NPCEditor, NVBG and SmartBody modules. Several related systems [40, 41, 42] have been deployed and evaluated.

5.2 Virtual Listeners

Virtual listeners are agents that aim to simulate human listening behavior, in particular verbal and nonverbal back-channeling. These systems are typically user-driven, one-to-one and face-to-face. Input is commonly nonverbal only, i.e. a user's head movements, length and prosody of speech, etc. The agent itself is often a rule-based system, matching certain input patterns to pre-defined character behavior or generated behavior (e.g. mirroring of head movements, nodding). Example virtual listeners are Rapport [43] and SEMAINE [20], and a public speaking trainer [51]. The Rapport system is included in the Virtual Human Toolkit, using a predecessor of GAVAM (now part of MultiSense) and a custom rule selector. It is based on psycholinguistic theory and was designed to create a sense of rapport between a human speaker and virtual human listener. It has been used in many studies to gather evidence that it increases speaker fluency and engagement. The Rapport system exemplifies how our approach enables both the re-use of existing knowledge and technologies as well as the sharing of results with the larger research community.

5.3 Virtual Interviewers

With *virtual interviewers*, the focus of the conversation lies on gathering information from or assessing the human user. Interactions are driven by the agent or are mixed-initiative. Shifting some of the conversational burden to the agent requires increased dialogue management and natural language understanding capabilities. Since the 'interview' is executed within a known domain, agent responses can be crafted up-front. An example of a virtual interviewer is the SimCoach system, a web-based guide that helps navigate healthcare related resources [44]. Its dialogue manager, FLoReS [45], allows for the creation of forward looking, reward seeking dialogues. An initial integration of FLoReS with the Toolkit has been completed and will be released shortly. An example of a hybrid virtual interviewer / listener is SimSensei, a virtual human platform to aid in recognition of psychological distress [35]. SimSensei enables an engaging face-to-face interaction where the virtual human automatically reacts to the perceived user state and intent, through its own speech and gestures. From the humans' signals and behaviors, indicators of psychological distress are inferred to inform a healthcare provider or the virtual human. SimSensei is in active development and advances several modules, including MultiSense, FLoReS, Cerebella (the successor to NVBG) and SmartBody.

5.4 Virtual Role-Players

Virtual humans can be particularly effective at facilitating interactive dramas or training scenarios in which a user (or player) must interact with other characters. They have advantages over human role-players due to their inherent consistency, 24-7 availability and ability to portray elements that are difficult or impossible for real humans (e.g. certain wounds, effects of a stroke, etc.). Examples include the INOTS [6] and ELITE training systems for the Navy and Army respectively, in which a single user practices interpersonal skills with a virtual human role-player as part of a larger class. It combines the NPCEditor and SmartBody with a branching storyline and intelligent tutor. Gunslinger [46] is a mixed-reality, story-driven experience, where a single participant can interact verbally and nonverbally with multiple virtual characters that are imbedded in a physical saloon. Together with the NVBG and SmartBody, it uses a version of the NPCEditor which has been extended to incorporate the notion of hierarchical interaction domains, comparable to a state machine. In addition, it receives perception input which is treated as an additional token in the statistical analysis.

5.5 Virtual Confederates

Finally, virtual humans are gaining interest as a methodological tool for studying human cognition, including the use of *virtual confederates*. Virtual humans not only simulate the cognitive abilities of people, but also many of the embodied and social aspects of human behavior more traditionally studied in fields outside of cognitive science. By integrating multiple cognitive capabilities to support real-time interactions with people, virtual humans create a unique and challenging environment within which to develop and validate cognitive theories [48, 49, 50].

6 Conclusions and Future Work

We have shown how the Virtual Human Toolkit helps to address several challenges in virtual human research. First, a full virtual human system requires many different capabilities, and the Toolkit contains modules that cover audio-visual sensing, nonverbal behavior understanding, speech recognition, natural language processing, nonverbal behavior generation & realization, text-to-speech and rendering. Second, individual capabilities need to be integrated into a larger framework, which the Toolkit offers in the form of a reference architecture and related APIs. As such, it provides a rich context in which to embed individual research efforts and delivers a flexible framework that enables the exploration of a wide range of different virtual human systems. Finally, it lowers the effort associated with creating virtual humans by providing a suite of modules and tools that facilitate the rapid development of new characters and by promoting re-use of assets. These reduce the required knowledge and resources to develop virtual humans, lowering the barrier of entry into further inter-disciplinary research. Empirical evidence suggests a computer literate user with no particular virtual human or computer science background can build a limited question-answering character in less than a day, with more advanced characters taking up to several weeks. We plan to more formally evaluate these efforts within the year. While creating new virtual human architectures and systems can be done by individuals, depending on the complexity this may require a small group of specialists with a computer science background.

The Toolkit and related technologies have already been used in several dozen research and applied projects at ICT. Since its release to the community, it has seen download requests from close to 400 individuals and has been used as a teaching tool in several classes.

While the Toolkit offers a comprehensive framework for virtual human research and development, it is not without its limitations. The provided NPCEditor component focuses on statistical text classification rather than deep understanding of natural language and dialogue management. This limits verbal interactions to mostly question-answering type conversations. We aim to address this by releasing both the hierarchical interaction domain plugin for the NPCEditor as well as the FLoReS dialogue manager shortly. In addition, the Toolkit is lacking certain abilities, in particular task driven behavior, emotion modeling and persistent memory. While these areas are of interest in our basic research, we feel they are currently less appropriate for inclusion with the more applied Toolkit.

Our current focus is on expanding the capabilities of the Toolkit. In addition to including more advanced dialogue management, we aim to more tightly integrate MultiSense, to expand the available character library, and to increase the number of supported platforms, including full support for Mac OS, Android, iOS and the web.

Future work is aimed at addressing many of the lessons learned over the past decade. With Cerebella, the successor to the NVBG, we aim to expand our model of nonverbal behavior and provide a more powerful way to describe and generate this behavior, including the ability for a character to keep a consistent gesture space. This will require the implementation of gesture co-articulation as well as extension to BML. In addition, we aim to expand the role of nonverbal behavior understanding in order to provide a richer context to the agent. Furthermore, we will continue to investigate methods that support exploration of different configurations within the modular architecture, both at the level of modules themselves as well as the implementation of different gradations of capability compliance (e.g. per utterance speech recognition results vs. partial speech results vs. continuous speech). This requires a refinement of our current distributed messaging model, creating a balance between a structured and well-defined API on the one hand and a non-restrictive and expandable infrastructure that allows for experimentation on the other. Finally, we aim to more concretely define separate categories, or genres, of virtual humans. This allows for the creation of best practices, methodologies and supporting tools that enable more rapid development of virtual human systems.

Acknowledgements. We like to acknowledge all the hard work of our many colleagues at USC ICT without whom the Virtual Human Toolkit would not have been possible. The effort described here has been sponsored by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government; no official endorsement should be inferred.

References

1. Rossen, B., Lok, B.: A crowdsourcing method to develop virtual human conversational agents. International Journal of HCS, 301–319 (2012)
2. Bickmore, T., Bukhari, L., Vardoulakis, L.P., Paasche-Orlow, M., Shanahan, C.: Hospital buddy: A persistent emotional support companion agent for hospital patients. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 492–495. Springer, Heidelberg (2012)
3. D'Mello, S.K., Graesser, A.C.: AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems 2(4), Article 23 (2012)
4. Lane, H.C., Noren, D., Auerbach, D., Birch, M., Swartout, W.: Intelligent Tutoring Goes to the Museum in the Big City: A Pedagogical Agent for Informal Science Education. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS, vol. 6738, pp. 155–162. Springer, Heidelberg (2011)
5. Johnson, W.L., Valente, A.: Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures, pp. 72–83 (2009)
6. Campbell, J., Core, M., Artstein, R., Armstrong, L., Hartholt, A., Wilson, C., Georgila, K., Morbini, F., Haynes, E., Gomboc, D., Birch, M., Bobrow, J., Lane, H., Gerten, J., Leuski, A., Traum, D., Trimmer, M., DiNinni, R., Bosack, M., Jones, T., Clark, R., Yates, K.: Developing INOTS to support interpersonal skills practice. In: Proceedings of the Thirty-second Annual IEEE Aerospace Conference, pp. 1–14 (2011)
7. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for hand-Held Devices, vol. 1, pp. 185–188 (2006)
8. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The Computer Expression Recognition Toolbox (CERT). In: Proc. IEEE International Conference on Automatic Face and Gesture Recognition (2011)
9. Bickmore, T., Schulman, D., Shaw, G.: DTask and LiteBody: Open Source, Standards-based Tools for Building Web-deployed Embodied Conversational Agents. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 425–431. Springer, Heidelberg (2009)

10. Stone, M.: Specifying Generation of Referring Expressions by Example. In: AAAI Spring Symposium on NLG in Spoken and Written Dialogue, pp. 133–140 (2003)
11. Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., De Carolis, B.: GRETA. A Believable Embodied Conversational Agent, Multimodal Intelligent Information Presentation (2005)
12. van Welbergen, H., Reidsma, D., Ruttkay, Z.M., Zwiers, J.: Elckerlyc - A BML Realizer for continuous, multimodal interaction with a Virtual Human. Journal on Multimodal User Interfaces 3(4), 271–284 (2010)
13. Shapiro, A.: Building a Character Animation System. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 98–109. Springer, Heidelberg (2011)
14. Heloir, A., Kipp, M.: EMBR – A Realtime Animation Engine for Interactive Embodied Agents. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 393–404. Springer, Heidelberg (2009)
15. Taylor, P., Black, A., Caley, R.: The architecture of the Festival speech synthesis system. In: Third ESCA Workshop in Speech Synthesis, pp. 147–151 (1998)
16. http://www.mindmakers.org/projects/saiba/wiki
17. Heylen, D., Kopp, S., Marsella, S.C., Pelachaud, C., Vilhjálmsson, H.: The Next Step towards a Function Markup Language. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 270–280. Springer, Heidelberg (2008)
18. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
19. van Welbergen, H., Xu, Y., Thiebaux, M., Feng, W.-W., Fu, J., Reidsma, D., Shapiro, A.: Demonstrating and Testing the BML Compliance of BML Realizers. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 269–281. Springer, Heidelberg (2011)
20. Schröder, M.: The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. In: Advances in Human-Machine Interaction (2010)
21. Quigley, M., Conley, K., Gerkey, B.P., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: ROS: an open-source Robot Operating System, ICRA Open Source Software (2009)
22. Leuski, A., Traum, D.: NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine 32(2), 42–56 (2011)
23. Leuski, A., Traum, D.: A statistical approach for text processing in virtual humans. In: Proceedings of the 26th Army Science Conference, Orlando, Florida, USA (December 2008)
24. Lee, J., Marsella, S.: Nonverbal Behavior Generator for Embodied Conversational Agents. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006)
25. Feng, A.W., Xu, Y., Shapiro, A.: An Example-Based Motion Synthesis Technique for Locomotion and Object Manipulation. In: I3D 2012 Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, pp. 95–102 (2012)
26. Thiebaux, M., Lance, B., Marsella, S.: Real-time Expressive Gaze for Virtual Humans. In: AAMAS, vol. 1, pp. 321–328 (2008)
27. Kallmann, M., Marsella, S.: Hierarchical Motion Controllers for Real-time Autonomous Virtual Humans. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 253–265. Springer, Heidelberg (2005)
28. Feng, A., Huang, Y., Kallmann, M., Shapiro, A.: An Analysis of Motion Blending Techniques. In: Kallmann, M., Bekris, K. (eds.) MIG 2012. LNCS, vol. 7660, pp. 232–243. Springer, Heidelberg (2012)
29. Feng, A., Huang, Y., Xu, Y., Shapiro, A.: Automating the Transfer of a Generic Set of Behaviors onto a Virtual Character. In: Kallmann, M., Bekris, K. (eds.) MIG 2012. LNCS, vol. 7660, pp. 134–145. Springer, Heidelberg (2012)
30. Wagner, J., Lingenfelser, F., Bee, N., Andre, E.: Social signal interpretation (ssi). KI - Kuenstliche Intelligenz 25, 251–256 (2011)

31. Baltrusaitis, T., Robinson, P., Morency, L.-P.: 3D constrained local model for rigid and non-rigid facial tracking. In: IEEE Computer Vision and Pattern Recognition, pp. 2610–2617 (2012)
32. Morency, L.-P., Whitehill, J., Movellan, J.: Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In: Automatic Face and Gesture Recognition, pp. 1–8 (2008)
33. Suma, E., Lange, B., Rizzo, A., Krum, D., Bolas, M.: FAAST: The Flexible Action and Articulated Skeleton Toolkit. In: Proceedings of IEEE Virtual Reality, pp. 247–248 (2011)
34. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A(S.), Morency, L.-P.: Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 455–463. Springer, Heidelberg (2012)
35. http://unity3d.com/
36. http://www.ogre3d.org/
37. http://www.cereproc.com/products/sdk
38. http://msdn.microsoft.com/en-us/library/ee125663(v=vs.85).aspx
39. http://activemq.apache.org/
40. Leuski, A., Kennedy, B., Patel, R., Traum, D.R.: Asking questions to limited domain virtual characters: how good does speech recognition have to be? In: 25th Army Science Conference, Orlando, Fl (2006)
41. Artstein, R., Gandhe, S., Leuski, A., Traum, D.R.: Field Testing of an interactive question-answering character. In: ELRA, LREC (2008)
42. Swartout, W., et al.: Ada and Grace: Toward Realistic and Engaging Virtual Museum Guides. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 286–300. Springer, Heidelberg (2010)
43. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, L.-P.: Virtual Rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 14–27. Springer, Heidelberg (2006)
44. Rizzo, A., Forbell, E., Lange, B., Buckwalter, J.G., Williams, J., Sagae, K., Traum, D.: SimCoach: An Online Intelligent Virtual Agent System for Breaking Down Barriers to Care for Service Members and Veterans. In: Scurfield, R.M., Platoni, K.T. (eds.) Healing War Trauma: A Handbook of Creative Approaches, pp. 238–250. Routledge (2012)
45. Morbini, F., DeVault, D., Sagae, K., Gerten, J., Nazarian, A., Traum, D.: FLoReS: A Forward Looking, Reward Seeking, Dialogue Manager, Spoken Dialog Systems (2012)
46. Hartholt, A., Gratch, J., Weiss, L., The Gunslinger Team: At the Virtual Frontier: Introducing Gunslinger, a Multi-Character, Mixed-Reality, Story-Driven Experience. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 500–501. Springer, Heidelberg (2009)
47. Kenny, P., Parsons, T.D., Gratch, J., Rizzo, A.A.: Evaluation of Justina: A Virtual Patient with PTSD. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 394–408. Springer, Heidelberg (2008)
48. Traum, D.R., Marsella, S.C., Gratch, J., Lee, J., Hartholt, A.: Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 117–130. Springer, Heidelberg (2008)
49. Gratch, J., Hartholt, A., Dehghani, M., Marsella, S.: Virtual Humans: A New Toolkit for Cognitive Science Research. In: CogSci (2013)
50. Khooshabeh, P., McCall, C., Gandhe, S., Gratch, J., Blascovich, J.: Does it matter if a computer jokes. In: Extended Abstracts on Human Factors in Computer Systems, pp. 77–86 (2011)
51. Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., Scherer, S.: Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training. In: Aylett, R., Krenn, B., Pelachaud, C., Shimodaira, H. (eds.) IVA 2013. LNCS (LNAI), vol. 8108, pp. 116–128. Springer, Heidelberg (2013)
52. http://www.microsoft.com/en-us/kinectforwindows/

Real-Time Behavioral Animation of Humanoid Non-Player Characters with a Computational Ecosystem

Rui Filipe Antunes and Frederic Fol Leymarie

Goldsmiths, University of London, United Kingdom
{rui.antunes,ffl}@gold.ac.uk

Abstract. A novel approach to a decentralized autonomous model of agency for general purpose Non-Player Characters (NPCs) is presented: Computational Ecosystems as a model of AI. We describe the technology used to animate a population of gregarious humanoid characters in the virtual world *Where is Lourenco Marques?* an ethnographic artistic work characterized as a virtual world inhabited by a population of NPCs interacting autonomously among themselves as well as with an audience of outsiders (human observers). First, we present the background and motivations for the project. Then, we describe the technical details about the algorithm that was developed to generate the movements and behaviors of a population of NPC 'storytellers'. Finally, we layout some of the critical aspects of this particular implementation and contextualize the work with regards to a wider usage in virtual worlds.

Keywords: Multi-agent systems, Simulation, modelling and visualization, Animation, Computational Ecosystem, Virtual Worlds.

1 Introduction

The Ecosystem as an Allegory. To animate a population of Non-Playing Characters (NPCs) we have used a Computational Ecosystem (CE) to play the role of an Artificial Intelligence (AI). This is based on a community of autonomous agents which are organized as a simulation of a food-chain while trading units of energy. Each of the individuals emulates a rudimentary life cycle of generic carbon-based life forms. Mendelian genetics informs the evolution of the community, and genetic characteristics such as the speed or size of an agent are inherited by children from their parents when couples engender in a process that evokes sexual reproduction. Energy is required for the activities these individuals perform in a virtual city, including: moving, running, or simply breathing. The dynamics of energy transfer occur in *predatory* acts when the population competes for energy and space. In particular, when the energy level of an individual is too low (*i.e.* below a pre-specified threshold), this is regarded as death at which point the individual is taken away from the population.

The idea of using a CE to animate a virtual population of humanoids came about to symbolize the social situation experienced in a colonial city. In the tradition of literary allegory, the behavior of the community carries a second, or connotative level of

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 382–395, 2013.

narrative. In the virtual city/ecosystem social groups become analogous to *species*. The performance in the habitat dictates the behavior of these storytellers in the virtual world. Nevertheless, instead of the conventional primal events of fierce animals that attack and devour each other, what is shown to the audience/users are animations of humanoids interacting with each other and gesturing in apparent conversation. Each action in the ecosystem is rendered visible in the virtual world as the animation of a corresponding gesture or movement. For instance, if two hypothetical individuals interact, and one "attacks" the other, the expressiveness of the movements of the arms is greater in the winning individual than the movements that its victim shows.

The Computational Ecosystem as an AI Engine. The use of the technology of the CEs as generative engines appears in contexts as diverse as audio-visual installations [1], music [2] or even for the choreography of characters in virtual worlds [3]. The innovation we introduce is the use of the CE as an AI to coordinate the social movements and behaviors of a community of humanoid NPCs, taking advantage of the complexity potential inherent in such a dynamic ecosystem.

By design, the behaviors observed in the community of individuals in a virtual city are translations in the form of visual representations of the inner-workings of the ecosystem. The spatial layout of the virtual city provides a support for the visualization of the underlying multi-agent state space, with a direct connection between the current state of the ecosystem and the configuration of the characters on the layout. Each NPC is an individual belonging to the population and we refer to each one as a 'character'. Each birth is represented by the new character appearing next to its parents, while its 'dematerialization' in thin air represents its 'death'. The interactions during the lifetime of the characters are translated into a series of movements and gesticulations while being constrained to the surface of the world. The CE establishes correspondences between states, movements and actions performed by each of the characters. For instance, the action of an hypothetical individual feeding in the virtual ecosystem might correspond to a certain gesticulation being performed by the character in the virtual world, while its escape after a fight with another creature will correspond to a different gesture being expressed by the character. A set of *nine base animations* were defined for this work. This small set was deemed sufficient for a prototype to allow to explore and demonstrate the potential of CEs in animating virtual populations of humanoid characters. This includes animations for walking (1) and of gestures (8) for the arms to be used during interactions. The implemented mechanism permitting the generation of each character behaviours is described later in § 2.

1.1 Background

The animation of populations in digital reconstructions of historical sites such as Petra, in Jordan, the former Pennsylvania train station in New York city [4], the ancient city of Pompeii [5], the Babylonian Uruk [6], or for theme parks [7], are just a few examples of a flourishing area of research that looks at modeling virtual spaces inhabited by communities of autonomous humanoid characters.

A few standards have been established with regards to modelling groups of NPCs. Three approaches prevail, when individuals: (a) are represented as particles subject to

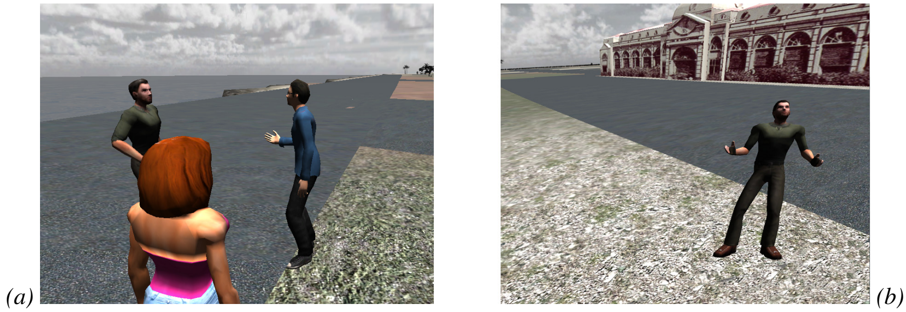


Fig. 1. (a) Storytellers in autonomous interaction. (b) Storyteller interacting with the audience.

physical forces [8]; (b) are represented as states of cells in cellular automata [9]; or (c) follow a rule-based scheme [10]. One of the merits of these approaches is in modeling in a realistic way at a macro-level the spatial-flow features of a crowd. More recently the individual's behavior within a multitude, at a micro-level, has received the attention of researchers. Some works attempt to recreate the spontaneity of behaviors visible in small groups of a few dozens of individuals. Shao and Terzopoulos define pedestrian behaviors in both (i) the former NYC Penn station and (ii) the theater of Petra's Great Temple, producing NPCs exhibiting heterogeneous and apparently spontaneous behaviors such as standing in queues or watching shop windows [4]. Pelechano *et al.* model a group of humanoids interacting socially at a cocktail party [11]. A good survey of the field of crowd animation up to 2008 can be found in the books by Pelechano *et al.* [12] and by Thalmann and Raupp Musse [13].

We describe later the generative system developed to produce the visual effects of a community of gregarious individuals in small interacting crowds (groups of 2 to 10 individuals, Figure 1). This work was motivated by an artistic piece which required, by design specifications, that in order to address *European colonialism* the animations of a humanoid population inhabiting a virtual city had to be driven by *predatory* behaviors as a metaphor/allegory for the social conditions lived in a colonial city. We are looking, in particular, at systems where agents are organized in the hierarchical structure of a *food-chain* while trading units of energy and biomass as a way of promoting community dynamics. A CE proves to be able to provide for complex environment simulations rich in the heterogeneity and spontaneity which we are targeting.

1.2 The Artwork *Where is Lourenço Marques? (WisLM)*

This paper describes the technical implementation of the CE that animates a virtual population in the artwork *Where is Lourenço Marques?* characterized as a representation of the former city of Lourenço Marques, the capital of the province of Mozambique, during the period of Portuguese colonial domination. The city later became known as Maputo after the independence in 1975. During the process of decolonization from 1974 to 1976, many of its citizens were forced to abandon the city for social and political reasons. This artwork presents the mediation of the memories of some of those who experienced the last period of its colonial times. The memorial takes expression in a 3D virtual world; an illustrative video can be found on YouTube (search: "Where is Lourenco Marques" by "xtnz").

From Interviews to Character 'storytellers'. The initial steps of the project entailed a process of interviews with a community who left the city in those dramatic days, and now mostly resides in Portugal. An artistic and subjective reconstruction of the city was built in a 3D virtual environment (using Unity3D) based on their accounts and their shared material forms of memorabilia. The focus of this paper is the community of humanoid NPCs which roams autonomously the city and interact with each others as well as with the human audience.

These characters become 'storytellers' (Figure 1) when the user selects them (*e.g.* by pointing or clicking on any of them). A selected NPC interrupts its current activity and looks at the camera. Then, while gesticulating expressively, the character provides an audio testimony by streaming one of the oral accounts recorded during the initial process of interviews. Thus, this population of animated NPCs can assist in the task of bringing the past experiences of some of the "expatriated" citizens. Each of the individuals in this population is the bearer of an excerpt from an interview, functioning as the carrier and mediator of real-life human stories. The audience is thus implicitly invited to seek-out the storytellers through the city and listen to their stories.

2 Technical Description

In order to elaborate our CE, we have selected some techniques with a proven record of successful results in animating populations of multiple individuals. These are: (i) a *hormonal system* as proposed by Daniel Jones in his framework for bio-inspired swarms [14], (ii) a *metabolic system* to define and restrict the diet of our characters based on the model of Saruwatari *et al.* [15], and (iii) a *classifier system* adapted from John Holland's model [16] which drives the action–selection mechanism of our characters. We now describe our implementation of each of these main techniques.

2.1 The Hormonal System

Jones' framework for sound-based performance using swarm dynamics, introduces a biomimetic design process that augments the classical rule-set for flocking, in part through the implementation of a hormone-like system which allows temporal modifications of the individual behavior of particles within a swarm [14]. We have adapted this architectural design to expand the classical energy paradigm used in more traditional CEs which typically operates as a thermostat-like regulator [17]. Adding an extra layer of hormonal regulators increases both the life-likeness of our model and its associated complexity. This layer is defined via five variables:

1. Testosterone: increases with age and crowdness, decreases upon giving birth, and also causes an increase in the likelihood of reproduction.
2. Adrenaline: increases with overcrowding, decreases as a result of internal regulation overtime, and also causes a greater rate and variance of movement.
3. Serotonin: increases with 'day' cycles, decreases during 'night' cycles and as result of hunger, and also causes a greater social attraction towards other characters.

4. Melatonin: increases during ‘night’ cycles, decreases during ‘day’ cycles and also decreases the rate of movement.
5. Leptin: increases upon eating, decreases steadily at all other times, also causes downward regulation of serotonin when depleted, and finally causes greater attraction to food.

The Blueprint Descriptor. The hormone-like system described above is initially configured by the genetic descriptor which contains a blueprint for the attributes of the character. This is a string with 15 binary digits, where different sections of the string code for a set of six specific features:

1. Age (**gAge**) — defines the rate at which the agent ages.
2. Introspection (**gIntr**) — establishes the level of gregariousness.
3. Hormone cycles (**gCycl**) — the strength or speed of the hormone cycle (*e.g.* how much will it increase per day).
4. Hormone uptakes (**gUptk**) — indicates the intake during an hormone cycle (*e.g.* how much will the hormone increase when a character meets another one).
5. Body chemistry (**gChem**) — defines the chemical components present in the body.
6. Metabolism (**gMetab**) — determines what chemicals can be ‘digested’.

In the *initial* population, these features are determined randomly. Afterwards, each subsequent generation inherits this information from their parents. This follows a Mendelian-like process of diploid reproduction, where crossover operators are used on Gtypes to promote variation. In the process some noise is added to the inherited information to mimic mutations (with an arbitrary small probability $Pm = 0.05$ of effective mutation on each bit).

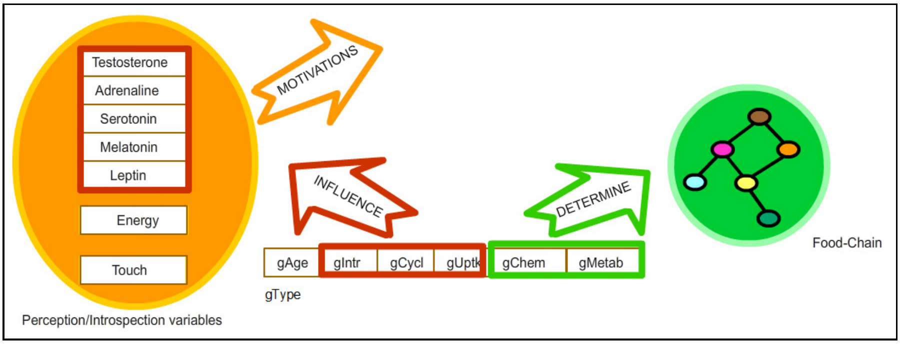


Fig. 2. Graph depicting the GType influence on the behavior of the character. The variables (gIntr, gCycl, gUptk) directly influence the hormonal system, which motivates the actions performed in the world; the variables (gChem, gMetab) set the body-composition and dietary specifications, which determines the environmental context for the actions.

2.2 The Metabolic System

Saruwatari *et al.* [15] provide a model we found useful to determine the dietary specifications. Their framework uses two strings, the first of which defines the body constitution of the character, while the second is used to describe its metabolism. Potential preys are those whose constitution-string matches the predator's metabolic-string. Saruwatari *et al.* have shown this simple mechanism potentially leads to the emergence of complex multi-trophic food-chains with variable depths, which in turn gives us the necessary differentiation and stratification required in our model.

The string of 3 binary digits present in the Gtype section **gChem** defines the first part of the dietary specification representing the character's body 'chemical' composition. These digits code for chemicals A, B and C. Each character is equipped with chemical repositories that are direct translations from the composition-Gtype **gChem**. Take for instance a character with **gChem** of "010". The only chemical present in the repository will be B. On the contrary, another character with Gtype "101" will have both chemicals A and C active. When an hypothetical character (X) preys another character (Y), X will fill its own repositories by extraction from the chemical repositories of Y. In this transfer process 90% of the value is wasted. Each repository only carries a maximum capacity directly related to the character's body size which is given by a direct translation of the binary value of **gChem**. These chemical attributes play a fundamental role in the ecosystem, since they determine part of the dietary specification in the community and thus the interactions between individuals.

The second part of the dietary specification is provided by **gMetab**, the component that defines the character's metabolism, *i.e.* what chemicals can be 'digested'. An example is an hypothetical character with **gMetab** 010, which will be able to prey individuals whose **gChem** codes the B component: 010, 110, 011, 111. Consequently the combination **gChem–gMetab** structures the essential interactions in the predator-prey relationships. This predator-prey mechanism of matching the metabolic-string with the composition-string provides an interaction space of size 8 x 8, which was wide enough for the current work (in terms of observed behaviors).

The Metabolic Rules. The metabolic system simulates a simplified food-energy conversion. Besides the described hormonal system and chemical repositories, another main structuring variable contributes to the behavior of the characters: energy. This is generated from the chemical repository of the character. To emulate a conversion from mass to energy, we have defined an arbitrary chemical reaction which requires three chemical units to produce one unit of energy (*e.g.* $1\,B + 2\,C \Longrightarrow 1\,energy\,unit$). Energy can be spent when breathing or in activities performed in the world such as moving, running away, attacking, playing, mating or eating. Below a pre-set level more energy needs to be produced from the chemical–repositories of the character. Below a pre-set threshold the character starts to feel tired, activating an internal sensor (to search for food). If the energy level reaches the value 0 the character dies and is removed.

2.3 Behavior of the Characters via a Classifier

The characters' behavior is defined by three main operational stages: perception, decision, and action. Each character is provided with: (i) internal sensors monitoring energy and hormonal levels, and (ii) sensors for contact which are triggered by the proximity of other characters (Figures 2 and 3). As a function of these inputs, the character will choose the action to take using a classifier system inspired by the description provided by John Holland [16], a model that allows autonomous agents exhibiting self-organization capabilities and temporal adaptation. Holland's model was used in the well-known Echo system [18] and also inspired artworks such as Eden [17].

The Classifier System. During the process of perception the system inspects the level of energy and the state of the hormonal variables, as well as if the body of a character is entering in contact with any other characters. When any of these variables is activated, such as when: (i) the energy or leptin levels are below pre-selected thresholds, (ii) the testosterone, adrenalin or serotonin levels are above some pre-fixed thresholds, or (iii) the character is touching another body, then an *action–message* is generated. This message takes the form of a string of length 6, and is composed from the grammar set $\{0, 1, \#\}$, identifying which sensor is active — binary values indicate the active and inactive states while # functions as a wildcard.

The active messages list. This is a list with messages queuing to be processed. If a message is new on the list, it will be inserted with an assigned priority of 0. If, on the contrary, the message already exists, meaning that the same sensor has already triggered one or more messages, then the priority of this existing message is incremented. During the decision stage, the message with highest priority on the list will be removed and processed against a table of rules to generate actions.

The table of rules. It describes a set of actions and their indices. The rules are initially similar for all characters. Each rule contains three parameters: index, action and priority. The index is again composed from the grammar $\{0, 1, \#\}$ and is used to match rules to a corresponding selected message being processed. Multiple rules can match one single message. The character #, which functions as a wildcard, implies that any value can be accepted for the corresponding particular character of the index. Furthermore, each of the rules has an assigned priority (initialized with a random value), and thus from all the candidate rules (with indices matching a selected message) only the one with highest priority is selected. The action to perform is coded on the second section of the rule, an alphanumeric code to be translated into a procedural action, such as an instruction to prey on any character that is within a certain distance, or an instruction to move towards the closest character, and so on (Figure 3).

The reward system. The priority of the rules is updated according to the consequences of the actions performed in the world. This translates into the character recognition of which actions are advantageous. An *ad hoc* reward was attributed to some of the possible actions such as eating when hungry, or victory or defeat on battle. If for instance the selected action is to feed, this implies a positive reward. On the contrary, being hit implies a negative value. The reward can be associated not only with the current rule which has triggered an event, but also preceding rules leading to the current action. Each character has a FIFO (First In First Out) memory stack which stores the last five rules performed. This block of memory is also rewarded accordingly, with the rules being

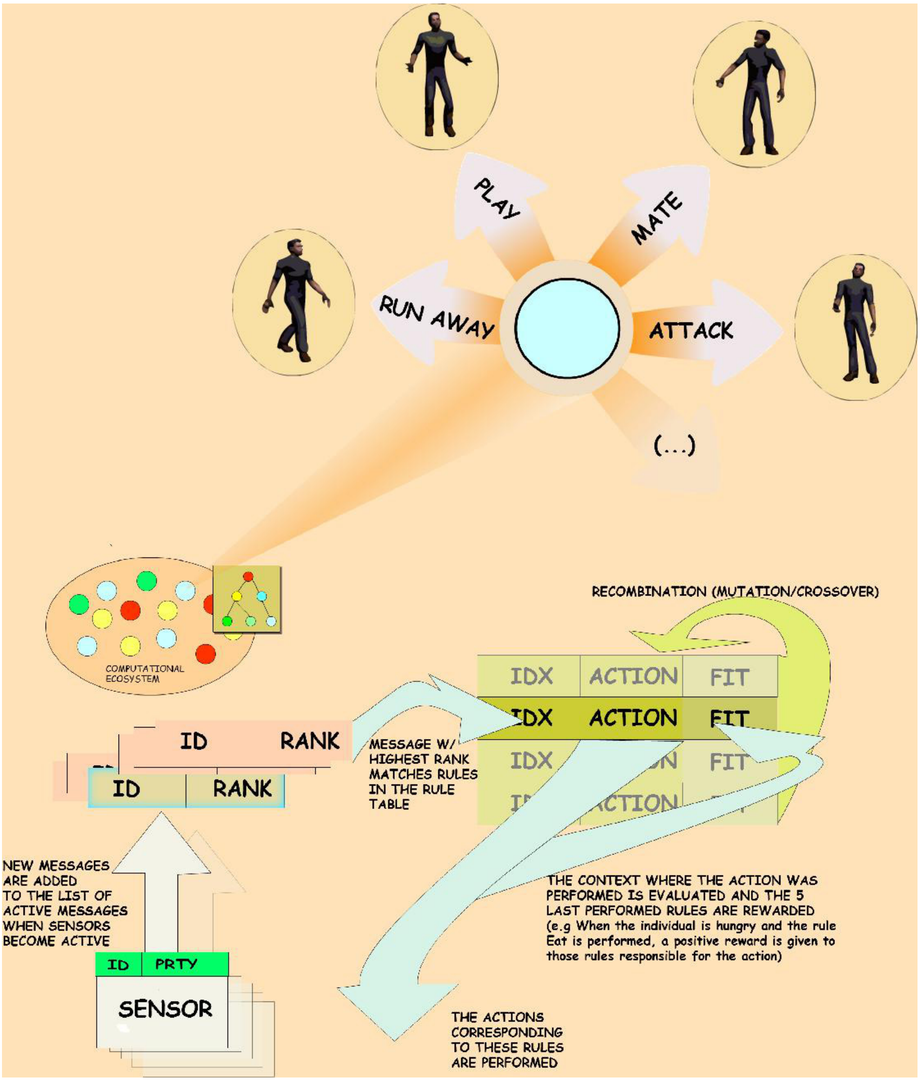


Fig. 3. Graphical depiction of the overall structure of the algorithm. On the bottom the action selection mechanism is illustrated: sensors trigger messages indicating a specific need; these messages are prioritized and form a buffer, which is ordered by associated priority. The message with the highest priority is selected to trigger an action. On the top of the graph the mechanism rendering actions is illustrated. For each syntactic action, an associated animation is played.

credited in a decremental way corresponding to the time they have been present on the memory. For instance, when a character manages to prey, the rule for the action which has triggered the preying event is rewarded with 5. The immediate rule–action prior to that one is rewarded with 4; the anterior with 3, and so on. When a new event occurs and a new action is performed, the oldest rule-action is removed from the stack.

Generation of new rules. As an outcome of reproduction, a newborn inherits from the current rule-table of the parents. To constitute the new rule-table, the rules with top priority from both progenitors are inherited in a 50/50 proportion. Each of the indices of the new set of rules then is perturbed by a process of (possible) mutations. These indices may suffer a transformation resulting from four possible attempts for digit mutation, each with a success probability of 50%. We set the mutation rate at such a high probability to ensure rapid variability in behaviors.

The Mapping of Behaviors. As mentioned earlier, actions are encoded in the second part of each rule which might trigger new messages or generate some physical action such as move or talk. To render visible each physical action in the virtual world one associated animation needs to be played. The relationship between the animations and the actions is rigidly defined *a priori*. For instance, for the rule triggering the action 'eat', the animation associated with 'eating' is performed (Figure 4 a)). However, these animations are not literal visualizations of these internal states, but rather were selected as interesting and playful in illustrating moments of conversation via gesticulating using well defined separate sets of body movements. The continuous dynamics of the virtual world is generated by the on-going displacements of the characters. These movements can be of three types: (i) in the direction of an arbitrary 'preferred location', *i.e.* a randomly selected layout point set of coordinates which, once reached, is re-initialised with a new target position; (ii) in the direction of other characters as a consequence of the internal needs as determined by the classifier system under the influence of the hormonal and metabolic sub–systems; and (iii) moving towards a member of the public in response to being selected for interaction (*i.e.* telling a story).

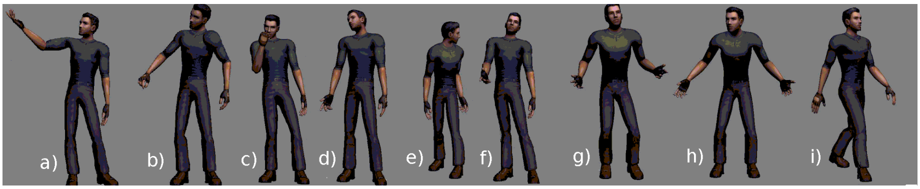


Fig. 4. Still images of the animations corresponding to each of the actions performed by the agent. From left to right: a) Eat a prey; b) attempt to mate with a partner that is not ready; c) reproduction or play alone; d) successful mating; e) attempt to mate but no compatible individuals exist on the vicinity; f) loosing an attack; g) play with happy mate; h) victorious attack; i) walking (move to mate, wander, move to prey).

3 Discussion

The *internal dynamics* of a functioning ecosystem is proposed as a way to structure and coordinate the animation of a population of humanoid characters. This *visualization* is exemplified with a population of NPCs simulating conversational behavior in the virtual world WisLM. The traditional approach of evolutionary art is to directly visualize the information defined on the Gtypes. In our work the information carried on the Gtype

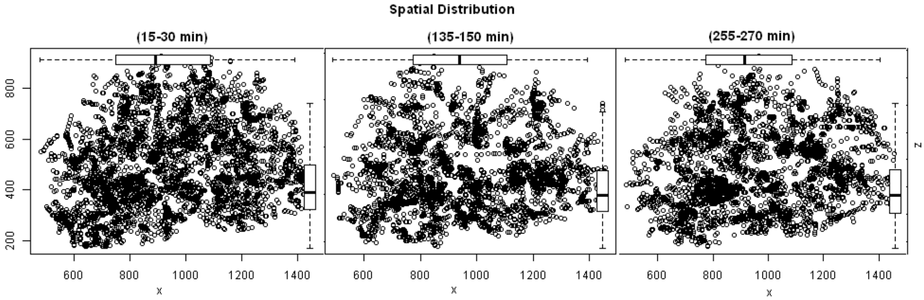


Fig. 5. Spatial distribution in time. At intervals of 15 seconds a snapshot of the system was captured, during a period of 6 hours (one typical day of exhibition). Each dot represents the location at that particular moment for one individual. The variables x and z stand for the coordinates in the Cartesian horizontal space. Each graph portrays a frame of juxtaposed 15 minutes of execution.

of the individuals describes features such as their body-composition and dietary constraints. However, in contrast to other approaches, what is emphasized and visualized in our system are the ephemeral states of the individuals within an ecosystem: *i.e.* the continuously changing actions they perform during their life-time reflecting the inner states of the CE, such as exchanges of energy. Similarly to the Gtype–Ptype paradigm from evolutionary computation, during the process of translation the linearity and the distance between the syntactic and the semantic levels can vary. For each action there is a corresponding symbol in the internal system and there is a posterior interpretation and visualization of that particular symbol. During this process of interpretation, converting symbols into visualizations, there is scope for creativity. The present research explores the ecosystem paradigm and the generative features implicit in CEs as a model of AI to develop new approaches for the animation of NPCs.

The result of this exploration is a populated landscape where individuals roam through the city, in a mapping of movements which is not random, with distinct emerging spatial attractor loci (Figure 5). Moreover there is place for heterogeneity and spontaneity and while ignoring some of the members of the crowd these autonomous individuals get together and aggregate in small groups in apparent expressive gesticulating dialogues (Figure 6). This system was exhibited to a public audience at the closing ceremony of the Watermans Festival of Digital Art 2012, and at the Tin Shed space, both in London, UK (Figure 7), with positive feedback from the audience. We obtained twenty four responses from anonymous users amongst the audience attending the shows. From an analysis of the responses to a questionnaire, 80% said the group formation and the talking movements in the simulation appeared to obey to an internal and coherent logic, whereas 65% said the behaviors to be believable in a cartoonish way.

To build our model of AI we have put together a set of established and efficient techniques for the agency of populations: Jones' hormone-framework, Saruwatu's model with dietary specifications and Holland's classifier system. Eventually, the complexity of the system could be reduced, *e.g.* events could be regulated by energy levels only. Our current design of a CE generates more noise in the system and consequent variability, than simpler implementations. Also, it provides a CE platform which is closer to a biology-based model, which we propose is of greater interest for an artwork such as WisLM which has its sources in human socio-history (of recent European colonialism). Nevertheless, our approach to a CE implementation is currently not computationally tractable for large crowds, when relying on the limited computer power of a single PC (our case and a typical situation when performing exhibits at small venues) — our model implemented a population restricted to a maximum of 200 simultaneous characters while still achieving real-time responses.

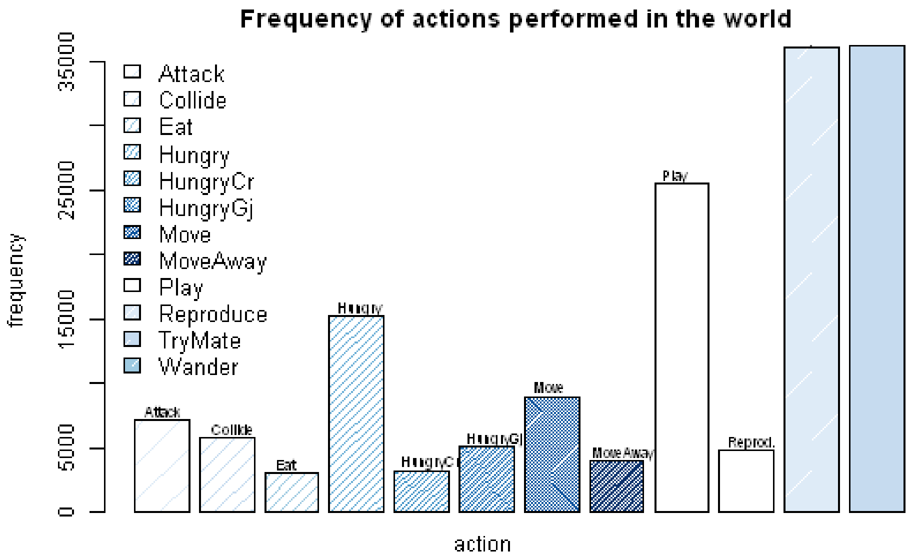


Fig. 6. The graph illustrates the number of times each action types was performed. At intervals of 15 seconds the system recorded what an individual was doing. From left to right, the actions are: Attack, Collide, Eat, Hungry, HungryCr (search for Carcasses), HungryGj (search for other creatures), Move, MoveAway, Play, Reproduce, TryMate, Wander.

Another aspect to take notice of, is that there is an implied semantic resulting from interactions of physical gestures which in our case was ignored. The resulting conversation is not immediately intelligible from the point of view of the actions in the ecosystem. Also, the small-number of pre-selected movements, as well as the lack of sophisticated blending which is limiting, hide the richness that could be effectively reached. An increased set of animations would make the system more flexible. To explore this potential further, it would be interesting to consider gestures and poses which better reflect the richness and combinatorics of the internal states of the characters. This could be enriched with a wider set of animations, which might reflect such nuances. Moreover, in contrast with our current deterministic approach of defining a limited set of animations, it would be interesting to break the animations into smaller bits. To explore the CE further, in terms of emergence and novelty this work would become richer with the incorporation of elements of a language of movements on which the CE could

(a) *(b)*

Fig. 7. *(a)* A perspective on the installation of WisLM during the exhibition at the Tin Shed, in London, in 2012. *(b)* A member of the audience interacts with one of the characters.

act. This would create the conditions for procedural movements with potential emergence of unexpected movements. Another interesting possibility would be of exploring the sonification of the CE internal states.

One advantage of the ecosystem approach to the animation of NPCs, over other established methods of population simulation, is its generative capacity. The system allows for a flexible representation of complex behavioral scenarios, and is relatively easily adaptable to different environmental contexts (*e.g.* buildings of various layouts and complexity levels) and varying numbers of individuals. The term "ecosystem" itself, and the associated terminology, are operative metaphors. The nature of the system is quite malleable. For example, instead of instances such as 'energy' the model admits variants such as 'currency', and then, events such as 'attack' or 'eat' could be mapped to become 'negotiate' or 'acquire'. Additionally, CEs, drawing on the ecosystem paradigm, provide natural fluctuations in the population density which might prove interesting from the point of view of the realism of the simulation. The drawback of the framework is however the difficulty in precisely controlling the behaviors observed, as the CE is, by definition, a complex system with a few variables upstream influencing potentially very complex detailed behaviors downstream.

4 Conclusions

Traditional genetic algorithms and evolutionary art are characterized by a process of interpretation of symbols between the Gtype and Ptype (*i.e.* genotype and phenotype). The approach we describe visualizes instead the ephemeral states of the individuals within the dynamics of an ecosystem implementation. The behaviors of the individuals during their normal activity, their primal movements and actions such as attack, flee, mate, prey, are taken as syntactic elements during a process of re-interpretation. For instance, the action of 'eating' produced at the inner ecosystem level might be translated, at the semantic level, as the animation of a wild animal chewing a carcass in the virtual world. However, as it happens with the original Gtype–Ptype paradigm, this process is also open to creativity and the linearity and distance of translation are subject to interpretation. For example, the action 'attack' rather than the animation of an animal fighting might instead correspond to a specific movement being choreographed to be performed by a dancer or, as determined in WisLM, might correspond to a conversational movement.

Our work explores these ideas with the animation of a population of humanoids NPCs in a virtual world playing the role of storytellers. However, the same specifications did also require the behavior of this community of humanoids to be the result of the dynamics of an ecosystem where individuals would seek out each other with predatory intentions. A CE, a system of agents organized in a hierarchical structure (of a food-chain) and trading units (of energy and biomass) generates such dynamics. The potential complexity of crowd interaction is revealed by the spontaneity of group formation of characters engaging in heterogeneous gesticulations during their conversations. The resulting populations offer spatial and behavioral distributions which are realistically far from uniform (Figures 5 and 6).

We have modeled a population of gregarious NPCs showing some of these spontaneous and conversational behaviors. Our approach took advantage of one of the fundamental properties of CEs: by relying on the variety and spontaneity of the elementary behaviors, the autonomy and self-organization of the agents generates ever-changing heterogeneous patterns at the global scale of a community. In other words, to build up this work we drew on the fact that the CE is, in essence, a dynamic generative framework which we have shown can be applied to animate NPCs.

Acknowledgments. This research is supported in part in the form of a PhD studentship to Mr. Antunes by the *Fundação para a Ciência e Tecnologia* from Portugal, contract reference SFRH / BD / 61293 / 2009. We also thank the telecom company Telefonica and in particular the VIDA competition for financial assistance under the award *Incentives for Ibero-American Production*. Finally, we thank the anonymous reviewers for their helpful comments and constructive criticism. Images 7a and 7b appear courtesy of Tin Shed Gallery.

References

1. Dorin, A.: Pandemic — Generative Software Installation, Exhibited: Bouillants 4, Vern-sur-Seiche, Brittany, France, Allin & Dupuis (artistic directors) (2012)
2. Eldridge, A., Dorin, A.: Filterscape: Energy Recycling in a Creative Ecosystem. In: Giacobini, M., et al. (eds.) EvoWorkshops 2009. LNCS, vol. 5484, pp. 508–517. Springer, Heidelberg (2009)
3. Antunes, R.F., Leymarie, F.F.: Generative Choreography: Animating in Real-Time Dancing Avatars. In: Machado, P., Romero, J., Carballal, A. (eds.) EvoMUSART 2012. LNCS, vol. 7247, pp. 1–10. Springer, Heidelberg (2012)
4. Shao, W., Terzopoulos, D.: Populating Reconstructed Archeological Sites with Autonomous Virtual Humans. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 420–433. Springer, Heidelberg (2006)
5. Maïm, J., Haegler, S., Yersin, B., Müller, P., Thalmann, D., Van Gool, L.J.: Populating Ancient Pompeii with Crowds of Virtual Romans. In: VAST, pp. 109–116 (2007)
6. Bogdanovych, A., Ijaz, K., Simoff, S.: The City of Uruk: Teaching Ancient History in a Virtual World. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 28–35. Springer, Heidelberg (2012)

7. Huerre, S.: Agent–Based Crowd Simulation Tool for Theme Park Environments. In: 23rd Inter. Conf. on Comput. Anim. & Social Agents (CASA), Bournemouth University (2010)
8. Helbing, D.: A Fluid Dynamic Model for the Movement of Pedestrians. Complex Systems 6, 391–415 (1992)
9. Banerjee, B., et al.: Advancing the layered approach to agent–based crowd simulation. In: 22nd Workshop on Principles of Adv. & Distrib. Simulation, PADS, pp. 185–192 (2008)
10. Reynolds, C.W.: Flocks, Herds and Schools: A Distributed Behavioral Model. ACM SIGGRAPH, Computer Graphics 21(4), 25–34 (1987)
11. Pelechano, N.: et al.: Being a part of the crowd: Towards validating VR crowds using presence. In: Autonomous Agents & Multiagent Systems (AAMAS), pp. 136–142 (2008)
12. Pelechano, N., et al.: Virtual Crowds: Methods, Simulation, and Control. Synthesis Lectures on Computer Graphics and Animation. Morgan and Claypool (2008)
13. Thalmann, D., Musse, S.R.: Crowd Simulation. Springer (2007)
14. Jones, D.: AtomSwarm: A Framework for Swarm Improvisation. In: Giacobini, M., et al. (eds.) EvoWorkshops 2008. LNCS, vol. 4974, pp. 423–432. Springer, Heidelberg (2008)
15. Saruwatari, T., Toqunaga, Y., Hoshino, T.: ADIVERSITY: Stepping up trophic levels. In: 4th International Workshop on the Synthesis and Simulation of Living Systems, pp. 424–429 (1994)
16. Holland, J.: Hidden Order: How Adaptation Builds Complexity. Helix Books (1996)
17. McCormack, J.: Eden: An Evolutionary Sonic Ecosystem. In: Kelemen, J., Sosík, P. (eds.) ECAL 2001. LNCS (LNAI), vol. 2159, pp. 133–142. Springer, Heidelberg (2001)
18. Forrest, S., Jones, T.: Modeling Complex Adaptive Systems with Echo. In: Stonier, R., Yu, X. (eds.) Complex Systems: Mechanisms of Adaptation, pp. 3–21. IOS Press, Amsterdam (1994)

Judging IVA Personality Using an Open-Ended Question

Kris Liu[1], Jackson Tolins[1], Jean E. Fox Tree[1],
Marilyn Walker[1], and Michael Neff[2]

[1] University of California, Santa Cruz
{kyliu,jtolins,foxtree,maw}@ucsc.edu, maw@soe.ucsc.edu
[2] University of California, Davis
mpneff@ucdavis.edu

Abstract. Judgments of personality typically employ ratings of Big Five scale items such as "How emotionally stable is this person?" with choices from 1 (least) to 5 (most). Such questions focus raters' attention on an experimenter's dimensions of interest. We show that the personality traits provided as a result of open-ended questions such as "What personality does this animated character convey to you?" can differ from those observed when raters are given scale questions. Using IVAs that gesture in ways associated with emotionally stable and unstable people, we showed that participants were more likely to describe the unstable agent as disagreeable and the stable agent as extraverted; emotional stability was not usually mentioned. However, a Big Five inventory showed that these agents differed on agreeableness and emotional stability. An open-ended question method of assessing what personality an IVA conveys can potentially be more informative than using scale-item inventories alone.

Keywords: personality, Big Five, methodology, gesture, non-verbal behavior.

1 Introduction

Measuring an individual's personality typically relies on the descriptive adjectives used to develop the Big Five or OCEAN model [13, 16, 19], which claims that personality varies on five factors: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (which we refer to as emotional stability, such that high neuroticism is low emotional stability and vice versa). The Big Five model of personality has become a standard in the field of psychology over the last 50 years. A large body of research has focused on the creation of personality questionnaires, inventories, and adjective rating scales designed to measure these broad dimensions. The most well used personality inventory to develop from this methodology is the Revised NEO Personality Inventory (NEO-PI-R), developed by Costa and McCrae [4]. But there have been other valid and successful measures such as the Big Five Inventory (BFI), developed by John, Donahue,

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 396–405, 2013.

and Kentle [8], and Goldberg's trait descriptive adjective [7] lists, which has high internal consistency and has been easily replicated [10].

Two related problems arise when using any of a number of Big Five inventories for the purpose of asking an observer to assess the behavior of another. First, the characterizations of personality that are spontaneously generated by people may be different from measurements gleaned from personality inventories, as people may be primed or influenced by the wording of scale rating items such as "This person is emotionally stable." Second, behaviors not associated with anything having to do with personality may appear diagnostic when considered in light of a personality trait. For example, touching one's hair is not always a sign of anxiety. But if an observer is asked whether one person who is touching his hair is more anxious than another person who is not, that behavior may now seem indicative of anxiety even if the observer did not previously mention hair-touching as something that anxious people do.

A linguistic analogue can be found in *um*, which is often believed to indicate anxiety [2, 3]. Listeners judge speakers who use *ums* with a particular topic as less comfortable than those who do not [5]. Speakers are also seen as more dishonest and more likely to be experiencing speech production problems. The fact that listeners are willing to judge those utterances as indicating more anxiety, dishonesty and production trouble than non-*um* utterances suggests not only that there are multiple impressions an *um* can convey, but that listeners are willing to make various judgments depending on the question asked.

A similar phenomenon has also been observed with judgments of sarcasm [1]. Utterances dripping with sarcasm were filtered to mask lexical content while retaining prosodic information. Listeners rated the filtered sarcastic utterances to be more sarcastic than non-sarcastic utterances, as well as angrier and more inquisitive. What is not answered by these measures is what listeners' judgments are when their attention is not focused on sarcasm, anger, or inquisitiveness. If listeners were simply asked what the response conveyed, would they infer different things? Do experimenters decrease the informativeness of their participants' responses with their questions?

The methodology in the present study attempts to circumvent the issue of an experimenter's unintentionally influencing and over-simplifying potentially sensitive judgments of behavior, specifically of personality gleaned from gesture. We tested what personality traits were conveyed by an animated agent by using both open-ended and closed, Likert scale questions. One way to characterize the two methods is that one assesses first impressions and the other assesses potentially primed impressions. Ideally, data obtained through open- and closed-ended questions should be complementary with open-ended question data enhancing the basic data provided by the scale inventory.

1.1 Seeing Personality

Visibility of a trait plays a role in how well a given trait is judged [6, 9]. For example, for extraversion, visibility is high and judgments are more accurate than for other personality traits [6,9]. Extraverts also tend to consistently display broader gestures, made further away from their bodies [12,18]. Their gestures are also more frequent and animated than the gestures of introverts [11]. Findings such as these have been applied to agents designed to portray extraversion [15].

In contrast, emotional stability (i.e., neuroticism) and agreeableness are less visible and are judged less accurately [9]. Nevertheless, there is some evidence that different non-verbal behavioral cues can predict personality. The Five Factor Nonverbal Personality Questionnaire (FF-NPQ) asks the examinee how likely they are to engage in a particular behavior, for example, riding a bucking horse [17]. The FF-NPQ and the NEO-PI-R have been found to reliably predict the Big Five factors to a similar extent, indicating nonverbal behavioral cues can also be used as predictive of personality. That is, the validity of both verbal measures (i.e. adjective checklists) and nonverbal measures (e.g., FF-NPQ) suggests one can also determine personality directly from nonverbal behavior practices.

1.2 Expressing Emotional Stability

Previous researchers summarized the findings on emotional stability most applicable to an animated agent and were able to show reliable variation of the trait through variation in language and through the use or absence of scratches and other self-touches (*self-adaptors*) [14]. The current work examines whether a user's impression an agent's emotional stability can be determinted solely based on its gestures and whether this impression changes depending on how the user is asked to evaluate the agent.

Fig. 1. The Shaky agent (proposed low emotional stability), shown on the left, was designed to display a stiff posture with rigidly held arms, raised shoulders and a narrow stance, whereas the Smooth agent (proposed high emotional stability) on the right was designed to appear more relaxed

Based on the findings reported in Neff et al. [14], two new animation clips were generated, designed to portray high and low emotional stability. Both clips consisted of the same gestures with the same timing. Variation was only allowed in the quality of the movement. In brief, the high emotional stability variant

Table 1. Motion edits applied to create the Shaky Agent and Smooth Agent clips

	Shaky Agent (Proposed Low Emotional Stability)	Smooth Agent (Proposed High Emotional Stability)
Stance	Narrower, asymmetric stance. Legs swiveled in.	Reduced knee bending.
Posture Timing	Time warped body turns to create more rapid jerks.	Used more smooth base motion data.
Collarbones	Collarbones brought up and more back.	Collarbones brought down and slightly back.
Arm Swivel	Elbow rotated 20°inward.	
Gesture Strokes	Gestures 20% smaller and brought more in front of body. Gestures cross in front of body. Jerks were added to the motion in the direction of the motion path.	Gestures more outward. No jerk was added to the motion, so the trajectory was smooth.
Gesture Retractions	Retract position is held out from the body, low and to the side of the character.	Hands are raised to allow arm bend and brought more in front of the body for a more relaxed appearance. Physical simulation is used on the retraction to add a more relaxed swing.
Gesture Phase Connections	Sharper, achieved by using lower weight tangents on motion curves.	More rounded.

was designed to appear more relaxed and comfortable, with gestures whose trajectories were smooth (Smooth Agent); in contrast, the low emotional stability variant had jerky gestures (Shaky Agent). Motion capture data was used for both clips, with edits applied on top to vary qualitative aspects of the motion. The changes used are summarized in Table 1 and frames from the two clips illustrating the base pose difference are shown in Figure 1.

2 Study

This study examined three questions. First, could gesture alone change an observers attribution of personality to a virtual agent? Second, was an observer's unprimed attribution of an agents personality congruent with its proposed, programmed personality? Finally, were basic assessments of agent personality using human-normed Big Five inventories in agreement with the relatively unbiased impressions of personality? For the purposes of this study, we tested one personality dimension: emotional stability, as people often have a relatively strong folk notions on how they neurotic vs. non-neurotic individuals gesture.

2.1 Method

Participants. There were a total of 74 participants: 12 (16.2%) were recruited through UCSCs participant pool for class credit, 52 (70.3%) were Mechanical Turk workers for $1, and 10 (13.5%) collected through convenience sampling of research assistants and their friends who were blind to the experiment. Thirty-five of the 74 participants completed a BFI inventory in addition to the single open-ended question.

Stimuli. Two 15-second clips were created using a single IVA with a covered face to avoid facial expression as a confound. He faced the participant and swayed at the identical times in both and gestured diagonally downwards from his shoulder with his arms. The agent that was programmed with proposed High Emotional Stability (non-neurotic) movement had smooth and sweeping (Smooth) movements, whereas the proposed Low Emotional Stability (neurotic) agent was characterized by jerky movements, reminiscent of someone who is shaking badly (Shaky).

Procedure. Prior to the start of the experiment, all participants were shown clips of recent video games to familiarize them with computer-generated animation that is not based on motion-capture, as our pilot data indicated that 10-15% of participants became preoccupied with how "unnatural" and "robotic" the virtual agent seemed. This habituation process increased the number of participants who were able to treat the agent as something that had a personality, though it did not reduce the number who provided a description about personality.

Using a between-subjects design to avoid carry-over effects, the participants watched either the Smooth or Shaky Agent clip. They were then immediately asked the open-ended question, "What personality does this animated character convey to you?" and given the option of watching the clip over again before answering this question. They were then given a modified version of the Big Five Inventory (BFI), a 44-item Big Five assessment that asks participants to rate the agent on perceived characteristics using 5-point Likert scales [8]. All items started with "If I had to guess, I would describe the character in the video as someone who" as opposed to the original "I am someone who" wording. They were not informed of the experimental aims until the end of the study.

2.2 Analysis

Open-Ended Question Coding. Participants' responses were coded using Goldberg's Big Five clusters, which consist of 339 trait adjectives [7]. If a response was already listed, then the factor it was associated with would be counted as the personality factor that the participant found most salient. If a participant's response was not found, two blind coders were asked to choose up to three of the closest Goldberg adjectives. All chosen adjectives had to apply to the same personality factor or they were excluded. For instance, the description "he has leadership qualities" was excluded because one coder chose "dominant" (high extraversion) and another chose "cooperative" (high agreeableness). The descriptor "arrogant" was kept because the one coder chose "boastful" and the other, "pompous" (both low agreeableness). They also excluded descriptions that did not describe personality (e.g., "is gesturing"), as well as those that defied a readily apparent single adjective description (e.g., "like a small boy"). This was done because 1. only half of participants listed multiple adjectives that all converged upon a single personality factor (some responses applied to as many as four) and 2., the experimenters wanted to avoid over-ascribing meaning to descriptions that may not have been intended by the participants.

Big Five Inventory Scoring. The BFI was scored according to John, Donahue, and Kentle [8] with personality scores calculated for all five dimensions: openness, conscientiousness, extraversion, agreeableness, and emotional stability (which they call "neuroticism", so that high neuroticism is equivalent to low emotional stability).

2.3 Results

Open-Ended Question. Sixty-one (28 Smooth Agent, 33 Shaky Agent) participants produced 121 descriptors. Thirteen participants (17.6%) were dropped because they did not ascribe any personality to the agent. Thirty-seven descriptors were dropped from the analysis because they were not Big Five personality traits or were too ambiguous to fit into a single Goldberg (1990) factor cluster. Only 16 exact descriptors were found in the Goldberg (1990) 339-adjective inventory, so the remainder of descriptors had to be matched by the blind coders. Additionally, 30 participants listed adjectives that were all aligned with the same personality factor (half were single-word responses).

Table 2. Frequencies of Goldberg (1990) adjectives and adjective-equivalents produced by participants by factor

	Openness		Conscientiousness		Extraversion		Agreeableness		Emotional Stability	
	High	Low	High	Low	High	Low	High	Low	High	Low
HES Agent	4	2	1	1	14	1	5	4	1	1
LES Agent	1	1	0	5	8	6	7	17	0	5

Fisher's exact test revealed that participants' spontaneous descriptions of a personality did differ depending on whether a participant watched a virtual agent designed to be either High Emotional Stability (Smooth) or Low Emotional Stability (Shaky) ($p = .01$). The modal personality trait by far for those observing the smooth agent was high extraversion while those observing the shaky agent mentioned low agreeableness the most. Emotional stability was infrequently mentioned, though the Shaky Agent was labeled as being low in emotional stability more often than the Smooth Agent, who was rarely talked about in those terms at all.

The open-ended question data suggested that the Smooth Agents gestures did not convey an emotionally stable personality but they did convey an extraverted personality. On the other hand, the Shaky Agent did seem to bear some gestural hallmarks of an emotionally unstable person, but most adjectives described him as having a disagreeable personality; that is, emotional stability was not the most salient trait conveyed.

Table 3. Structure matrix

	Function
Emotional Stability	0.90
Agreeableness	-0.74
Conscientiousness	-0.57
Extraversion	0.38
Openness	-0.29

Big Five Inventory. Examination of the BFI scale scores suggest that participants ascribed different personality traits to Smooth and Shaky Agents. T-tests revealed that there were significant differences between the agreeableness and the emotional stability ratings between the Smooth and Shaky Agents. The Shaky Agent ($M = 2.58$, $SD = 0.93$) was found to be significantly lower in agreeableness than the Smooth Agent ($M = 3.10$, $SD = 0.86$), $t(19) = -2.42$, $p = .02$, Cohen's $d = 0.70$. He was also found to be less emotionally stable (or more "neurotic") than the Smooth Agent, ($M = 3.42$, $SD = 0.95$ and $M = 2.81$, $SD = 0.80$ respectively), $t(72) = 2.92$, $p = .01$, Cohen's $d = 0.69$. These results suggest that a difference in gesture alone can influence perceptions of personality as measured by the BFI.

A discriminant analysis was run in order to ascertain whether one agent's gestures were better than the other in communicating personality. Echoing the t-tests, the structure matrix of correlations (see Table 3) suggested that the best predictor for distinguishing between Smooth and Shaky Agents was emotional stability, followed by agreeableness. The remaining BFI sub-scales are considered poor predictors. However, the agents were not equally effective in conveying personality. Classification results show a high rate of correct condition classification for participants who saw the agent that was meant to be low emotional stability, i.e., the Shaky Agent (86%). The success rate for the Smooth Agent participants was much lower (41.9%). Ideally, the patterns of BFI ratings would clearly indicate which video the participant watched, as the differences in gestures should reliably convey two distinct personalities. However, only the Shaky Agent conveyed a specific personality in a more consistent fashion. Specifically, his gestures appeared to communicate a less agreeable and emotionally unstable personality.

2.4 Discussion

The results of this study show that the gestures and postures that a virtual agent is programmed to use can influence the perception of its personality, if people make the initial leap to ascribe a personality to it, which over 80% of the participants in this study did. The results also show that open-ended questions and Big Five inventories can sometimes yield conflicting information about what personality viewers ascribe to an IVA. The influence of gesture was not the same for both Smooth and Shaky Agents. Overall, the Smooth Agent's gestures were somewhat unsuccessful in communicating a personality that was consistently identified by participants. The Shaky Agent, however, had his own distinct personality that came through from first impressions and in the BFI ratings.

The Shaky Agent was rated as being low in emotional stability using the BFI, which was the personality trait he was intended to evoke, though participants did not describe him as such in the open-ended question. On the other hand, he was also rated as being disagreeable on the BFI and spontaneously described as disagreeable in the open-ended question. This indicates that the participants' first impressions of him were of a person who is hard to get along with socially, not because he is anxious or negativistic (traits of low emotional stability) but because he is "angry" and "someone of bad intent."

Participants' first impression of the Smooth Agent indicated that they thought he was highly extraverted, suggesting that the gestures he was programmed with primarily communicated extraversion rather than high emotional stability. This is not necessarily surprising as the outward trajectory of the arm movement has been shown to be related to extraversion [18]; there is no reason to believe one gesture cannot be related to more than one trait. Yet the BFI ratings showed that when participants were asked to speculate in a more directed fashion using a closed-ended question on his personality, he was rated no more extraverted than the Shaky Agent. He was, however, judged to be more agreeable and emotionally stable than the Shaky Agent.

3 General Discussion

In reality, gestures are often an accompaniment to language and people are almost never without some environmental or interactional context, all of which will weight interpretation of personality. Nevertheless, personality can be distinguished via gesture alone, even in a completely silent, faceless, decontextualized virtual agent. This suggests that gesture can be an effective way to subtly convey personality in an agent without having to provide any dialogue or backstory, though determining the gestures that go with specific personality traits may be more problematic.

Previous work done by Neff et al. [14] found that the manipulation of an agent's non-communicative self-adaptor gestures (such as scratching) did influence perception of emotional stability. They also found that emotional stability and agreeableness were highly correlated linguistically but not gesturally; in fact, they found no effect of agreeableness for the non-verbal factors [14]. Our stimuli were completely non-verbal, slightly exaggerated versions of the Neff et al. stimuli. Differences in agreeableness were not intended or expected, although we did in fact find them.

It is possible that the nonverbal behaviors of agents are judged differently than those of humans, either because the variation in stylistic expression is viewed differently or because humans are reluctant to attribute personality traits based on an agent's hand and body movements. It could also be that judgments of agreeableness and emotional stability are highly correlated, particularly if both are negative: for instance, if someone is disagreeable, one might be more prone to labeling him emotionally unstable and vice versa.

The BFI scores were useful in quickly rating an IVA's personality, but it is only useful in comparison to another IVA. Both agents essentially scored neutral (between 2.6-3.6 range) ratings on all the BFI sub-scales: this is expected but not particularly informative. The analysis of the open-ended question was a more laborious process but provided data that gave a more vivid sense of what people thought of the agent's personality, without priming them with the assumption that any Big Five traits were present or forcing those who did not think the agent had a personality to choose one for it. The BFI tells us that the Shaky Agent is less agreeable than the Smooth Agent, but agreeableness has many facets (warmth, generosity, empathy, temperament, etc.) and the BFI cannot narrow it down further. On the other hand, the open-ended responses tell us that he is disagreeable because he is most often described as "angry."

Likewise, the most common responses for the Smooth Agent were "confused" and "open," two words that do not converge on any Big Five categories. Only "open" (if indicative of high extraversion) could have been anticipated as a perceived trait *a priori* based on the research that created the stimuli (as outward gestures are also characteristic of emotional stability). Yet the same Smooth Agent who was rated as more emotionally stable than the Shaky Agent was also frequently described as being "confused." This unanticipated result may have been overlooked if we had not given participants an open-ended question about their perception of the IVA's personality.

The use of both qualitative and quantitative data in early stages of developing an IVA designed to evoke a certain personality can be useful in understanding a user's perception. Future work will include a version of this study that uses gestures relating to extraversion, which is thought to manifest more physically than the other factors.

Acknowledgments. This research was supported by NSF grant IIS-1115742. We would like to thank our research assistants: Anna Wartan, Andrew Banakus, Jasmine Norman, Nicolas Russell, and Megan Kostecka.

References

1. Bryant, G., Fox Tree, J.: Recognizing verbal irony in spontaneous speech. Metaphor and Symbol 17, 99–119 (2002)
2. Christenfeld, N.: Choices from identical options. Psychological Science 6, 50–55 (1995)
3. Christenfeld, N., Creager, B.: Anxiety, alcohol, aphasia, and ums. Journal of Personality and Social Psychology 70, 451–460 (1996)
4. Costa, P., McCrae, R.: 'Normal' personality inventories in clinical assessment: General requirements and the potential for using the neo personality inventory: Reply (1992)
5. Fox Tree, J., Schrock, J.: Basic meanings of you know and i mean. Journal of Pragmatics 34, 727–747 (2002)
6. Funder, D., Dobroth, K.: Differences between traits: Properties associated with interjudge agreement. Journal of Personality and Social Psychology 52, 409–418 (1987)
7. Goldberg, L.: An alternative "description of personality": the big-five factor structure. Journal of Personality and Social Psychology 59, 1216–1229 (1990)
8. John, O., Donahue, E., Kentle, R.: The big five inventory—versions 4a and 54, Institute of Personality and Social Research, Berkeley (1991)
9. John, O., Robins, R.: Determinants of interjudge agreement on personality traits: the big five domains, observability, evaluativeness, and the unique perspective of the self. Journal of Personality 61, 521–551 (1993)
10. John, O., Srivastava, S.: The Big Five Trait taxonomy: History, measurement, and theoretical perspectives, pp. 102–138. University of California, Berkeley (1999)
11. LaFrance, B., Heisel, A., Beatty, M.: Is there empirical evidence for a nonverbal profile of extraversion?: A meta-analysis and critique of the literature. Communication Monographs 71, 38–48 (2004)
12. Lippa, R.: The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: Model analysis. Journal of Research in Personality 32, 80–107 (1998)
13. McCrae, R., Costa, P., Busch, C.: Evaluating comprehensiveness in personality systems: The california qset and the fivefactor model. Journal of Personality 54, 430–446 (1986)
14. Neff, M., Toothman, N., Bowmani, R., Fox Tree, J.E., Walker, M.A.: Don't scratch! self-adaptors reflect emotional stability. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 398–411. Springer, Heidelberg (2011)
15. Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 222–235. Springer, Heidelberg (2010)
16. Norman, W.: Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66, 574–583 (1963)
17. Paunonen, S.: Big five factors of personality and replicated predictions of behavior. Journal of Personality and Social Psychology 84, 411–424 (2003)
18. Riggio, R., Friedman, H.: Impression formation: the role of expressive behavior. Journal of Personality and Social Psychology 50, 421–427 (1986)
19. Tupes, E., Christal, R.: Recurrent personality factors based on trait ratings. Tech. rep., Lackland US Air Force (1961)

An Affective Virtual Agent Providing Embodied Feedback in the Paired Associate Task: System Design and Evaluation

Christian Becker-Asano[1], Philip Stahl[1], Marco Ragni[2], Matthieu Courgeon[3], Jean-Claude Martin[3], and Bernhard Nebel[1]

[1] University of Freiburg, Department of Computer Science, Georges-Köhler-Allee 52, 79110 Freiburg, Germany
{basano,stahl,nebel}@informatik.uni-freiburg.de
[2] University of Freiburg, Center for Cognitive Science, Friedrichstr. 50, 79098 Freiburg, Germany
ragni@cognition.uni-freiburg.de
[3] LIMSI-CNRS, Rue von Neumann, Batiment 508, 91403 Orsay Cedex, France
{martin,courgeon}@limsi.fr

Abstract. An affective, virtual agent is presented that acts as a teacher in the classical paired associate task. It is explained, why and how the virtual agent framework MARC was combined with the cognitive architecture ACT-R, the affect simulation architecture WASABI, and the voice-synthesis module OpenMARY. The agent's affective feedback capabilities are evaluated through an empirical study, in which participants had to solve association tasks. We expected that (1) the presentation of the task by a (neutral) virtual agent would change a learner's performance and that (2) the additional simulation and expression of emotions would impact a learner's performance as well. Finally, we discuss reasons for the lack of statistically significant differences as well as planned future application scenarios of our affective agent framework.

1 Introduction

In the domain of pedagogical agents [1] it has long been claimed beneficial to equip virtual agents with the capability to both convey as well as elicit emotions. The visual quality of interactive 3D computer graphics has increased dramatically since then and also the state-of-the-art in emotion simulation has significantly advanced. Although some evidence has been gathered for particular agents in particular scenarios [2–5], answering the general question of whether or not an agent should show emotions to improve a human's performance in human-computer interaction remains an open challenge.

We set out to test the influence of affective behavior shown by a virtual agent in face-to-face interaction with a human. We used the paired associate task (introduced by [6]) to see, if the participants' performance changes, when the task is presented by an affective as compared to an unemotional agent[1]. In addition,

[1] See also: http://www.youtube.com/watch?v=3BYTNxMs028

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 406–415, 2013.

our results are compared to those of the original study, in which the 20 associations between a single monosyllabic word and a digit between zero and nine were presented as text only.

In doing so, a novel combination of several independent software components was devised as to create high-quality, convincing agent behavior. Accordingly, after an overview of related work in the next section, the system is described in Section 3. Section 4 details an empirical study together with a presentation of its results. The latter are being discussed in Section 5, in which possible directions for future research are presented as well.

2 Related Work

Virtual agents are designed to capture the richness and dynamics of human behavior [7]. Different frameworks for virtual agents exist, e.g., the Virtual Human Toolkit [8], Greta [9], and MARC [10]. They differ in complexity, graphical output and application domains.

Concerning the applications of virtual humans, a study on emotional contagion between virtual and real humans [11] suggests that, although virtual agents can elicit this effect, it is dampened when humans have to make strategic decisions at the same time. Furthermore, in a conversational setting the smiling behavior of the virtual agent MAX has been shown to be mimicked by humans [12], but its likeability was not rated higher when it smiled more often.

Taking these limitations into account, we decided to integrate our system into a task that affords only very limited interaction capabilities of our virtual human, namely, the "paired associate task" [6, 13]. It is presented by Anderson as a common task to test the capabilities of human working memory.

We extend this previous work by adding a virtual agent's emotional feedback to it. The consequences of a virtual agent's affective feedback on students have previously been investigated in a learning environment [14]. However, they focused on "empathetic feedback" that was realized by "short, text-based responses" and not, as in the system described here, in terms of facial and vocal expressions. One study investigated how positive, neutral, and negative feedback responses from AutoTutor influenced learners' affect and physiology [15]. It was found that AutoTutor's feedback correlated with the learner's affect: after positive feedback from AutoTutor, learners mostly experienced delight, while surprise was experienced after negative feedback.

3 System

Our system realizes a virtual agent with the ability to show different emotions, produce verbal output, recognize speech input and the ability to predict human input in the paired associate domain.

To this end, the following five software components were combined (cp. Fig. 1):

- The MARC framework [10] realizes the visual output by providing an affective agent, which can be controlled in real-time.

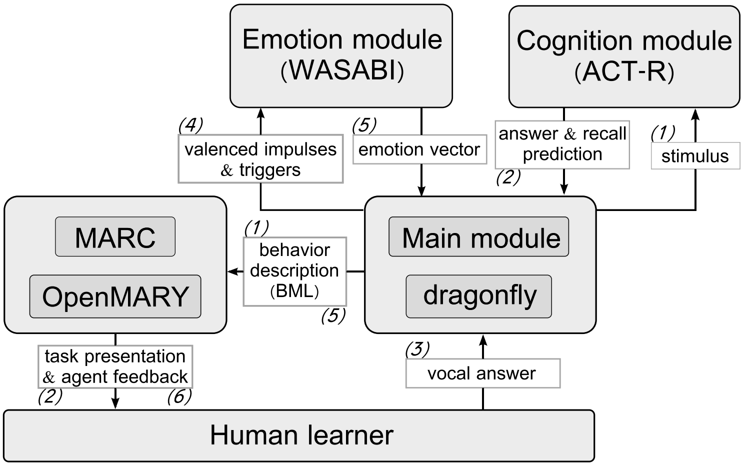


Fig. 1. System overview showing the interactions between the different modules

- The open source TextToSpeech software OpenMARY [16] generates verbal output with different affective connotations.
- The emotion module is realized by the open source, affect simulation architecture WASABI [17].
- The cognition module is a reimplementation of the memory activation function (probability of retrieval; [6]) that evaluates the likelihood to recall a specific memory chunk (and to give the correct answer). This gives an estimate of the difficulty of a retrieval.
- The main module uses Microsoft Windows 7 speech recognition via dragonfly [18] to receive user input.

In the following, the single modules and their interaction are described.

3.1 MARC and OpenMARY

MARC is a virtual agent framework for creating agents that allow real-time affective behavior. The agents' dynamic facial expressions [19] can be driven by emotions such as those provided by the WASABI architecture. The graphical output is rendered online, so that all predefined animations can be called in real-time. The agents within MARC are controlled via a dialect of the Behavior Markup Language (BML, [20]; see also *(1)* and *(5)* in Fig. 1). In combination with OpenMARY optimized lip synchronization of the rendered audio is achieved.

For the emotion mapping and the specification of animation parameters the system designer is provided with a graphical user interface. Animations can be assigned to each of the WASABI emotions, namely *Happy, Concentrated, Bored, Annoyed, Angry, Surprise, Fear, Hope, Relief* and *Fears-Confirmed*, and all their

Fig. 2. The virtual agent expressing *anger*, *neutral*, and *joy* (left to right)

MARC specific parameters like *intensity* and *interpolation* can be specified. For the empirical study reported here we limited MARC's expressions to *neutral*, *anger*, and *joy* (cp. Fig. 2), of which the latter two had been evaluated to be easily recognizable [19]. These particular emotions are also most relevant to the task in terms of influencing a student's motivation. WASABI's emotion *happy* is mapped to MARC's emotion *joy*. Furthermore, although WASABI distinguishes *annoyed* and *angry* (see Table 1), these emotions are both expressed by *anger* in the MARC framework.

OpenMARY is an open source text to speech system characterized by a modular design, an XML-based system-internal data representation and an easy to use interface which can be accessed via network protocols [16]. An important factor for achieving a realistic affective agent is the use of emotional speech synthesis. Thus, OpenMARY's German voice [21] was used to realize the affective states *anger*, *neutral*, and *happy* in the agent's vocal expressions.

3.2 WASABI and ACT-R

The agent's emotions are simulated dynamically by the WASABI affect simulation architecture [17]. Two types of input signals (cp. *(4)* in Fig. 1) are sufficient for WASABI to simulate the time course of primary and secondary emotions (see also [17, pp.148ff]):

1. *Valenced impulses* ranging from -100 to 100 are necessary to drive the internal emotion dynamics.
2. *Emotion triggers* are needed to maximize the intensity of either *angry*, *annoyed*, or *happy*. Any such maximized intensity drops off linearly to zero within ten seconds, before it is automatically reset to its predefined base intensity of 0.75.

These inputs are realized in terms of messages sent via UDP by the *main module*. WASABI, in turn, sends UDP-based messages back to the *main module* once per second containing the actual set of emotion/intensity pairs.

ACT-R functions about retrieval probability and latency are used to give the agent an understanding about the difficulty of recalling a certain item at a

Table 1. Derivation of the triggered emotion and the valenced impulses send to the emotion module depending on the discrepancy between expected and received answers. In the last column the facial expressions of the MARC agent are given associated with the WASABI emotion presented in the third column.

Expectation	human answer	emotion triggered	impulse	associated facial expr.
negative	none	none	0	neutral
negative	incorrect	annoyed	-30	anger
negative	correct	happy	80	joy
none	none	annoyed	-20	anger
none	incorrect	angry	-50	anger
none	correct	happy	50	joy
positive	none	annoyed	-50	anger
positive	incorrect	angry	-80	anger
positive	correct	happy	30	joy

certain time in the experiment. These estimations enable the agent to predict how likely the participant will give the correct answer.

When the *main module* receives the learner's answer, the *valenced impulses* are derived from this answer and the agent's expectation; cp. Table 1. If the learner's answer is correct and this was highly expected, the resulting impulse is only slightly positive and the emotion *happy* is triggered. An unexpected, correct answer, however, would also trigger *happy*, but the positive impulse would be very strong. Incorrect answers are treated in a similar fashion. The system designer can specify the mapping parameters from recall probabilities provided by ACT-R to expectation values to impulse intensities in the overall framework. The probability values are mapped to the three types of expectations as follows:

- Probabilities less than 60%: negative expectation, i.e. the answer is expected to be incorrect
- Probabilities between 60% and 90%: no expectation, i.e. the agent is unsure about which answer to expect
- Probabilities greater than 90%: positive expectation, i.e. the answer is expected to be correct

To compute the recall probabilities the base-level activation function and the probability function are reimplemented as presented by Anderson in [13, pp.74/124]. The threshold τ and the noise variable s are set to -2.0 and 0.5, respectively. These values correspond to those proposed in the ACT-R tutorial.

3.3 Interaction of the Modules

The *cognition module* receives a task description from the *main module* and predicts the human learner's answer. After the learner's answer has been recognized by *dragonfly*, the *main module* triggers emotions and sends *valenced impulses* to the *emotion module*. In parallel, it listens continuously for WASABI messages containing an *emotion vector* to update the agent's affective state. The emotion with the highest intensity is encoded into a BML message to update the MARC agent's facial display accordingly. Furthermore, MARC passes on the emotion information to OpenMARY such that its synthesis is changed accordingly.

The result of the implementation is an affective agent that acts in the domain of paired associate learning as an affective tutor. Instead of providing feedback about possible error sources, it shows affective behavior according to the participant's performance hopefully further motivating him to improve performance.

4 Empirical Study

We tested this system on the paired-associate task [6]. In the original study 20 association pairs are displayed for eight rounds on the screen in random order. First, the word is displayed for five seconds, then the number. From round two to eight the participants have to recall the number within five seconds after the associated word has been presented by pressing the correct key on the keyboard.

4.1 Hypothesis and Research Question

We set out to investigate two questions regarding the paired-associate task:

1. Does multimodal task presentation performed by a virtual agent change a human learner's performance?
2. How does an agent's emotional feedback impact the learner's performance, if at all?

Previous research has shown that the effects of an agent's empathetic feedback depend on whether or not a strategic task has to be solved by the human at the same time [11]. As the paired associate task is a learning and not a strategic task, we expected a significant change of the learner's performance, when this task is presented by an agent. We were unsure, however, if the integration of emotional feedback into this non-strategic task further impacts performance.

Accordingly, we established a *neutral condition*, in which the participants are expected to achieve different correctness rates than the ones reported in the original (no-agent) study. For the emotional feedback we compared a simple, rule-based approach, i.e. the *reflexive condition*, with the dynamic emotion approach realized by the integration of WASABI, namely the *WASABI condition*.

4.2 Design and Procedure

The three conditions differ in the following way:

1. *Neutral condition*: the agent shows no emotional expressions at all. This serves as a control condition testing the impact of the task presentation by an embodied agent.

2. *WASABI condition*: the emotion module is enabled. As a result the agent can change its mimic and connotation of speech according to the changing affective course in WASABI as described in Section 3.2 taking ACT-R-driven expectations into account.
3. *Reflexive condition*: the agent shows a strictly rule-based behavior by always showing an *angry* emotion in response to an incorrect answer and a *joy* emotion in response to a correct answer.

The general procedure of the paired-associate task remained the same. However, in addition to a visual presentation on the screen, the MARC agent says the word and the number at the time of their resp. presentation. In addition, our participants were requested to respond verbally from round two instead of using the keyboard. The agent then reacted to the correctness of the answer by nodding or shaking its head. The experimental setup is presented in Fig. 3, left, and the agent as it is displayed in Fig. 3, right.

Fig. 3. The experimental setup (left); the agent as it is displayed on the screen with subtitles in addition to its verbal output

4.3 Participants

Data of 60 participants were collected, of which two had to be excluded, because they did not achieve a correctness rate of 60 percent or higher in any of the seven runs (similarly to the original study [6]). The data of two more participants were removed, because they had misunderstood the instructions.

The remaining 56 participants were on average 23.3 years old with 32 of them being female. They were randomly assigned to the neutral condition (22 subjects, 14f, *mean* = 21.5 years), the WASABI condition (18 subjects, 8f, *mean* = 24.3 years), and the reflexive condition (16 subjects, 10f, *mean* = 23.7 years).

4.4 Results and Discussion

The results of the experiment are summarized in Table 2 comparing the values of the original study [6] with our conditions. The data sets of the conditions

Table 2. Correctness rates (in percentages) with standard deviations for the different conditions. For the original study [6] only the overall average of the standard deviations 2.6% is available.

	ORIG. STUDY		NEUTRAL		WASABI		REFLEXIVE	
Run	mean	std	mean	std	mean	std	mean	std
2	52.6	(2.6)	59.6	24.4	58.1	20.8	61.3	14.1
3	66.7	(2.6)	69.8	22.1	66.9	25.9	70.7	16.3
4	79.8	(2.6)	82.1	17.0	76.9	23.2	81.7	11.2
5	88.7	(2.6)	88.0	14.4	80.6	20.9	92.0	9.2
6	92.4	(2.6)	92.7	10.1	86.7	17.4	92.3	8.8
7	95.8	(2.6)	92.7	10.7	88.6	12.5	95.7	5.1
8	95.4	(2.6)	92.7	10.1	93.1	11.7	98.3	2.9

were analyzed by computing repeated measures ANOVAs between subjects with the statistical software *R*. All pairwise comparisons between conditions reveal no significant difference (*neutral* against *WASABI condition* $F(7, 296) = 0.112$, n.s.; *neutral* against *reflexive condition* $F(7, 272) = 0.052$, n.s.; *WASABI* against *reflexive condition* $F(7, 240) = 0.571$, n.s.). The comparison of our conditions with the original study's data has to remain on an informative level in lack of the raw data of the Anderson study. The participants' overall correctness rate in the *neutral condition* was on average 0.9% better than that reported in the original study. The correctness rates achieved in the *WASABI condition* were on average 2.9% worse than that of the original study and the one of the *reflexive condition* was 2.9% better. With regard to our research questions, the presence of an agent as teacher in the paired associate task seems not to change a learner's performance significantly. Surprisingly, a learner's correctness rates seem to benefit from an agent's affective feedback, only if it is achieved by a rather simple, reflexive emotion simulation.

A number of reasons for the insignificant differences can be speculated about:

– The task was so simple that most students reached very high recall rates early on, so that the agent's display of negative emotions could not really interfere with the memory task.
– The choice of parameters within our framework might not be optimal. The total presentation time of *joy* and *anger* was very different: In the Wasabi condition *anger* was shown on average for 46.3 seconds (3.2%) during the whole experiment, while *neutral* was presented for 493.9 seconds (34.6%) and *joy* for 887.8 seconds (62.2%). In the reflexive condition learners were on average confronted with *anger* for 212.4 seconds (14.9%), with *neutral* for 217 seconds (15.2%) and with *joy* for 998.6 seconds (69.9%). Maybe showing *anger* more often would enlarge the agent's impact on human learners.
– The learners might have been so focused on their task that an agent's affective feedback was largely ignored similar to the effects reported in [11] for strategic tasks.

5 General Discussion

We set out to investigate (a) the effect of a virtual agent's presence in and active presentation of a learning task and (b) the additional effects of the agent's affective feedback during task presentation. In doing so, we implemented a new experimental agent framework that combines the state-of-the-art virtual agent framework MARC with an affective component based on WASABI, a cognitive component based on ATC-R, the voice synthesis component OpenMARY, and a voice recognition component based on dragonfly.

Although the results of the empirical study remain inconclusive, it needs to be pointed out that the experimental framework can easily be modified and extended. The following tasks will be approached next:

- Extending the number of emotions as provided by WASABI to be integrated into the simulation and displayed by the agent.
- Online emotion recognition by means of physiological sensors, facial features and/or eye tracking.
- Changing the task to one that affords more direct interaction with the agent, possibly a game like chess, e.g., similar to [22].

In conclusion, we believe that this framework can serve as a flexible test environment for further psychological studies on the effects of emotions in human-computer interaction.

References

1. Johnson, W.L., Rickel, J.W., Lester, J.C.: Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education 11, 47–78 (2000)
2. Conati, C.: Probabilistic assessment of user's emotions in educational games. Applied Artificial Intelligence 16, 555–575 (2002)
3. Qu, L., Wang, N., Johnson, W.L.: Pedagogical agents that interact with learners. In: AAMAS 2004 Workshop on Balanced Perception and Action in ECAs (2004)
4. Prendinger, H., Becker, C., Ishizuka, M.: A study in users' physiological response to an empathic interface agent. Intl. Journal of Humanoid Robotics 3(3), 371–391 (2006)
5. Hall, L., Woods, S., Aylett, R.: Fearnot! involving children in the design of a virtual learning environment. Intl. Journal of Artificial Intelligence in Education 16(4), 327–351 (2006)
6. Anderson, J.R.: Interference: The relationship between response latency and response accuracy. Journal of Experimental Psychology: Human Learning and Memory 7(5), 326–343 (1981)
7. Gratch, J., Rickel, J., André, E., Badler, N., Cassell, J., Petajan, E.: Creating interactive virtual humans: Some assembly required. IEEE Intelligent Systems 17, 54–63 (2002)
8. Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S.C., Piepol, D.: Building Interactive Virtual Humans for Training Environments. In: Interservice/Industry Training, Simulation and Education Conference (2007)

9. Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C.: Greta: an interactive expressive eca system. In: Proc. Intl. Conf. on Autonomous Agents and Multiagent Systems, AAMAS 2009, Richland, SC, pp. 1399–1400 (2009)
10. Courgeon, M., Martin, J.C., Jacquemin, C.: MARC: a Multimodal Affective and Reactive Character. In: Proc. 1st Workshop on Affective Interaction in Natural Environements (2008)
11. Tsai, J., Bowring, E., Marsella, S., Wood, W., Tambe, M.: A study of emotional contagion with virtual characters. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 81–88. Springer, Heidelberg (2012)
12. Krämer, N., Kopp, S., Becker-Asano, C., Sommer, N.: Smile and the world will smile with you – the effects of a virtual agent's smile on users' evaluation and behavior. Intl. Journal of Human-Computer Studies 71(3), 335–349 (2013)
13. Anderson, J., Lebiere, C.: The Atomic Components of Thougt. Taylor & Francis (1998)
14. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems. In: Affective Computing and Intelligent Interaction and Workshops, pp. 1–6 (2009)
15. Aghaei Pour, P., Hussain, M.S., AlZoubi, O., D'Mello, S., Calvo, R.A.: The impact of system feedback on learners affective and physiological states. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 264–273. Springer, Heidelberg (2010)
16. Schröder, M., Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology 6, 365–377 (2003)
17. Becker-Asano, C.: WASABI: Affect simulation for agents with believable interactivity, vol. 319. IOS Press (2008)
18. Butcher, C.: Dragonfly speech recognition source code (April 2013), `http://code.google.com/p/dragonfly/`
19. Courgeon, M., Clavel, C., Tan, N., Martin, J.C.: Front view vs. side view of facial and postural expressions of emotions in a virtual character. In: Pan, Z., Cheok, A.D., Müller, W. (eds.) Transactions on Edutainment VI. LNCS, vol. 6758, pp. 132–143. Springer, Heidelberg (2011)
20. Vilhjálmsson, H.H., et al.: The behavior markup language: Recent developments and challenges. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 99–111. Springer, Heidelberg (2007)
21. Steiner, I., Schröder, M., Charfuelan, M., Klepp, A.: Symbolic vs. acoustics-based style control for expressive unit selection. In: ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-7), Kyoto, Japan, ISCA (2010)
22. Leite, I., Martinho, C., Pereira, A., Paiva, A.: iCat: an affective game buddy based on anticipatory mechanisms. In: Proc. 7th Intl. Joint Conf. on Autonomous Agents and Multiagent Systems, AAMAS 2008, Richland, SC, vol. 3, pp. 1229–1232 (2008)

Expressive Animation

Towards a Procedural Solution

Ewan McGregor and Robin J.S. Sloan

University of Abertay, Dundee
{0402429,r.sloan}@abertay.ac.uk

Abstract. This project aimed to establish the feasibility of creating a procedural system for generating expressive facial animation based on an affective agent. A procedural system supporting a limited number of emotional expression changes was created alongside keyframed animations of these same emotional expression changes, and audience response to these two approaches was tested empirically. Results seem to partially support the procedural animations generated being comparable with keyframed, in terms of perceptual validity.

Keywords: Animation, Emotional Expression, Facial Expression, Affective Agents.

1 Introduction and Background

Many recent commercial interactive media titles (and a number of notable research projects) have made use of affective agent driven characters rather than relying upon prewritten/scripted character interaction.

This use of procedurally driven characters has been variable in success within entertainment, but seems likely to increase in use as a technique in future. However, the reliance of these systems on pre-established libraries of animation seems to be a restriction on the freedom from predetermination gained through the use of an affective AI character system.

Given this, would it be possible to establish a procedural system for the generation of facial expression, based upon an existing affective agent system, which is capable of competing with manually keyframed or motion captured animation in terms of perceptual validity?

2 Project Methodology

The approach taken to the development of this project was primarily practice based. A basic procedural animation system based on the researcher's own knowledge of animation, and established animation principles (e.g. Thomas and Johnston[1]) was designed and then iteratively improved upon. A series of keyframed animations depicting the same emotional reactions was produced concurrently, going through a similar process of iterative development.

This system did not attempt to be an exhaustive solution to the expression of human emotion via facial expression, but focused on a small number of pre-established emotional expressions. These were selected based upon the expressions identified as having evidence of being universal by Ekman[2].

The affective agent serving as an input to the procedural animation system was based on the FAtiMA[3] system's emotion model, itself derived from the Ortony, Clore and Collins model of emotion activation[4]. This was done in order to allow the procedural system to act as though it were driven by an affective agent, even if connecting it to a real agent was beyond the scope of this project.

After a number of iterations, the procedural animation system's output was tested against the keyframed animations. This test was done using a large sample group and a simple online questionnaire, testing the effectiveness of the animations in terms of whether they were convincing expressions of an emotion. As the iterative improvements were based upon established animation principles, the researcher's own responses to the animation system, and that of fellow animation practitioners, this test hoped to establish the merits of the system as seen by a larger, less expert audience.

3 Results

Animations for expressions of Anger, Happiness, and Sadness[1] were created. These were then tested using the questionnaire system discussed in the previous section, and results of the test were evaluated using a Chi-squared goodness-of-fit test, considering responses to the procedural as the observed, and the keyframed as the expected.

For expressions of happiness and sadness, no significant difference was found in audience perception between the keyframed and procedural animations.

For the expression of anger, the procedural animation was considered to be significantly less convincing expression of emotion than the keyframed.

This would seem to partially confirm that procedural systems for facial animation can be comparable to keyframed animation, although further investigation into the expression of anger is indicated.

References

1. Thomas, F., Johnston, O.: The Illusion of Life: Disney Animation. Hyperion, New York (1995)
2. Ekman, P.: Facial Expression. In: Seigman, A., Feldstein, S. (eds.) Nonverbal Behavior and Communication. Lawrence Erlbaum Association, New Jersey (1977)
3. FearNot! http://sourceforge.net/projects/fearnot/
4. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)

[1] Mapping onto the OCC model emotions of Anger, Joy and Distress, respectively.

Interactive Machine Learning for Virtual Agents

Marco Gillies[1] and Andrea Kleinsmith[2]

[1] Embodied Audio-Visual Interaction Group, Goldsmiths' Digital Studios,
Department of Computing, Goldsmiths, University of London, UK
m.gillies@gold.ac.uk
http://www.doc.gold.ac.uk/eavi
[2] Virtual Experiences Research Group
Department of Computer and Information Science and Engineering
University of Florida

Machine learning has become a popular method for developing virtual agents as it makes it possible to develop realistic behaviour models based on real human behaviour. It is also well suited to situations where finding explicit algorithms for simulating behaviour is difficult, such as simulating non-verbal interaction. Many researchers have used learning algorithms to approximate examples of real behavior. For example, Lee and Marsella [4] have used a corpus of conversational behavior to model non-verbal aspects of speaker behaviour. Similarly, Gillies [2] has created agents that respond expressively to game events (in this case, spectators at a football match). This was learned from motion capture of a person responding to example events.

However, while machine learning means a behaviour model can closely simulate the behaviour of a real person, it is difficult for designers to have control of the learning process in order to shape the design of the behaviour. This means that designers have little scope for designing the model so that the agent behaves in the way they want it to, beyond a painstaking and difficult process of data gathering and labelling. Even if the data was gathered very carefully the results can be hard to predict and difficult to edit. We propose to use a new approach called interactive machine learning, in which the human is not simply a source of data but is actively involved in guiding the learning process.

Interactive Machine Learning is a new approach to machine learning in which user interaction is central to the learning process. Rather than simply providing a fixed data set to a batch process, users add data and tune parameters interactively, progressively refining the machine learning model based on interactive testing. This approach has the potential to fundamentally change the use of machine learning in interface design. The interaction and progressive refinement can make machine learning into a genuine design tool in which a designer has fine control of the resulting interface, as opposed to a batch approach where the designer has to simply prepare the data and hope that the result comes out as expected. The term Interactive Machine Learning (IML) was introduced by Fails and Olsen [1] who saw it as a way of involving users more closely in the machine learning process by interactively supplying and editing training data. Their Crayons system enables non-expert users to create image processing classifiers in an iterative process by drawing on images that they provide as training data.

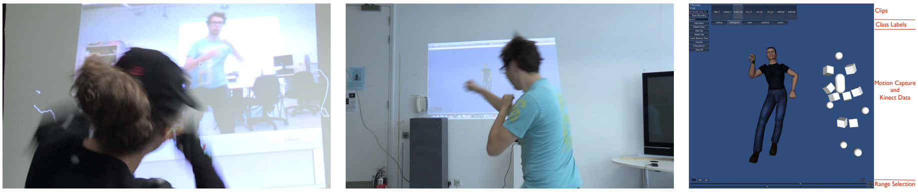


Fig. 1. A participant interacting with the virtual agent (centre), an actor controlling the virtual agent via motion capture (left), and the labelling interface

We have developed a system based on Interactive Machine Learning that allows participants to design the behaviour of a virtual agent that interacts with a real human using full body non-verbal actions. It uses an interactive machine learning methodology, in which participants examples of behaviour in order to train the model. A player interacts with a virtual agent (Figure 1, left) that is either controlled live by a performer via motion capture (Figure 1, centre) or animated based on the output of our machine learning engine. The performer is able to improvise actions and responses with the player via the virtual agent, by controlling it in real time from motion capture. The system is trained based on the results of these improvisations. The movements of both player and performer are recorded. The performer then uses an editing interface (Figure 1, right) to segment particular movements and label them as examples of particular actions and reactions. The results of this labelling operation are used to train a machine learning model, which is then able to control the virtual agent autonomously, classifying new actions by the player and using this classification to select a suitable response animation, from the data recorded by the performer.

References

1. Fails, J.A., Olsen Jr., D.R.: Interactive machine learning. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, IUI 2003, pp. 39–45. ACM, New York (2003), `http://doi.acm.org/10.1145/604045.604056`
2. Gillies, M.: Learning finite-state machine controllers from motion capture data. IEEE Transactions of Computational Intelligence and AI in Games 1, 63–72 (2009)
3. Kipp, M.: Creativity meets automation: Combining nonverbal action authoring with rules and machine learning. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 230–242. Springer, Heidelberg (2006), `http://dx.doi.org/10.1007/11821830_19`
4. Lee, J., Marsella, S.: Modeling speaker behavior: A comparison of two approaches. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 161–174. Springer, Heidelberg (2012), `http://dx.doi.org/10.1007/978-3-642-33197-8_17`

An Eye Movement Model for the Purpose of Emotional Expression

Chen Jiang[1], Jun Yu[1,3], and Zengfu Wang[1,2,3,⋆]

[1] Department of Automation, University of Science and Technology of China, China
[2] Institute of Intelligent Machines, Chinese Academy of Sciences, China
[3] National Laboratory of Speech and Language Information Processing , University of Science and Technology of China, China
jason001@mail.ustc.edu.cn, {harryjun,zfwang}@ustc.edu.cn

Abstract. This paper presents an eye movement model which can be used by virtual agent to express different emotions through vivid eye movements. First, on the basis of statistical data and empirical studies from ophthalmology and psychology, we build our eye model with detailed motion parameters and functions which can show subtleties of the spatial and temporal aspects of eye movement under neutral emotion condition. Second, we perform eye tracking experiment to find the impact of emotional change on eye movements, and analysis results are used to model eye movements under different emotions condition.

1 System Overview

Our eye movement model describes movements for eyeballs and eyelids respectively. For eyeballs, the movements involve eye saccades, and parameters for describing eye saccades include amplitude, direction and duration. For eyelids, the movements involve lid saccades and blinks, and parameters for describing lid saccades and blinks include amplitude, duration, lid position and blink rate. Eyeballs movement is concomitant with eyelids movement at some aspects, which is illustrated by our model.

2 Eye Movement Model under Neutral Emotion Condition

On the basis of previous work including some statistical data and empirical studies, we give detailed equations to describe those parameters. For eyeballs, the equations for amplitude, direction and duration of eye saccades are given by [1]. For eyelids, the equations for amplitude, duration, lid position of lid saccades can be found at [2,3], and equations for amplitude, duration, rate of blinks can also be found at [2].

3 Eye Movement Model under Different Emotions Condition

But all these descriptions from previous work were not associated with emotional change, thus in order to fill this research gap, we implement our own eye-tracking experiment.

⋆ Corresponding author.

The tracking data demonstrates emotional change is closely related to three parameters, which respectively are duration of eye saccades, eyelid position and blink rate. Through analyzing tracking data, we give three equations to describe these three parameters.

The description of emotion in our system is based on Whissel's measuring emotions theory [4], which proposed a two-dimensional emotion-space, the activation dimension (aroused-sleepy) and the evaluation dimension (happiness-sadness). In our system, activation value a and evaluation value e are ranged from 0 to 10, while the median value of activation value $\overline{a}$ and that of evaluation value $\overline{e}$ are both 5.

Duration of Eye Saccades: According to the tracking data, the saccades duration related to emotional change can be represented as equation below, which denotes the more active emotion is (activation value is bigger), the faster eye saccade is, and vice versa.

$$D_{Eye} = \frac{15 - a}{2 * \overline{a}} * (A_{Eye} + 20) \tag{1}$$

where D_{Eye} is the duration of eye saccade (in millisecond), A_{Eye} is the amplitude of eye saccade (in degrees).

Position of Eyelid: According to the tracking data, the position of lid related to emotional change can be represented as equation below, which denotes the more aroused emotion is (activation value is bigger), the larger eyelid opening is, and vice versa.

$$P_{UpperLid} = P_{Eye}(vertical) + 20 + (a - 5) \tag{2}$$

where $P_{UpperLid}$ and $P_{Eye}(vertical)$ are the upper lid position and the vertical position of eyeball respectively.

Blink Rate: According to the tracking data, blink rate related to emotional change can be represented as equation below, which denotes the more positive emotion is (evaluation value is bigger), the higher blink rate will be, and vice versa.

$$Rate_{Blink} = \frac{e + 5}{2 * \overline{e}} * rand(8.0, 21.0) \tag{3}$$

where $rand(8.0, 21.0)$ means a rand value produced between 8.0 and 21.0.

Acknowledgements. This paper is supported by the Fundamental Research Funds for the Central Universities of China (No. WK2100100009),and the China Postdoctoral Science Foundation under(No. 2012M521248).

References

1. Lee, S., Badler, J., Badler, N.: Eyes alive. ACM Transactions on Graphics, TOG (2002)
2. Evinger, C., Manning, K., Sibony, P.: Eyelid movements. Mechanisms and normal data. Investigative Ophthalmology & Visual Science 32, 387–400 (1991)
3. Becker, W., Fuchs, A.: Lid-eye coordination during vertical gaze changes in man and monkey. Journal of Neurophysiology 60, 1227–1252 (1988)
4. Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. Australian Journal of Psychology 51, 154–165 (1999)

Effects of Users' Social Skill on Evaluations of a Virtual Agent That Exhibits Self-adaptors

Tomoko Koda and Hiroshi Higashino

Graduate School of Information Science and Technology
Osaka Institute of Technology, Osaka, Japan
koda@is.oit.ac.jp

Abstract. Self-adaptors are bodily behaviors that often involve self-touch. Our continuous evaluation of the interaction between an agent that exhibits self-adaptors and without indicated that there is a dichotomy on the impression on the agent between users with high social skills and those with low skills. People with high social skills feel more friendliness toward an agent that exhibits self-adaptors than those with low social skills. The result suggests the need to tailor non-verbal behavior of virtual agents according to user's social skills.

Keywords: gesture, self-adaptors, non-verbal behavior, social skills, evaluation.

1 Introduction

Self-adaptors are non-signaling gestures that are not intended to convey a particular meaning. Because of its non-relevance to conversational content, there has not been much IVA research done on self-adaptors, compared with nonverbal communication with high message content. Neff et al. reported that an agent performing self-adaptors was perceived as having low emotional stability and suggests the importance of self-adaptors in conveying a personality of an agent [1]. However, self-adaptors are not always the sign of emotional unstableness or stress. Blacking [2] states self-adaptors also occur in casual conversations, where conversant are very relaxed. If those relaxed self-adaptors occur with a conversant that one feels friendliness, one can be induced to feel friendliness toward a conversant that displays self-adaptors. We apply this to the case of agent conversant, and assume that users can be induced to feel friendliness toward the agent by adding self-adaptors to IVAs. We made the following hypothesis: "Compared with people with low social skills, people with high social skills have a greater sense of friendliness toward an agent that exhibits self-adaptors."

We conducted an experiment with IVA in order to verify the hypothesis. The agent character and animation of the three types of self-adaptors were created. Figure 1 shows the agent carrying out the movements of "touching hair", "touching nose", and "touching face". Participants were asked to carry a pseudo-conversation with the agent five times (one per day). 24 Japanese male students' (aged 19-24 years) social skills were measured beforehand using KiSS-18 (Kikuchi's Scale of Social Skills: 18 items) [3]. Before the start of the experiment, they were separated into a high social skills (HSS: n=11) group and a low social skills (LSS: n=13) group. The conditions of the experiment were social skills (HSS group, LSS group), type of agent (with self-adaptors, without self-adaptors), and trial number (1st, 2nd, 3rd, 4th, 5th). After each conversation, the participated rated their impressions on the agent using a semantic differential method on a scale from 1 to 6.

Fig. 1. Agents that exhibit "touching hair", "touching nose", and "touching face" self-adaptors

2 Results and Discussion

We ran three-way ANOVA with factors "social skills" (HSS, LSS), "self-adaptors" (with, without), and "number of trials" (1st, 5tt). We found a dichotomy between the use's social skills on the perceived friendliness of the agent. Compared with the LSS group, the HSS group rated significantly higher friendliness toward the self-adaptor-performing agent after both the 1st and the 5th trial. The HSS group evaluated agents that performed self-adaptors to be significantly friendlier after the 5th trail than after the 1st trial. The LSS group rated agents that did not perform self-adaptors to be significantly friendlier after the 5th trial than after the 1st trial. Regardless of the number of trials in this experiment, the HSS group had a significantly higher sense of friendliness toward the agent that performed self-adaptors than the LSS group did. Also, because there was not much difference between the LSS group's scores for the condition of self-adaptors and number of trials, we believe that it was not the case that the LSS group did not have a sense of friendliness toward the agent with self-adaptors; rather, the HSS group felt a stronger sense of friendless toward the agent.

From these results, our hypothesis was supported. Our results suggest the importance of changing the level of displaying self-adaptors of IVAs according to the users' social skills. Also suggested by the results is the possibility that a sense of friendliness toward the agent can be increased in a continual manner by taking into account the level of the users' social skills and whether or not to have the agent perform self-adaptors during continued interactions.

References

1. Neff, M., Toothman, N., Bowmani, R., Fox Tree, J.E., Walker, M.A.: Don't Scratch! Self-adaptors Reflect Emotional Stability. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS (LNAI), vol. 6895, pp. 398–411. Springer, Heidelberg (2011)
2. Blacking, J. (ed.): The Anthropology of the Body. Academic Press, London (1977)
3. Kikuchi, A.: Notes on the researches using KiSS-18 Bulletin of the Faculty of Social Welfare. Iwate prefectural University (2004) (in Japanese)

Activity Planning for Long-Term Relationships*
(Extended Abstract)

William Coon, Charles Rich, and Candace L. Sidner

Worcester Polytechnic Institute
Worcester, MA, USA
{rich,sidner}@wpi.edu

We have implemented a general-purpose algorithm for planning appropriate joint activities in the context of an interactive system that has a long-term relationship with its user. The algorithm is data-directed and explicitly models the difference between relationship stages, such as stranger, acquaintance and companion. We have conducted a short laboratory evaluation of the algorithm that demonstrates the plausibility of its results according to the judgements of participants.

One of the main methods for developing closeness is appropriate shared activities. For example, an electronic home companion for isolated older adults might support a wide range of activities, including chatting about the weather or sports, assisting with the maintenance of a personal appointment calendar, and coaching the user to get more exercise. Furthermore, the user might interact with the system several times per day for weeks or months or more. Certain activities, such as chatting about the weather, are appropriate on the very first day of interaction, while other activities, such as exercise coaching, should wait until the system and user develop a closer relationship. Even with a close friend, however, you don't normally start a conversation with a very difficult topic, such discussing a serious illness, but rather build up to it with social chit-chat first.

Our basic approach is a planning algorithm, in which both the specifics of the activities and interaction with the user are abstracted, as shown in Fig. 1, so that it can be applied to any system that seeks to develop a long-term relationship with its user. Whereas other parts of the interactive system are busy managing the moment-by-moment details of interaction, the role of the activities planner is to take a long-term view of the relationship with the user. In particular, the activities planner is

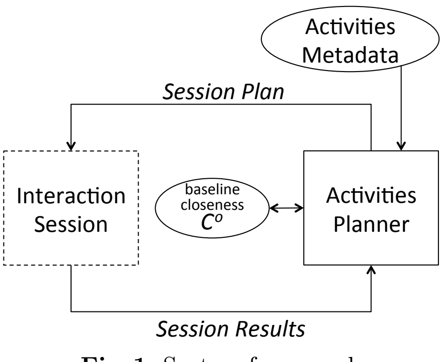


Fig. 1. System framework

* See http://www.cs.wpi.edu/~rich/always/pub/CoonRichSidner2013_IVA.pdf for full paper. This work is supported in part by the National Science Foundation under award IIS-1012083. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

concerned with increasing the closeness of the relationship through appropriate choice of joint activities. Closeness is important both for its own sake and because it is a prerequisite for some useful activities.

At the start of each session, the activities planner produces a conditional high-level *session plan*, that specifies the order of possible activities to suggest, such as "baseball chat" followed by "calendar help," but not the specific details of the activity, such as the dialogue, that will occur.

The key state variable in the framework in Fig. 1 is $C^o \geq 0$, the *baseline closeness*. This is the closeness level at the start of the next interaction session. It is used by the planner in the planning process and updated at the end of each interaction session, based on the results of the session. The main input to the planning process is a database of *activities metadata*. For each activity, the planner needs to know the following five items:

- C_A^o, the minimum baseline closeness at which this activity becomes available,
- C_A, the minimum closeness required to start this activity,
- Δt_A, the expected duration of activity in minutes,
- I_A, the expected instrumental utility of this activity, and
- ΔC_A, the expected relational utility of this activity.

The *instrumental utility* is intended to capture the practical benefit of the activity, whereas the *relational utility* is intended to capture the contribution of an activity to increasing the closeness of the relationship, i.e., its purely social benefit. For example, watching TV might have no instrumental utility and a small relational utility, whereas discussing a serious medical diagnosis may have both high instrumental utility and high relational utility.

The ultimate evaluation of the activities planner will be to use it in a long-term field study. In the meantime, however, we undertook a laboratory user study to evaluate whether the planner produces plans that at least are consistent with general expectations of how relationships develop. Our experimental approach was to see whether changing key features of the planner's algorithm had a measurable effect on the naturalness and plausibility of the resulting plans. Specifically, we compared our planning algorithm (using manually created metadata) to a version that was modified to be "antisocial" by violating three key features of the planning algorithm.

Table 1. Evaluation results

	social > anti	anti > social	p-value
$C^o = 0$ *(stranger)*	40	5	$7.88e^{-8}$
$C^o = 2$ *(acquaintance)*	38	5	$2.5e^{-7}$
$C^o = 4$ *(companion)*	27	17	0.174
Overall	105	27	$2.07e^{-11}$

We conducted a within-subject study with 12 participants and 6 conditions. The participants completed questionnaires evaluating the naturalness and plausibility of hypothetical scenarios automatically generated by the normal (social) and antisocial planners describing interactions between a community worker, Samantha, visiting an older adult, Katherine, in her home. We scored the data by counting the number of times that a social scenario was preferred to an antisocial scenario and vice versa (see Table 1). According to our hypothesis, the social scenarios should be preferred to the antisocial scenarios. This is strongly supported overall and for both the stranger and acquaintance conditions individually.

Finally, we want to point out that this work only scratches the surface of computationally modeling long-term human-computer relationships. For example, we have ignored the important effects of factors such as relative status, gender and personality, to name just a few. There is a rich literature on all of these topics that is waiting to be adapted into practical algorithms.

Realistic Facial Animation Driven by a Single-Camera Video

Changwei Luo[1], Jun Yu[1,2], and Zengfu Wang[1,2,3]

[1] Dept. of Automation, University of Science and Technology of China
[2] National Laboratory of Speech and Language Information Processing, University of Science and Technology of China
[3] Institute of intelligent machines, Chinese Academy of Sciences, Hefei 230031, China
luocw@mail.ustc.edu.cn, {harryjun, zfwang}@ustc.edu.cn

Abstract. We describe a system for performance-based facial animation, the system enables any user to control the facial expressions of a digital avatar by performing facial actions in front of a video camera. Firstly, a muscle-based face model is created and muscle actuation parameters are used to animate the face model. Next, a real-time facial tracking algorithm that incorporates geometric priors is employed to track facial features of a performer in the video. Finally, tracking results are converted to muscle actuation parameters to drive the face model. Experimental results show that the synthesized facial animations are quite realistic. Compared with most existing performance-based animation systems, ours only requires a video camera for performance capture, which makes the system easy to use for ordinary users.

Intuitive control over three-dimensional facial animations is an important problem in human computer interaction and virtual reality. There have been numerous algorithms for performance based facial animation. However, most of them are not easy to use for ordinary users. As it is inconvenient for performers to wear facial markers, 3D scanners are usually unavailable, some of the algorithms require complex pre-processing. Our goal is to develop a system that is computational inexpensive and easy to use from the user point of view.

Our face model is built from video images, the surface of the skin is represented by a mesh. 15 facial muscles are placed underneath the skin surface. These muscles are classified into linear and sphincter muscles. For linear muscles, we use the model proposed by Waters. Orbicularis oris is a sphincter muscle. We divide it into two parts: inner part and outer part. The inner part simulates those fibers proper to the lip. The outer part simulates those fibers derived from other facial muscles. The jaw rotates around an axis which connects the two ends of the mandible, this rotation is weighted. We develop a weighting function so that the vertices both in the lower lip and upper lip deform properly.

To robustly track facial features of a performer in the video, we first learn the statistical model of face shape. Then we use this model to guide the tracking. Our tracking method is as follow. A user specifies the initial positions of the N feature points in the first frame. Then, each feature point is tracked using KLT tracker. The initial tracking results of the N feature points from KLT tracker are not very reliable because of the drifting problem, we use the statistical shape model to refine the initial tracking results. The procedure is solving for the rigid pose and the linear weights by minimizing the difference between the statistical shape model and the initial tracking results. We use a coordinate-descent method to solve the problem. We left one parameter vary at a time by fixing the other to its current guesses. This transforms the nonlinear problem into a linear one. The tracking results are converted to muscle actuation parameters using a procedure similar to that described in [1].

The Experiments conducted on 8 videos show that our tracking method is able to track facial motions of different human subjects. We compared our tracking method with KLT and ASM. The performance is measured by the normalized point-to-point error. The comparison results for one of the videos are show in Figure 1(a). It is shown that tracking errors accumulate quickly for KLT. The tracking errors of our method are comparable to that of ASM. By incorporating geometric priors, the drifts of KLT are successfully removed.

We implemented the animation system on a PC platform and test the system on different users. Figure 1(b) shows several snapshots taken from original single-camera video and retargeted results, the synthetic animations are realistic and quite similar to the facial actions in the video. Compared with the animation system described in [2], ours only requires a video camera and can be directly used without pre-processing.

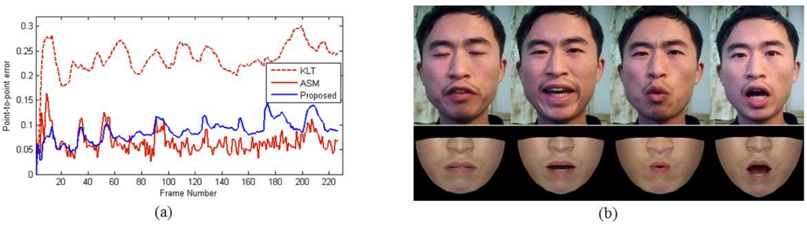


Fig. 1. (a)The normalized point-to-point errors as a function of frame number. (b)Example frames from a video sequence retargeted onto our 3D face model.

Acknowledgements. This work was supported by the Fundamental Research Funds for the Central Universities of China (No. WK2100100009), and the China Postdoctoral Science Foundation under (No. 2012M521248).

References

1. Choe, B., Lee, H., Ko, H.S.: Performance-driven muscle-based facial animation. The Journal of Visualization and Computer Animation 12, 67–79 (2001)
2. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation. In: Proceedings SIGGRAPH (2011)

The INGREDIBLE Database: A First Step Toward Dynamic Coupling in Human-Virtual Agent Body Interaction

Igor Stanković[1], Pierre De Loor[1], Virginie Demulier[2], Alexis Nédélec[1], and Elisabetta Bevacqua[1]

[1] UEB, Lab-STICC, ENIB
[2] LIMSI-CNRS, France
{stankovic,deloor,nedelec,bevacqua}@enib.fr, demulier@limsi.fr

1 Introduction

This paper describes a human-avatar interaction database, built from the recorded interactions of two actresses wearing motion capture (MoCap) suits. They were located in separate rooms, and communicated only via each other's avatar. This form of interaction is noteworthy for two reasons: firstly, due to the use of avatars, no facial expressions nor gaze were visible, ensuring that the only interaction cues were body movements; secondly, from a psychological point of view, this database can tell us how people communicate with virtual representations of other individuals.

To our knowledge, none of the existing multi-modal expressive behaviour databases meet our requirements. Most databases, such as [1], do not focus on interaction or have been collected for different reasons. Other databases, for example [2], contain recordings of human-human interactions and focus mainly on verbal communication limited to the upper part of the body (i.e. face, head, and hands).

The database presented here is part of the INGREDIBLE project, which aims to develop a virtual agent capable of maintaining a gestural coupling [3] with a human. The assumption is that the increased persistence of coupling – with some evolving rules yet to be defined – will heighten the realism of a virtual character's presence [4] and benefit their use in applications such as personal assistants and intelligent virtual tutors, and provide more realistic interactive behaviours in video games.

2 Technical Description and Collected Data

MoCap recordings for the database were carried out with the aid of two professional actresses from the theatre company Derezo[1]. The actresses were located in different rooms, and only able to see each other's avatars. In order to make the

[1] http://www.derezo.com

R. Aylett et al. (Eds.): IVA 2013, LNAI 8108, pp. 430–431, 2013.

INGREDIBLE database more widely usable, the database utilizes two different MoCap suits and systems to collect data: Art-Human and Moven.

We have two main reasons for requiring a MoCap database: firstly, the database will be used to develop feature analysis tools able to recognize users' gestures; secondly, MoCap recordings are necessary for the animation of a virtual agent. We also recorded synchronised videos of the two actresses while interacting in order to annotate their movements and find cues of dynamic coupling.

There are several types of recordings in the database. Recordings were either non-interactive or interactive. *Non-interactive* means that the actresses did not interact, and so no avatar was displayed in front of them. Instead, they were asked to perform a series of predefined gestures, repeating each one with variations in three dimensions. The modification criteria were amplitude (narrow, medium, wide), speed (slow, medium, fast), and fluidity (staccato, medium, fluid). In the *interactive* approach, the actresses communicated with each other through their human-size avatars, which were displayed on a screen in front of them. They were introduced to this environment by being encouraged to interact freely for as long as they wished. These first recordings often provided us with very interesting and spontaneous data, but the actresses rapidly grew bored without a prescribed task to perform. To add artistic, gestural, and expressive details to the interactions (according to the requirements of the project), we defined two interacting situations: 1.) imitation and 2.) bodily emotional dialogue.

The resulting dataset consists of 114 different recordings, 57 captured by each suit, with more than 150 hours of recording and 27 GB of data. The database stores recorded Art-Human and Moven data converted to the *.bvh* format. It also holds Art-Human raw data in *.txt* files and Moven *.mvn* data.

3 Future Work

The database contains some limitations (e.g. only two participants, both participants are female, both participants are actresses), so as a part of our future work, more recordings would be necessary. At present, the videos are being annotated by a team of psychologists aiming to extract cues of dynamic coupling.

Acknowledgement. This work has been funded by the ANR INGREDIBLE project: ANR-12-CORD-001 (`http://www.ingredible.fr`).

References

1. Rett, J., Faria, D., Neves, A., Simplicio, C.: HID - Human Interaction Database (2007), `http://paloma.isr.uc.pt/hid`
2. Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP database. Journal of Language Resources and Evaluation 42(4), 335–359 (2008)
3. Warren, W.: The dynamics of perception and action. Psychological Review 113, 358–389 (2006)
4. Forgas, J.P.: Handbook of affect and social cognition. Erlbaum, New Jersey (2001)

OpenBMLParser: An Open Source BML Parser/Analyzer

Herwin van Welbergen[1,2,⋆]

[1] Sociable Agents Group, CITEC, Fac. of Technology, Bielefeld University
[2] Human Media Interaction, University of Twente

1 Introduction

The Behavior Markup Language (BML) [1] has become the de facto standard for multimodal behavior specification for virtual agents. A BML block describes the behaviors (e.g. gaze, speech, gesture) a virtual agents should realize, the synchronization constraints between them, and how the block is to be composed in the ongoing behavior plan. Besides the synchronization constraints that are explicitly defined in BML, several implicit constraints act upon a BML block [2]. Parsing BML is not a trivial endeavor, since the language is highly redundant and extensions can be hooked up at several points. Properties that are useful for both BML scheduling and queuing can be obtained by analysis of a BML block. Our open source Java BML parser/analyzer OpenBMLParser provides authors of new and existing BML Realizers with a building block that can handle BML parsing and BML block analysis and allows them to easily hook up their own BML extensions.

2 Parsing, Analyzing and Extending BML

To schedule a BML block, the realizer needs a list of behaviors in it, the set of its 'at' constraints and the set of its 'after' [1] constraints[2]. After parsing a BML block, OpenBMLParser provides exactly that information. We have also implemented the BML cluster property predicates defined in [2]. These predicates can be used by the realizer to resolve or maintain the block's implicit constraints. Both BML behaviors and constraints may be wrapped inside `required` elements, to indicate that the BML block should fail as a whole if required behaviors or constraints fail. OpenBMLParser captures this information by listing the requiredness for each behavior and constraint.

The BML specification requires that BML blocks are scheduled in the order they are submitted to a realizer. BML blocks that are waiting to be scheduled

⋆ An early version of what is now OpenBMLParser was implemented by Ronald C. Paul. This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster Competition and managed by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the contents of this publication.

[1] Before constraints can be rephrased as after constraints.

form a scheduling queue. More reactive behavior realization can be achieved by allowing a realizer to schedule multiple blocks in parallel. The scheduling of a block may start as soon as no dependent blocks are before it in the scheduling queue. OpenBMLParser supports such multi-threaded scheduling by providing the realizer with the dependency information of a BML block. A BML block is dependent on another block if 1) its constraints refer to that block 2) it has behaviors referring to that block or 3) the `bml` element of the block defines dependencies on other blocks.

OpenBMLParser allows users to register their own classes that parse behaviors or description levels, and allows the registration of custom attributes on existing behaviors. It also provides users with functionality to register one or more handlers for the parsing of their own custom attributes and attributes values within the `bml` element itself, e.g. to parse custom block composition values or to specify that a block should be preplanned for later execution.

3 Using OpenBMLParser

OpenBMLParser[2] can help the authors of new Java[3] realizers, realizers that do not support BML 1.0 yet, or other applications that make use of BML to focus on the development of new scientific or business value, rather than on the tedious parsing and analysis of BML blocks. OpenBMLParser is fully BML 1.0 compliant: it parsers and analyzes all BML 1.0 Core behaviors, Core Extensions, constraint definitions and BML 1.0 feedback. It is currently used as a BML parser/analyzer in AsapRealizer [3], allowing its extensions for BML block composition and preplanning, and the parsing of its over 20 custom behaviors and description levels; and as a BML feedback parser in RealizerTester [4] –an integration testing tool employed by multiple realizers.

References

1. Kopp, S., Krenn, B., Marsella, S.C., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
2. van Welbergen, H., Reidsma, D., Zwiers, J.: Multimodal plan representation for adaptable BML scheduling. In: Autonomous Agents and Multi-Agent Systems, pp. 1–23 (2013)
3. van Welbergen, H., Reidsma, D., Kopp, S.: An incremental multimodal realizer for behavior co-articulation and coordination. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 175–188. Springer, Heidelberg (2012)
4. van Welbergen, H., Xu, Y., Thiebaux, M., Feng, W.-W., Fu, J., Reidsma, D., Shapiro, A.: Demonstrating and testing the BML compliance of BML realizers. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 269–281. Springer, Heidelberg (2011)

[2] Available at `https://github.com/saiba/OpenBMLParser`
[3] Or any other JVM-based language.

A Companion Robot That Can Tell Stories

Carole Adam[1] and Lawrence Cavedon[2]

[1] Grenoble Informatics Laboratory - Joseph Fourier University, Grenoble, France
carole.adam@imag.fr
[2] RMIT University, Melbourne, Australia
lawrence.cavedon@rmit.edu.au

Keywords: Interactive Storytelling, artificial companion, personalisation, engagement, believability.

Introduction. Telling engaging stories is an interesting ability for an artificial companion for children. Three features can be made engaging: the **agent** itself (which is addressed by the field of Embodied Conversational Agents); the **content** of the story (*e.g.* interactive stories [1]); and the way the story is **narrated**. We designed and implemented engaging narrative strategies for Reeti, an affective expressive robot with a wide range of emotional facial expressions.

Strategies. We used: a corpus of parent-child interactions [2] showing a tendency to use personalisation to engage children; guidebooks for human storytellers [3] highlighting the importance of interactivity and tailoring the story to the audience; and literature in HCI [4] and relational agents [5] revealing specific requirements such as customisation, interactivity, and user control. We thus identified the following engaging narrative strategies for Reeti:

- Embody the different characters by changing **voice**, for livelier narration;
- Adapt **vocabulary** to the child's age, use simpler synonyms or definitions;
- Show emotional intelligence: **express** emotions consistent with the story, and **detect** and react to the child's emotions triggered by the story;
- Make **random** changes in the text of the story to avoid boredom;
- Make **personal** comments relating story to child's profile and context;
- Offer to play **interactive** games to favour engagement (quiz, guess...):
- Offer multiple **choices** at some points of the story to give a feeling of agency;
- Insert relevant **diversions** (jokes, anecdotes...) to prevent boredom;
- Refrain from interrupting the story to **focus** on key moments (immersion);

SMILE language. To perform these, the storyteller needs two types of information: about the user profile and context (already available to companions); and about the story, triggers for strategies and additional scripted content (provided as story annotations with our SMILE language [6]). For example the annotated snippet below tells the storyteller that *wolf* is an emotional word, and provides scripted comments to react to two emotions (as deduced from user profile).

> *When Little Red Cap arrived in the woods, she met the* <emoword>*wolf* <comm emo="fear">*Lucky there are no wolves around here right?*</comm> <comm emo="excitement">*You like scary animals don't you?*</comm> </emoword>. *But she did not know it was a nasty animal and was not afraid.*

Implementation. We implemented several modules in Java: a **SMILE parser** for the annotated stories; a **GUI** using Google speech recognition and/or text input to let the user interact with the robot during the narration; and a basic **storytelling engine for Reeti**, with only the "change of voice" strategy so far.

Pilot studies. To inform the implementation of our storytelling module in the Reeti expressive communicating robot, we conduced two pilot studies.

We **first** had 22 visitors at the Innorobo robotic show (Lyon, France, March 2013) play a game with a robot and rate the acceptability of robots in different hypothetical roles with a child. Users explicitly stated that they could accept a robot only as a complement but not as a substitute; they would not trust the robot with responsibilities; and they were reluctant to letting it create a relationship with their child. The physical appearance of the robot was found to influence its perceived credibility in its role (*e.g.* too small to have authority). The storyteller and playing buddy roles were both considered as very acceptable.

We **later** had 25 students and staff at Grenoble Informatics Laboratory rate our list of strategies on two criteria: believability (likeliness that a human storyteller would use it) and engagingness (efficiency to captivate a child). The users insisted on the importance of **interactivity**, in particular the storyteller's ability to understand and answer the child's questions, but also to itself ask questions about the child's opinions and feelings. They found most strategies engaging, even when not human-like, except for changing the story (undesirable to modify the author's work) and forcing focus (harsh to not let the user in control).

Conclusion. The aim of our approach is to make it possible for an artificial companion to use strategies to modify a story (or another text), in order to really **personalise** its narration, not to a group or category of users, but to one specific user that it gets to know over time. More details can be found in [7].

References

1. Cavazza, M., Donikian, S. (eds.): ICVS 2007. LNCS, vol. 4871. Springer, Heidelberg (2007)
2. Adam, C., Cavedon, L., Padgham, L.: "Hello Emily" - Personalised dialogue in a toy to engage children. In: Companionable Dialogue Systems, ACL (2010)
3. Hostmeyer, P., Kinsella, M.A.: Storytelling & QAR Strategies. Libr. Ultd. (2011)
4. Brandtzaeg, P.B., Folstad, A., Heim, J.: Enjoyment: Lessons from karasek. In: Funology - From Usability to Enjoyment. HCI, vol. 3, pp. 55–65. Springer (2005)
5. Bickmore, T.W., Picard, R.W.: Establishing and maintaining longterm human-computer relationships. ACM Trans. on Comp.-Human Interactions 12(2) (2005)
6. Adam, C.: Il était une fois... un robot compagnon qui racontait des histoires. In: WACAI. LIG research reports, vol. RR-LIG-039. LIG (2013)
7. Adam, C., Cavedon, L.: Once upon a time... a companion robot than can tell stories. Technical Report RR-LIG-??, LIG, Grenoble, France (2013)

Towards Mapping and Segmentation of Very Large Scale Spaces with Intelligent Virtual Agents*

Katrina Samperi, Nick Hawes, and Russell Beale

School of Computer Science, The University of Birmingham, UK
{k.samperi,n.a.hawes,r.beale}@cs.bham.ac.uk

1 Introduction

Intelligent virtual agents are increasingly required to generate and use maps in very large scale environments. One such environment is the virtual world Second Life [1]. Second Life is an ever growing large scale virtual environment where users can interact with, create objects and change the environment in a variety of ways. Creating a virtual agent that is able to build, maintain and use a map of an environment of this scale is a challenging problem.

In very large scale environments it can become necessary to segment the map used by the agent to improve performance. A single monolithic map will often be too large to fit in the memory the agent has available. Attempting to update or use this map for path planning will take more time as the map size increases. We want to segment this map intelligently to reduce the time required for the agent to plan a path between any two points in the environment.

Our hypothesis was that we could improve the segmentation of the map by using trails. Trails are a set of observations of how other avatars, human and AI controlled, move around an environment. It has been shown in previous work that the movement of other avatars can provide information about the structure of the environment [2] and that trail information can be used to help improve the generation of roadmaps in these types of environments [3]. This paper describes our preliminary findings on using this approach for segmentation.

2 Approach

To investigate our hypothesis we compared a single monolithic map with three different segmentation methods. The environment used was a combination of four regions in Second Life, a space 512m^2 in size. To evaluate we compared the time required to finish the segmentation, the planning time between two given points, and the length of the planned path. A good segmentation method would take a short period of time and allow for fast path planning. As segmentation

* The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement No. 600623, STRANDS.

restricts the options available for path planning, a route based on a segment map will usually be longer than one planned using a single monolithic map. In these cases the route planned should still be as short as possible.

The four different segmentation methods compared were:

– No segmentation - the base case
– Regular segmentation at various resolutions
– Quadtree segmentation [5] at various resolutions
– Voronoi segmentation [4] using different seed points

Trails were used in conjunction with Voronoi Segmentation. We were able to identify cluster points in the trails, the places where many avatars gather together, and use these for seed points in the algorithm.

3 Results and Future Work

Using no segmentation the total time required to plan the route was 631.01s. The length of the planned route was 1299.11m and the success rate 95%.

We found that the fastest method for segmenting the map was to use regular segmentation. However, routes planned using these maps were nearly 50% longer than with no segmentation. Trail based segmentation generated the map and planned a route quicker than using no segmentation at all, taking 59.64s. These maps planned shorter routes than maps generated using regular segmentation, but the success rate was reduced to 70%. Quadtree segmentation took a long time to complete but, on average, planned the shortest routes.

These results are promising, but not conclusive as to whether trail based segmentation is better for dividing a large scale environment than other methods.

We will to investigate our results further and find out if trails become more useful as the size of the environment increases. Trail based segmentation may be especially useful in dynamic environments, as the environment structure being taken into account may lead to a reduction in the number of segments that need updating as the world changes.

References

1. Linden Research Inc. Second life official site (2012), `http://secondlife.com`
2. Samperi, K., Beale, R., Hawes, N.: Please keep off the grass: individual norms in virtual worlds. In: Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers, BCS-HCI 2012, pp. 375–380. British Computer Society (September 2012)
3. Samperi, K., Hawes, N., Beale, R.: Improving map generation in large-scale environments for intelligent virtual agents. In: The AAMAS 2013 Workshop on Cognitive Agents for Virtual Environments. LNCS. Springer (2013)
4. Thrun, S.: Learning metric-topological maps for indoor mobile robot navigation. Artif. Intell. 99(1), 21–71 (1998)
5. Wurm, K., Hennes, D., Holz, D., Rusu, R., Stachniss, C., Konolige, K., Burgard, W.: Hierarchies of octrees for efficient 3d mapping. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS. IEEE (2011)

A Character Agent System for Promoting Service-Minded Communication

Kaoru Sumi and Ryuji Ebata

Future University Hakodate
041-8655 Hakodate, Japan
kaoru.sumi@acm.org

Abstract. This paper describes a dialogue system for learning a Japanese style of service-mindedness. The system provides educational training for users in a customer service. It consists of a facial expression recognition system using brain wave measurement equipment, a speech recognition system, a speech synthesis system, and a dialogue control system.

Keywords: A character agent system, Service-Minded Communication.

1 Character Agent System for Service-Minded Communication

This paper introduces a system for providing educational training in hospitality through dialogue with a character agent [1]. This system focuses on the Japanese style of service-mindedness, which is typified by paying attention to individual customers. “Service-mindedness” means a technique or mentality for promoting customer satisfaction. Smiling is one such important technique for serving customers [2]. Our group has researched persuasive technology in which a technical artifact persuades a user to do something. We evaluated users’ impressions of a character agent by setting up an emotion-arousing scenario and observing how the users reacted to various patterns of character agent reactions [3][4][5]. Based on these research results, we have developed the system for facial expression training.

Our system displays a 3D model character agent, and it can perform facial expression recognition by using brain wave measurement equipment and conduct dialogue via speech recognition and speech synthesis systems and a dialogue control system (Figure 1). The system shows a 3D character agent that can exhibit the following facial expressions: "smiling," "laughing," "angry," "sad," "disgusted," "frightened," and "surprised,". These facial expressions can be digitally controlled in intensity. The agent mouths the shapes of vowels in each facial expression to achieve a lip-synching effect. Morphing technology enables fluid movement of the facial expressions and lip-synching.

For the facial expression recognition system, we use the facial recognition application of a brain wave measurement device. This equipment digitally recognizes the intensity of facial expressions.

Fig. 1. Usage example

By using our system, the user can talk to the character agent and learn from the character agent how to interact with customers. The character agent appears on the screen, and sometimes the system displays lines in the scene, which the user should practice with an appropriate facial expression. Within the overall system, the speech recognition system judges whether the user's speaking of the lines is appropriate by comparison with dialogue templates. Similarly, the facial expression recognition system judges whether the user's facial expression is appropriate.

2 Discussion

We conducted an experiment to examine the effectiveness of a system designed for learning about service-mindedness, which is found in Japanese hospitality and is exemplified by consideration for the customer. The system operates through dialogues between a user and a character agent. We think this system should facilitate the development of service-mindedness without requiring an instructor.

References

1. Sumi, K.: Human Agent Interaction for Learning Service-Minded Communication. In: 1st International Conference on Human Agent Interaction, iHAI 2013 (to appear, 2013)
2. Morishita, H.: Shinri Sekkyaku Jutsu (Cognitive Service for Customers), Soshimu (2009) (in Japanese)
3. Sumi, K., Nagata, M.: Evaluating a Virtual Agent as Persuasive Technology. In: Csapó, J., Magyar, A. (eds.) Psychology of Persuasion. Nova Science Publishers (2010)
4. Sumi, K.: Human Interface of Robots or Agents via Facial and Word Expression. In: International Symposium on Artificial Life and Robotics (AROB), Invited Talk (2012)
5. Sumi, K., Nagata, M.: Characteristics of Robots and Virtual Agents as a Persuasive Talker. In: Stephanidis, C., Antona, M. (eds.) UAHCI/HCII 2013, Part II. LNCS, vol. 8010, pp. 414–423. Springer, Heidelberg (2013)

A Speech-Enabled Intelligent Agent for Pedestrian Navigation and Tourist Information

Srinivasan Janarthanam[1], Oliver Lemon[1], Phil Bartie[2], Tiphaine Dalmas[2], Xingkun Liu[1], William Mackaness[2], and Bonnie Webber[2]

[1] Heriot-Watt University
[2] University of Edinburgh

Abstract. We present an intelligent virtual agent that assists pedestrian users in navigating within urban environments and acquiring tourist information by combining spoken dialogue system, question-answering (QA), and geographic information system (GIS) technologies. In this paper, we present the architecture and features of our latest system, extended from an earlier version which was built and evaluated with real users.

Keywords: mobile, speech, dialogue, geographical, visual, question-answering, GIS.

1 Introduction

We present an interactive intelligent virtual agent (an Android application) that addresses the problems of pedestrian navigation and tourist information in urban environments. There has been little prior work that addresses these two problems - navigation and tourist information provision - in an integrated way. An agent such as this could serve as a personal tour guide to pedestrian tourists as they walk around unknown cities. With the proliferation of smartphones, there has been a number of mobile applications developed to address these problems. However these applications have the following problems: first, they demand the user's visual attention because they predominantly present information on a mobile screen. This can be dangerous in urban environments, as well as being distracting. Second, these applications address the problems of navigation and tourist information independently and therefore do not have a shared interaction context. This means that users cannot switch between the tasks in a natural and fluid manner.

2 Architecture

The architecture of the current system is shown in figure 1. Our architecture brings together Spoken Dialogue Systems (SDS), Geographic Information Systems (GIS), and Question-Answering (QA) technologies. Its core is a spoken dialogue system (SDS) consisting of an automatic speech recogniser (ASR), a semantic parser, an Interaction Manager, an utterance generator and a text-to-speech synthesizer (TTS). Users' spoken utterances are recognised and parsed into dialogue acts and are sent to the interaction manager (IM) along with GIS information. The IM provides three services: navigational instructions, pushing point of interest (PoI) information and managing user questions. It manages the conversation using five threads: dialogue control, respond to user queries, and one each for the three services. Some of key features include interleaving conversational threads, the use of landmarks for navigation, use of social media data for identifying landmarks and PoIs, and user interest modelling.

The GIS modules in this architecture are the City Model that consists of information concerning the location, names and types of entities, the Visibility Engine that models which entities are visible to the user from a given location and orientation, and the Pedestrian tracker that improves the user location information from raw GPS coordinates. Users communicate with the system using a smartphone-based client app (an Android app) that sends the user's position, pace rate, and spoken utterances to the system, and delivers synthesised system utterances to the user. An earlier version of the system is described and evaluated in [1].

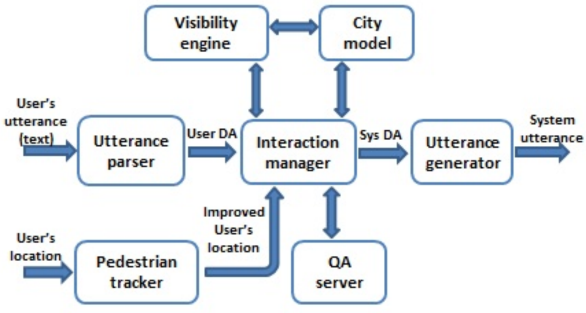


Fig. 1. System Architecture

Acknowledgments. The research leading to these results was funded by the European Commission's Framework 7 programme under grant agreement no. 270019 (SPACEBOOK project).

Reference

1. Janarthanam, S., Lemon, O., Bartie, P., Dalmas, T., Dickinson, A., Liu, X., Mackaness, W., Webber, B.: Evaluating a city exploration dialogue system combining question-answering and pedestrian navigation. In: Proc. ACL 2013 (2013)

A Question-Answering Game for Preschool Children

Anuj Tewari[1], Peiyi Ko[2], and John Canny[1]

[1] Electrical Engineering and Computer Science
[2] School of Public Health
University of California,
Berkeley
{anuj,jfc}@cs.berkeley.edu, peiyiko@berkeley.edu

Abstract. The preschool literacy gap is one of the most difficult challenges for education in the US. Children in the lowest SES (Socio-Economic Status) quartile have less than half the working vocabulary of those in the top quartile at age 3. On the other hand, preschool children are incessantly inquisitive, and will readily engage in question answering and asking activities if given the opportunity. We argue here that question asking/answering technologies can play a major role in early literacy. We describe the early evaluation of a conversational agent, with the goal of engaging children in a 20-questions game. Towards this goal, we conducted a feasibility study to determine if childrens questions are on-topic and suitable for ASR/dialogue systems. We evaluated the agents performance at conducting a game of 20-questions against that of a human partner.

1 Introduction

A large body of research has shown that the literacy gap between low and high socio-economic status children is well-established before formal schooling begins, that it is enormous, and that it predicts academic performance throughout primary, middle and secondary school. Indeed rather than closing this gap, there is much evidence that formal schooling exacerbates it: once behind in reading and vocabulary, children read with lower comprehension, learn more slowly and have lower motivation than their more language-able peers. The greatest impact on child literacy will come from intervention at pre-school ages. This explains the need for expert interactive systems that can work as engaging question-answering agents.

2 Envisaged Solution

We envision a projection system with a virtual character that acts as the question-answering agent. The aim is to create a virtual play space that can keep the conversation grounded in a context. The virtual character (usually an animal), rendered over a wide area using a projection system, will engage children in language games that use question-answering as the primary dialogue structure. For example, the character will show the child several objects, then hide one and ask the child to guess what it is by asking questions about it.

3 Current Work

We conducted a two-phase study, one phase using a human language partner, and the second using an agent. Rather than relying on speech recognition and dialog interpretation, we used a Wizard-Of-Oz system. The goal of the studies was to explore the feasibility of the envisaged solution: whether students would ask on-topic questions, whether the questions matched some templates, and whether they would be engaged by the game. Phase 1 involved 20 children studying at the same preschool, playing a 20-questions game with a familiar researcher. Phase 2 involved the same participants as phase 1. Half of them played the game with the same researcher. The other half played the same with an agent that we designed and implemented.

3.1 Results: Phase I

Participants asked a total of 210 questions, which means an average of 10.5 questions per child in a span of 20 minutes. Analysis of the coded data revealed that 80% of the questions asked by our participants were about the part, property or function of objects. The remaining 20% were guesses about what the object could be, example ”Is it a cat?” Participants in our study needed limited explanation. On average researchers had to intervene only 1.67 times per child. Children were adept at inferring the objects. In phase 1, they were able to successfully solve 104 out of the 126 (80%) trials conducted. During the study, sometimes a child would talk about objects or contexts that were not grounded in the environment. It was noticed that approximately 5% of the utterances by the participants were not grounded in the environment.

3.2 Results: Phase II

Across 7 trials in the session, children in the human condition asked 78 questions and this number was 123 for the agent condition. A two-tailed t-test for the total number of questions asked, pointed towards statistical significance (p-value = 0.03, Cohens d = 0.94). Handling redirection and off-topic conversations is an important characteristic of any question-answering system. Interactions with the agent, just like interactions in phase 1 did not deviate from the topic much. Children generally stayed focused and there was no significant difference in the number of off-topic dialogues in the human and agent condition ($p = 0.813$). It is clear that children follow a similar pattern as phase 1 in talking to the agent, and there was a limited need for explanation, hints and off-topic dialogue handling.

4 Conclusion

Despite the hype around how conversational systems can make information more accessible, more rigorous research is clearly called for. This paper presents an important step towards this goal. We have identified need and opportunities for question-answering based conversational games in the everyday lives of preschool children. More detailed description of the agent and automated evaluation will be published as future work.

A Software Framework for Social Cue-Based Interaction with a Virtual Recruiter

Ionut Damian[1], Tobias Baur[1], Patrick Gebhard[2],
Kaśka Porayska-Pomsta[3], and Elisabeth André[1]

[1] Human Centered Multimedia, Augsburg University
{damian,baur,andre}@hcm-lab.de
[2] DFKI GmbH, Saarbrücken, Germany
patrick.gebhard@dfki.de
[3] London Knowledge Lab, Institute of Education
K.Porayska-Pomsta@ioe.ac.uk

1 Introduction and Motivation

In this paper we present a software framework which can be used to generate job interview simulations using a virtual recruiter and social cue-based interaction. We use the term social cues to describe conscious or unconscious behavioural patterns which have a specific meaning in a social context. The main goal of this endeavour is to help youngsters improve social skills pertinent to job interviews.

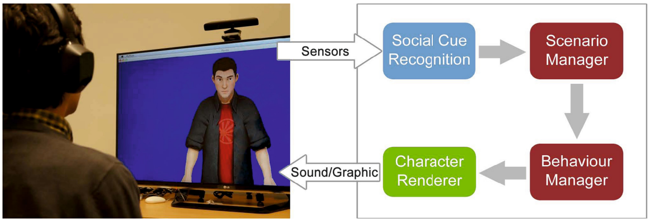


Fig. 1. General setup and main software modules of the system

2 The System

We propose a job interview simulation environment featuring a social virtual character in the role of the recruiter and using signal processing techniques to generate seamless interaction between the user and the character. The system consists of three major modules: a behaviour manager, a social cue recognition module and a scenario manager (Fig. 1).

At the backbone of the system lies the EMBOTS behaviour manager [1]. It supports fine grained multimodal behaviour control for virtual characters and offers various functions which are needed in an interactive character system (e.g. TTS, Character Rendering, Emotion Simulation).

The social cue recognition module is based on the SSI framework[1] [2]. The main strength of SSI lies in its ability to record and process human behaviour data and

[1] http://openssi.net

social signals in real time. In particular, SSI supports the parallel and synchronized processing of data from multiple sensor devices, such as web/dv cameras, multichannel microphones and various physiological sensors. For our system, we rely on the Microsoft Kinect[2] sensor. Its main advantages are its low cost, its robustness towards environmental noise, such as lighting and background, and the availability of various software toolkits which provide skeleton and face tracking. Using SSI and the Microsoft Kinect, we implemented several social cue recognizers which are able to automatically detect various behaviours, such as head gaze, postures, gestures and voice activity, in real time [3]. The data of each interaction is also recorded and can be used to derbief the user during a post-hoc analysis of the interview simulation using the NovA behaviour analysis tool[3].

The third module we use in our system is the SceneMaker scenario manager [4]. It allows us to model and to execute behavioural aspects at different levels of abstraction. Using SceneMaker we modelled a prototypical job interview scenario consisting of various dialogue scenes as well as reactions to different behaviours recognized by the social cue recognizers.

3 Conclusion

We presented an approach to a job interview simulation environment featuring a virtual recruiter and social cue based interaction. The virtual character's behaviour is controlled by the EMBOTS behaviour manager [1] and the interaction is facilitated by several social cue recognizers using the SSI framework [2] and a Microsoft Kinect. The recognizers are able to automatically analyse the behaviour of the user in real time. Preliminary tests suggest that the system generates credible job interview simulations giving users a realistic experience.

Acknowledgement. This work was partially funded by the European Commission within FP7-ICT-2011-7 (Project TARDIS, grant agreement no. 288578).

References

1. Heloir, A., Kipp, M.: Real-time animation of interactive agents: Specification and realization. Applied Artificial Intelligence 24(6), 510–529 (2010)
2. Wagner, J., Lingenfelser, F., André, E.: The social signal interpretation framework (SSI) for real time signal processing and recognition. In: Proc. Interspeech 2011 (2011)
3. Damian, I., Baur, T., André, E.: Investigating social cue-based interaction in digital learning games. In: Proceedings of the 8th International Conference on the Foundations of Digital Games, SASDG (2013)
4. Gebhard, P., Mehlmann, G., Kipp, M.: Visual scenemaker - a tool for authoring interactive virtual characters. Journal on Multimodal User Interfaces 6(1-2), 3–11 (2012)

[2] http://www.microsoft.com/en-us/kinectforwindows
[3] http://openssi.net/nova

Size Certainly Matters – At Least If You Are a Gesticulating Digital Character: The Impact of Gesture Amplitude on Addressees' Information Uptake

Mette Clausen-Bruun, Tobias Ek, and Magnus Haake

Cognitive Science, Department of Philisophy, Lund University
Mette.clausen-bruun@comhem.se, tobias.ek@telia,se,
magnus.haake@lucs.lu.se

In order to create trust and rapport, one goal in multi-modal virtual character research is to determine how to vary expressive qualities of the characters so that the addressee will perceive them in the desired way. This pilot study investigates the impact of gesture amplitude on addressees' information uptake. As digital characters used in pre- and early school years' learning games tend to be less human-like and more fanciful, we measured participants viewing a humanoid or an 'alien' character, respectively.

Speakers' gestures have been shown to contribute to addressees' semantic and pragmatic comprehension and memory [1, 2, 3], although eye-tracking studies have found that only a minority of gestures draw fixations (defined as instances where the fixation marker remains for at least 120 ms directly on a fixated object) [4]. Nevertheless, it has been shown that gestures that are incongruent with the semantic meaning of the spoken word have hindered addressees' comprehension [5]. Although there is little evidence of a correlation between addressees' direct fixations of gestures and their uptake of gesture information [6], gesture amplitude has been shown to be a key indicator of extraversion, a personality trait that is often considered advantageous for people conveying a message [7].

In our study, we randomized a convenience sample of 120 undergraduate university students to four groups: (i) 'normal gesture amplitude, alien character'; (ii) 'high gesture amplitude, alien character'; (iii) 'normal gesture amplitude, humanoid character', and, (iv) 'high gesture amplitude, humanoid character'.

The characters were created by means of motion capture data from a woman sitting on a chair, retelling two children's narratives presented in animated cartoons. The data was used to animate a simplified cartoonish Motionbuilder character ('*alien'*) as well as a character resembling a female human being (*humanoid*).

We opted to create a high gesture amplitude condition by manually extending and fine-tuning the movements based on a case-by-case assessment of each gesture at their initial phase as well as in their striking phase. Doing this algorithmically would risk jeopardizing naturalness of movement and synchronization with speech. Four videos for the above mentioned groups were created, each with a duration of two minutes and 20 seconds, using the soundtrack from the original recording.

As hardware tends to get increasingly smaller, the survey was conducted on 15-17" laptops. Each participant was introduced to one of the four video conditions, and then asked to fill in 22 multiple-choice statements about the story and a questionnaire with subjective ratings of the characters.

Participants rated the humanoid significantly more natural and less distracting than the 'alien' and tended to perceive the gestures of the humanoid more facilitating on information recall than those of the 'alien'. As one reviewer points out, one possible reason for this could be the limited face and finger motions in the characters. The face and fingers of the 'alien' are larger and therefore more salient than those of the humanoid and the lack of motion could therefore be more disturbing.

We also found that size actually does matter, although significance was modest. This may be attributed to the 'normal gesture amplitude' having a rather high baseline. During a motion capture recording one may become more conscious about one's gestures and tend to either restrict or exaggerate them. Future studies should add a 'minimized gesture amplitude' condition to control for this possible confound and, to eliminate yet another possible confound, complexity of the narratives should be adapted to the age of the participants. Furthermore, it would be of value to designers of teaching games to investigate whether gesture amplitude has the same effect on deep, vs. shallow learning, that is contents vs. form. Also, further research is needed in respect of gesture amplitude and screen size.

Interdisciplinary studies may contribute to the establishment of proper criteria and guidelines for developers to follow and there is definitely a length of parameters to examine across screen sizes – as well as across cultures!

References

1. Hostetter, A.B.: When do gestures communicate? A meta-analysis. Psychological Bulletin 137, 297–315 (2011)
2. Cutica, I., Bucciarelli, M.: The deep versus the shallow: Effects of co-speech gestures in learning from discourse. Cognitive Science 32, 921–935 (2008)
3. Kendon, A.: Do gestures communicate? A review. Research on Language and Social Interaction 27, 175–200 (1994)
4. Gullberg, M., Holmqvist, K.: What speakers do and what addressees look at. Visual attention to gestures in human interaction live and on video. Pragmatics & Cognition 14, 53–82 (2006)
5. Habets, B., Kita, S., Shao, Z., Özyurek, A., Hagoort, P.: The role of synchrony and ambiguity in speech-gesture integration during comprehension. Journal of Cognitive Neuroscience 23, 1845–1854 (2011)
6. Gullberg, M., Kita, S.: Attention to speech-accompanying gestures: Eye movements and information uptake. Journal of Nonverbal Behavior 33, 251–277 (2009)
7. Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 222–235. Springer, Heidelberg (2010)

A Procedural Approach to Simulate Virtual Agents Behaviors in Indoor Environments

Laura Mattos Flach, Vinícius Jurinic Cassol,
Fernando Pinho Marson, and Soraia Raupp Musse

Graduate Course in Computer Science
Pontifícia Universidade Católica do Rio Grande do Sul
Poarto Alegre, Brazil
{laura.flach, vinicius.cassol, fernando.marson}@acad.pucrs.br,
soraia.musse@pucrs.br
http://www.inf.pucrs.br/~vhlab

1 Introduction

The problem to be dealt in this paper is related to the automatic generation of indoor environments, groups and families of virtual agents which should serve as background characters in games and other computer graphics applications. Firstly, environments should be generated in a coherent way in order to be populated. Secondly, the population inside a building should be coherent to the environment. In summary, the main contributions of this paper are: i) fully automatic method to generate families, members and behaviors coherent to their environment, ii) provide persistence of generated data, considering the possibility of re-generation taking into account past status and current time, and finally iii) describe an action selection model coherent with the members of the family, past actions, time and environment.

2 The Procedural Model

Our model uses the connection between environment [1] and population as a key element in the definition of agents and their behaviors in a procedural way. Seeds are used in order to keep the persistence of information, and the model also considers the time, in order to keep behaviors coherence performing two important features in procedural models: *persistence* and *coherence*. An environmental seed is created an it is is composed by *number of rooms and bedrooms, total area, minimum and maximum coordinates*. This seed works as input for family and agents creation. In order to generate a family, we take into account he environmental seed in order to control the randomness of generation using acceptable rules as parameters. Such rules have been specified using statistics available about real cities [1] defining the percentages of families size existent until 5 members and the presence of a couple. The family seed is used to generate characteristics of family members, that are stored into the agent seeds. An agent seed is composed by:

[1] http://www.ibge.gov.br/cidadesat/link.php?uf=rs

Agent $ID(unique)$, *age*, *gender*, *role in the family*, *schedule* (preferred period of the day to be at home) as well as a *family seed*. Once environment, family, members and their attributes and status are created, we are able to select their behaviors. The fixed attributes $\boldsymbol{a}$ of family members are used to basically two functions. Firstly, age and gender are used only to determine the visualization aspects, while role and schedule are used to define which agents should be at home during different periods of the day. Once one agent should be instantiated by first time its behavior is generated based on their standard behaviors and a random process using the agent seed. Then, simulation starts by placing agents at home and initializing a FSM with first state chosen from a list of possible states. When agents are in the determined places, their attributes change as well as attributes can change according time evolution. A transition function δ is applied by detecting levels of status that are lower than a threshold. When it happens, the states are changed. In order to compute the agents motion we consider the agent's current positions and a position in the room where the action should be executed. Both positions are used for path planning and agents motion as in [2].

3 Experimental Results

Figure 1 illustrates some results obtained with our model. It is possible to observe a house populated by four people, where an agent leaving to work at morning (a), as well as the agent backing home at night (c). During this period, the agent is out of house and the other three agents keep evolving cohrerently at home(b).

(a) (b) (c)

Fig. 1. User observation during one day on a family composed by 4 people. An agent ID=1 leaves the home at 10:00 (a); situation of the house, with only 3 agents at home while agent ID=1 is working (b) and agent ID=1 evolving at home at 18:00 (c).

References

1. Marson, F., Musse, S.R.: Automatic real-time generation of floor plans based on squarified treemaps algorithm. International Journal of Computer Games Technology 1, 10 (2010)
2. Cassol, V.J., Marson, F.P., Vendramini, M., Paravisi, M., Bicho, A.L., Jung, C.R., Musse, S.R.: Simulation of autonomous agents using terrain reasoning. In: International Conf. on Computer Graphics and Imaging, Innsbruck, Austria (2011)

Animating Mixed Emotions: Can Basic Emotion Blends Be Recognized from a Virtual Face?

Meeri Mäkäräinen, Jari Kätsyri, and Tapio Takala

Department of Media Technology,
Aalto University School of Science and Technology, Espoo, Finland
{meeri.makarainen,jari.katsyri,tapio.takala}@aalto.fi

Abstract. We explore whether pairwise blends of basic emotions can be recognized from an animated face. Our results demonstrate that the mixture of joy and surprise can be expressed unambiguously by a virtual face, and several other basic emotion blends are recognizable at a more coarse classification level.

1 Introduction

Facial expressions of virtual agents are often based on the basic emotions anger, disgust, fear, joy, sadness and surprise. In real life, however, rich social interaction often requires more complex facial expressions. Several techniques have been introduced for creating complex facial expressions for virtual agents, most often by blending two basic emotions in one facial expression.

Previous research has shown that the expressions of pure basic emotions can be recognized from a virtual face, even though the recognition accuracy does not usually reach the level of natural faces. Confusions are most likely occur between fear and surprise, and between anger and disgust. So far, no studies have examined all pairwise blends of basic emotions in a similar manner. Our goal is to find out which pairwise blends can be recognized.

2 Methods

We prepared 42 videos of facial expressions, 21 with a virtual face and 21 with a natural face. They included the six basic emotions and all 15 pairwise blends. The virtual face videos were created using our animation model [1], in which blending is based on summing the muscle forces of two expressions. The natural face videos were created by morphing photographs of expressions of basic emotions.

29 volunteers evaluated the facial expressions online. Each video was rated on all six basic emotion dimensions using visual sliders. We considered the recognition of a pure emotion as being correct when the emotion with the highest rating was the targeted emotion. For the blends, recognition was considered to be correct when the two highest ratings were given to the targeted emotions. To determine statistical significance, we used Wilcoxon signed-rank tests with FDR correction. In addition, the participants were able to answer an open question "What other emotions do you see in the facial expression (if any)?"

joy+sur ang+dis ang+joy ang+sur dis+joy dis+sur fea+sad fea+sur

Fig. 1. The facial expression blends that were recognizable when the confusions between fear and surprise and between disgust and anger were allowed

3 Results and Discussion

Our virtual animation was able to successfully communicate all pure basic emotions except fear, while the natural face succeeded in communicating all pure basic emotions except fear and sadness. Most confusions occurred, as expected, between anger and disgust, and between fear and surprise.

The blend of joy and surprise was recognized from the virtual face, and the blend of fear and surprise from the natural face. We also notice that the confusions between anger and disgust and between fear and surprise cause many of the false recognitions. If we allow these confusions, the following blends become recognizable for the virtual face: anger+joy, anger+surprise, disgust+joy, disgust+surprise, fear+sadness, fear+surprise and anger+disgust. The recognizable blends are shown in Fig. 1. Similarly, six more blends become recognizable on the natural face.

In the open question, several blends of dissimilar emotions were characterized with emotion words that clearly describe mixed or ambivalent feelings. Malicious joy and fake joy were mentioned several times when joy was combined with a negative emotion. Also words like strained, baffled, uneasy, embarrassed and ambivalent were used for blends of dissimilar emotions.

Our results demonstrate that many of the blended expressions are recognizable at some level, but it is difficult to make the distinctions between anger and disgust and between fear and surprise. Moreover, we found that our facial expression blends succeeded in communicating emotional ambivalence arising from blending of two emotions.

Acknowledgments. This research was supported by the Graduate School in User Centered Information Technology, the Academy of Finland (project Enactive Media), and the aivoAALTO project of Aalto University.

Reference

1. Mäkäräinen, M., Takala, T.: An approach for creating and blending synthetic facial expressions of emotion. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 243–249. Springer, Heidelberg (2009)

A Thousand Words Paint a Picture - Examining Interviewer Effects with Virtual Agents

Jennifer Klatt, Astrid Marieke Rosenthal-von der Pütten, Laura Hoffmann, and Nicole C. Krämer

University Duisburg-Essen, Department for Social Psychology: Media and Communication, Forsthausweg 2, 47048 Duisburg, Germany
{jennifer.klatt,a.rosenthalvdpuetten,laura.hoffmann, nicole.kraemer}@uni-due.de

Abstract. Research in human-human context shows that gender and communication style of an interviewer affect interviewees´ answers. In this paper, we address the question of whether and how these rules apply to human-agent interaction. In a 2x2 between subjects design interview study (N=81), we investigated the influence of the agents´ wordiness and gender on participants´ self-disclosure and their evaluation of the interview and agent. High wordiness increased self-disclosure whereas the agent`s gender showed no influence. Also, the taciturn agent was perceived as more competent to conduct interviews on partnership. Overall, the interview was evaluated more positive compared to the wordy agent. Only one effect was found for the agent`s gender: the male agent was evaluated as being more competent in an interview on partnership.

Keywords: male and female agents, experimental study, gender effects, linguistic alignment, reciprocal self-disclosure, social effects, virtual agent.

Gender is one of the most crucial social cues in human-human interaction. Gender stereotypes raise expectations towards a person and dependent on the fact if those expectations are met or not, the evaluation of that person differs. However, there are conflicting empirical findings under which circumstances the evaluation is best: on the one hand, it is considered to be advantageous to act conform [1], on the other hand there are findings in which non-conformity can be seen as a positive surprise and be therefore evaluated especially positive, e.g. men who smile a lot [2]. In the context of interview systems that use a virtual agent it is crucial to know about the agent`s design aspects and their effect on the outcome of the interview. Those aspects can both be the agent`s appearance (like gender) and its behavior, like the dialogue design. In previous studies, the wordiness of the agent`s question has shown an impact on the interviewee`s self-disclosure [3]. Therefore, in this study the influence of the agent`s gender and wordiness will be examined and their effect on the interview behavior and evaluation. Thus, we ask: A) To which agents will participants disclose most information and B) which agents are evaluated most positive?

The study was designed as a 2x2 between-subjects design with the agent´s gender (male, female) and the agent´s wordiness (talkative, taciturn) as independent variables. Eighty students (39 females and 41 males) between 19 to 36 years (M=25.19 years, *SD*=3.36) participated in the study. They answered several questions in an interview with an agent on love, the role of men and women, and occupational topics (based on [3]). Also, a questionnaire was given to them including an evaluation of the interview and the agent. By means of a MANOVA, we identified two main effects for the wordiness on disclosure: First, when being interviewed by the talkative agent ($F(80,1)=14.47$, $p<.001$, *part.* $\eta^2=.232$), more than three times as many words were used ($M=696.12$, $SD=506.13$) than in the interview with the taciturn agent ($M=293.25$, $SD=190.34$). The second effect concerns the disclosure in a work-related content ($F(80,1)=6.43$, $p=.015$, *part.* $\eta^2=.118$): Toward the talkative agent, participants tended to disclose more information ($M=1.90$, $SD=.21$) than toward the taciturn one ($M=1.75$, $SD=.21$). Regarding the gender of the agent, no significant effects were found at all. In another MANOVA, there were main effects for both the agent`s gender and wordiness on the agent`s evaluation: Participants rated the male agent as being more competent ($M=3.35$, $SD=.95$) in conducting an interview on partnership ($F(80,1)=4.43$, $p=.039$, *part.* $\eta^2=.055$) than the female agent ($M=2.93$, $SD=.89$) and the taciturn agent as more competent ($M=3.35$, $SD=.83$) than the wordy one ($M=2.93$, $SD=1.00$; $F(80,1)=4.43$, $p=.039$, *part.* $\eta^2=.055$). Additionally, the interview in general was evaluated more positively after interacting with the taciturn one ($M=3.37$, $SD=.77$) than after the interview with the talkative agent ($M=2.95$, $SD=.96$; $F(80,1)=6.38$, $p=.014$, *part.* $\eta^2=.077$).

It was found that the wordiness of the agent had a stronger influence both on the answers given and the evaluation of the interview than the agent`s gender. A wordy agent leads to more words and to a greater disclosure in a work-related context, but is considered less competent in conducting an interview on partnership than a taciturn one. Also, the interview with the taciturn agent was evaluated better. Also, the male agent was perceived as more competent in conducting an interview on partnership. Again it was found that wordiness leads to more self-disclosure (like in [1]), but this does not lead to a more positive evaluation. Probably, both the male and the taciturn agent (with taciturn being a rather masculine stereotypical behavior) were evaluated more positive because it came as a positive surprise that they talk about topics like partnership. However, for the outcome of the interview, the wordy agent led to more disclosure. Therefore, dependent on the main purpose of the interview, either a male/taciturn agent should be chosen, if mainly the interview(er) should be evaluated positively or a female/wordy agent should be taken for the most disclosure.

Acknowledgments. We thank Brenda Benninghoff, Henning Geßner, and Philipp Kulms for their help on this project.

References

1. Diekman, A.B., Eagly, A.H.: Stereotypes as dynamic constructs. Women and Men of the Past, Present, and Future. Personality and Social Psychology Bulletin 26(10), 1171–1188 (2000), doi:10.1177/0146167200262001

2. Deutsch, F.M., LeBaron, D., Fryer, M.M.: What is in a Smile? Psychology of Women Quarterly 11(3), 341–352 (1987), doi:10.1111/j.1471-6402.1987.tb00908.x
3. von der Pütten, A.M., Hoffmann, L., Klatt, J., Krämer, N.C.: Quid Pro Quo? Reciprocal Self-disclosure and Communicative Accomodation towards a Virtual Interviewer. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 183–194. Springer, Heidelberg (2011)

Deep Conversations on a Flat Screen? A Comparison of Virtual Agents in Videos and Immersive Virtual Environments

Jennifer Klatt and Nicole C. Krämer

University Duisburg-Essen, Department for Social Psychology:
Media and Communication, Forsthausweg 2, 47048 Duisburg, Germany
{jennifer.klatt,nicole.kraemer}@uni-due.de

Abstract. Research suggests that the evaluation of an agent depends on characteristics like appearance, verbal or nonverbal behavior. It remains unclear how the medium used for presentation affects the evaluation and how this connects to different levels of presence. 161 participants were shown either a video or used an Immersive Virtual Environment (IVE) in which a virtual agent took on the role of a leader giving a presentation. Main results show that the IVE created a stronger sense of presence and the evaluation of the agent was more positive compared to the video. Shown in IVE, the virtual agent was perceived as more intelligent, competent, trustworthy, polite, reliable, and as having a better professional reputation.

Keywords: experiment, IVE, video, virtual agent, social presence, presence.

When using virtual agents, it is crucial to know how they are perceived by the user. One variable that may influence the evaluation of virtual agents is the presentation medium. Different media types have attributes that change the user experience. Presence has been described as the sense of being in an environment [1] and e.g. was found to be higher when a more natural view for a virtual environment is used [e.g. 2]. In an IVE "a user is perceptually surrounded" by the virtual environment [3]. Most common ones use Head Mounted Displays which display images and sounds of a virtual environment. In IVEs, the movement of the head is tracked and the information rendered accordingly. Research found that IVEs create a strong sense of presence and that they affect the experience [4]. However, virtual agents are often displayed on usual computer screens (e.g. on websites). Therefore, we examined whether the IVE creates a stronger sense of presence than a classic presentation of an agent (video on computer screen) and how it affects the evaluation of an agent.

In order to test our hypotheses, we chose a 1x2 between-subjects design with either a video on a computer screen or the IVE condition. Participants used an IVE that was displayed on the Sony Head Mounted Display HMZ-T1. 161 persons (80 females and 81 males) between 18 and 37 years (M=24.71, SD=3.74) participated in the experiment. In the virtual environment, a female virtual agent gave a presentation for about five minutes. The agent was presented as being a leader which was superior to the participant. Afterwards, a questionnaire on presence and person perception followed. By means of ANOVAs with the presentation medium (IVE vs. video) as a fixed factor, a main effect for presence was found (F(160,1)=53.22, p<.001, *part.* η^2=.251). The IVE created a stronger sense of presence (M=4.08, SD=1.08) as the video condition (M=2.97, SD=0.82). With regard to the evaluation of the agent, a MANOVA revealed five main effects of 15 bipolar items on person perception. Overall, the virtual agent was evaluated more positively in the IVE condition, she was perceived as more intelligent (F(160,1)=5.40, p=.021, *part.* η^2=.033), trustworthy (F(160,1)=4.41, p=.037, *part.* η^2=.027), competent (F(160,1)=3.93, p=.049, *part.* η^2=.024), reliable (F(160,1)=7.29, p=.008, *part.* η^2=.044), and polite (F(160,1)=4.13, p=.044, *part.* η^2=.025). Also, the perceived professional reputation of the virtual agent was measured. In an IVE, the virtual leader was perceived as having a higher reputation than in the video condition (F(160,1)=15.209, p<.001, *part.* η^2=.087). For all mean values, see Table 1.

Table 1. Mean values and standard deviations of the agent`s evaluation

Item/Scale	IVE condition		Video condition	
	M	*SD*	*M*	*SD*
Person perception				
dumb-intelligent	3.68	1.03	3.62	1.45
threatening-trustworthy	4.11	1.18	3.76	0.90
incompetent-competent	3.14	1.56	2.64	1.63
unreliable-reliable	3.90	1.31	3.34	1.33
impolite-polite	4.40	1.19	4.04	1.04
Professional reputation	4.87	1.09	4.16	1.23

In general, it was found that the IVE created a stronger sense of presence and lead to a more positive evaluation of the agent compared to the video. Significant differences occurred both for characteristics of her assumed personality like trustworthiness, reliability, and politeness, as well as her professional skills like intelligence, competence, and the perceived professional reputation. Therefore, an IVE can be an advantageous presentation medium with regard to the agent`s evaluation which may be caused by the nature of the medium and its high degree of presence. Thus, the presentation medium should be considered as well when working with virtual agents.

Acknowledgments. We thank Brenda Benninghoff, Ingo Börsting, Heike Choi, Christine Gallus, Maria Hertzenberg, Franziska Hauser, Silke Müller, Johanna Schäwel, Meike Schmidt, and Leonie Zeller for their help on this project.

References

1. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
2. Hoffmann, L., Haferkamp, N., Klatt, J., Lam-Chi, A., Krämer, N.C.: A Matter of Perspective: The Impact of First- and Third-Person Perspective on the Perception of Virtual Group Discussions. Journal of Gaming & Virtual Worlds 4(3), 239–257 (2012)
3. Loomis, J., Blascovich, J.: Immersive Virtual Environment Technology as a Basic Research Tool in Psychology. Behavior Research Methods, Instruments, & Computers 31(4), 557–564 (1999)
4. Johnsen, K., Lok, B.: An Evaluation of Immersive Displays for Virtual Human Experiences. In: IEEE Virtual Reality (2008)

yaPOSH Action Selection

Jakub Gemrot, Jan Havlíček, Michal Bída, Rudolf Kadlec, and Cyril Brom

Faculty of Mathematics and Physics, Charles University in Prague,
Malostranske namesti 25, 118 00, Prague 1, Czech Republic

Abstract. We present evolution of yaPOSH action selection system tailored for the development of intelligent virtual agents behaviors for the Unreal Tournament 2004 videogame. It was developed based on the data from the previous usability studies, in which we compared experiences of participants developing IVA behaviors in plain Java to those using Java+POSH.

Keywords: Virtual agents, Action selection, Empirical studies, Tools.

1 Action Selection Matters

One of the key aspects of intelligent virtual agents (IVAs) is their ability to convey human or animal like behaviors; ability to act and react in human or animal like fashion. Those illusions of intelligent behaviors relate greatly to the graphical representation of IVA bodies and the smoothness of their animations. However, such behavior must also be contextually appropriate, which has to be supported by IVAs hidden-to-user robust action-selection (AS).

In last four years we have started to address the issue of AS systems comparison systematically. We have been running a course on IVAs development for computer science students at Charles University in Prague since 2005. Since the academic year 2009/10 we have been conducting scientific experiments as part of the final exam for the course, in which students are creating IVA behaviors for game-like scenarios. The data collected are used to drive improvements of the AS system students are working with; yaPOSH (dialect of POSH AS system [1]). yaPOSH AS system can be seen as an implementation of Behavior Trees [2] or decision trees. Results of our two studies can be found in [3, 4].

Concerning behavior-tree-like tools we have found out, and improved yaPOSH AS system accordingly that: 1) all tree nodes must be parameterizable to spare user of creating duplicate branches that differ only in one or few parameters used in leafs (actions and senses) and to help the overall reusability of created sub-trees, 2) AS system must recognize at least three return values from IVA action; `FINISHED` (an action has been finished successfully), `RUNNING` (an IVA's action is still being executed within the environment), `FAILED` (an action execution has failed), 3) tree editor must be tightly coupled with the editor of IVA behavior primitives (actions and senses), so it does not slow user down, 4) AS system needs to be accompanied with online debugger, that helps user to understand decisions made by the system during runtime.

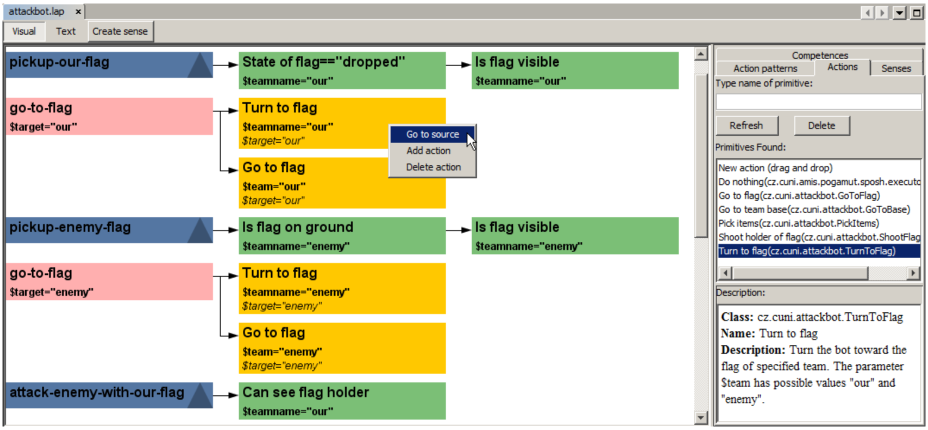


Fig. 1. The screenshot from the yaPOSH plan graphical editor displaying the example part of the plan for the UT2004 IVA capable of playing capture-the-flag game

The yaPOSH AS system brings academic tools closer to the gaming industry, where such tools are commonly used during the production. We are also experimenting with this system in the context of AAA game title that is being commercially developed by WarHorse studio inc. using CryEngine. The yaPOSH AS system can be downloaded as the part of Pogamut platform [5].

Acknowledgement. This research is supported by the Czech Science Foundation under the contract P103/10/1287 (GACR), by student grant GA UK No. 655012/2012/A-INF/MFF. This research was partially supported by SVV project number 267 314.

References

1. Bryson, J.J.: Intelligence by design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agent. PhD Thesis, MIT, Department of EECS, Cambridge, MA (2001)
2. Champandard, A.J.: Behavior Trees for Next-Gen Game AI. Internet presentation (April 21, 2013), `http://aigamedev.com/insider/presentations/behavior-trees`
3. Gemrot, J., Brom, C., Bryson, J., Bída, M.: How to compare usability of techniques for the specification of virtual agents' behavior? An experimental pilot study with human subjects. In: Beer, M., Brom, C., Dignum, F., Soo, V.-W. (eds.) AEGS 2011. LNCS, vol. 7471, pp. 38–62. Springer, Heidelberg (2012)
4. Gemrot, J., Hlávka, Z., Brom, C.: Does high-level behavior specification tool make production of virtual agent behaviors better? In: Dignum, F., Brom, C., Hindriks, K., Beer, M., Richards, D. (eds.) CAVE 2012. LNCS, vol. 7764, pp. 167–183. Springer, Heidelberg (2013)
5. Pogamut 3 platform (June 9, 2013), `http://pogamut.cuni.cz/`

Social Identity Bias in Agents' Rational Decision

Joana Dimas, Phil Lopes, Guida Preto, Pedro A. Santos, and Rui Prada

INESC-ID and Instituto Superior Técnico, Technical University of Lisbon,
Av. Prof. Cavaco Silva, 2744-016 Porto Salvo, Portugal
{joana.dimas,plopes}@gaips.inesc-id.pt,
guida.preto@ist.utl.pt, pasantos@math.ist.utl.pt, rui.prada@ist.utl.pt

Abstract. The Dynamic Identity Model for Agents allows simulating the influence of social context in autonomous agents' identity. Social context changes the way people perceive their or others identity (either as part of a social group or as unique and distinctive individuals). This perception tends to bias rational decision by leading to more cooperation with the members of a group, even when the groups goals contrast from the personal goals. The process is dynamic since the identity perceived shifts as the social context changes and its salience (strength) changes as well. We believe this is crucial for autonomous agents that face scenarios with social dilemmas, where group and personal interests are in conflict.

Keywords: context-situated agents, cooperation, dynamic identity, social bias, social identity, social dilemma.

1 Introduction

One of the processes that greatly influences a person's identity is how one sees oneself and others regarding the membership of social groups [5,6]. When in the presence of an out-group, the perception as group member strengthens, because a person tends to focus his or her perception on the shared features with other in-group members. The person sees itself as less distinctive from the rest of its own group, and when that occurs, there is a shift in the identity (e.g. motives, values and interests) from self (personal) to the group's. But, in the absence of a strong out-group, a person becomes aware of each member's uniqueness and specific personal attributes, relating to others in an interpersonal manner, dependent on their personality traits and close personal relationships, thus, using a more personal identity [1]. For the above reasons, this process of social identification often leads to bias in rational decision directing people to cooperate more with members of their in-group when the social group's identity is salient, even if, from the individual's perspective, the rational decision would be not to do so [3].

2 DIMA and Rational Decision

The Dynamic Identity Model for Agents (DIMA) allows agents to have their identity associated with several social identities (one for each group membership) besides their personal identity and also have their behaviour influenced by the one that has more salience (strength) according to the social context (see more in [2]). DIMA was implemented on the Project INVITE's research tool[1] [4] that allows the configuration of a myriad of game theory paradigms. In one of the possible situations, the players' goal is to escape an island before a volcano erupts. Players are assigned into teams and each team must build a raft in order to survive. Consequently, players must gather wood for their team, throughout several days. However, gold can also be found scattered all over the island. In the end, the player that survives with more gold wins. Players are then faced with the dilemma of either helping everyone by collecting wood (team's interest) or gathering gold and thus become rich when they escape (personal interest).

Using DIMA, the process of social identification can lead to a bias in the agent's decision. Depending on each team's members characteristics (e.g. shirt colour) and relevant aspects from the environment, the agent's identity can shift from personal to social. The team which players identify with the most, have a stronger social identity salience. Thus, while the rational decision is prioritizing self gain, this bias shifts the decision into more favourable outcomes to the team, resulting in more wood gathered. The higher the salience is, the strong the effects of the bias are.

Acknowledgements. This work was partially supported by the INVITE project (ref. UTA-Est/MAI/ 0008/2009) funded by FCT under the UT-Austin/Portugal cooperation agreement and by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2013, the PIDDAC Program funds.

References

1. Brewer, M.: In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. Psychological Bulletin 86(2), 307 (1979)
2. Dimas, J., Lopes, P., Prada, R.: One for all, all for one: Agents with social identities. To Appear in Proceedings of the CogSci 2013 - 35th Annual Conference of the Cognitive Science Society (2013)
3. Kollock, P.: Social dilemmas: The anatomy of cooperation. Annual Review of Sociology, 183–214 (1998)
4. Prada, R., Raimundo, G., Dimas, J., Martinho, C., Peña, J.F., Baptista, M., Santos, P.A., Ribeiro, L.L.: The role of social identity, rationality and anticipation in believable agents. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, vol. 3, pp. 1175–1176. International Foundation for Autonomous Agents and Multiagent Systems (2012)
5. Tajfel, H.: Differentiation between social groups. Academic Press, London (1978)
6. Turner, J., Oakes, P., Haslam, S., McGarty, C.: Self and collective: Cognition and social context. Personality and Social Psychology Bulletin 20, 454 (1994)

[1] `http://project-invite.eu/`

Tell Me a Story: A Comparative Study of Child Interaction with Virtual Characters and Adults

Ovidiu Şerban[1,2], Alexandre Pauchet[1], Anne Bersoult[3], and Émilie Chanoni[3]

[1] INSA de Rouen - LITIS, France
{forename.surname}@insa-rouen.fr
[2] "Babeş-Bolyai" University, Romania
[3] Rouen University - PsyNCA, France
{forename.surname}@univ-rouen.fr

Abstract. A storytelling setup gives the opportunity to conduct interaction studies in a familiar environment for children. By applying a Wizard of Oz technique, we simulate intelligent behaviour for an Embodied Conversational Agent, which allows children to be engaged into an interaction similar to the one observed in the presence of an adult. We present an experimental protocol, formalized into a narrative scenario, which is executed through the platform we propose, the Online Annotation Toolkit (OAK). Moreover, we conduct a comparative study over a series of interaction features, observed in the context of a virtual character and an adult, in video conference mode.

Keywords: User Study, Child-Computer Interaction, Storytelling Environment, Wizard of Oz.

1 Introduction

Building a virtual natural environment, in which the participants can interact without any difficulty, is very challenging. Moreover, introducing a virtual conversational agent into this kind of environment increase the expectations of the human participants, up to the point where they are disappointed by the agent's capabilities [1]. Building such an environment for children is even more difficult. Providing them a familiar environment, with natural reactions from a conversational agent, becomes critical.

Our purpose is to create a new environment, centred around the story telling activity, which allows all the participants to act naturally, even if the new dialogue partners are not their usual ones. This setup has two types of participants: a listener (the child) and a storyteller. The narrator (storyteller), can be either a psychologist present in a video conference mode or an avatar (an animated virtual character, driven by a psychologist). During the activity, the child is interacting with one of the partners for the first half of the story and it continues with the other. Our goal is to compare the two situations and to measure the difference between the two interaction environments. This is done by setting up a Wizard of Oz scenario, in the context of a storytelling activity.

2 Project Results

The group selected for this analysis consists of 20 children (7 girls and 13 boys), which were chosen due to their age homogeneity and development, providing statistically relevant results, as well.

Except for the number of pauses, there is not statistical relevant difference between the two modalities, which means that the children do not feel an actual difference between the two modalities. The only significant difference is the number of pauses taken to respond to the questions. This cannot be linked with an attention deficit, because all the children participating on the experiment successfully responded to the final survey, which consists in describing some specific aspects of the story. We believe this could be linked with the style of the narrator, as the avatar tends to be more monotonous than the psychologist. Nevertheless, this offers a good feedback to the new system design.

During the communicative error states, we observed an interesting difference in the interaction modality. The children use face gestures or postures more often to indicate that something went wrong when talking to the adult in video conference mode. While discussing with the virtual character, this tendency migrated to verbalisation, rather than using gestures. Moreover, during the interaction with the avatar, they used shorter and more concise sentences.

Based on the selected statistical results, we can conclude that children are able to adapt to the system, and that they enjoy the interaction with it, even if it is not as natural as with a real narrator. Moreover, due to our avatar interactivity, only very few children compared it with a cartoon like character.

3 Conclusion

During the experiment, the children interacted with the virtual character similarly to how they interacted with a human in video conference mode. This can be explained by their high ability to adapt to this kind of systems and that they have lower expectations than the adults. This result sustains our initial hypothesis, presented in the introduction, and offers the possibility to test new interactive models with children, in future.

Building OAK allowed the psychologists to model the protocol and scenario very easy. Moreover, the selected results show that the children are able to adapt to the new environment well, without making any effort. The experiments show several difference in the interaction modality and a low level of disfluencies when the children are interacting with the virtual character.

Acknowledgement. This work part of the ACAMODIA project, funded by the French CNRS (PEPS INS2I-INSHS 2011).

Reference

1. Mori, M.: The uncanny valley. Energy 7(4), 33–35 (1970)

A Dynamic Persuasive Dialogue Model for Encouraging Social Interaction for Older Adults*

John Paul Vargheese, Somayajulu Sripada, Judith Masthoff, Nir Oren, and Matt Dennis

Computing Science & dot.rural Digital Economy Hub,
University of Aberdeen, Aberdeen, AB24 5UA
{jpvargheese,yaji.sripada,j.masthoff,n.oren,m.dennis}@abdn.ac.uk

1 Introduction and Background

Telecare systems support independent living for older adults. However there is a concern that telecare systems may increase social isolation [1]. Telecare systems have been criticised for not addressing the importance of social connections and social activities for older adults [2]. Persuasive influences have been suggested as a possible solution for increasing the effectiveness of telecare systems [3]. Our research is focused on developing a persuasive agent, for encouraging social interaction amongst older adults. During a previous study, we identified six persuasive strategies used by formal carers for older adults in residential care to encourage social interaction [4]. We also developed a profile assessment model (PAM)[1] for determining which strategy to apply [4]. In this paper we describe a validation study, to confirm that the persuasive strategies discovered in [4] are used to encourage social interaction amongst older adults and to validate the PAM.

2 Methodology and Results

We conducted a semi-structured interview study with 12 formal carers for older adults in residential care. We used two groups. Group A, which consisted of seven participants who participated with our previous study and group B, five participants who did not. All participants were asked to confirm whether the persuasive strategies discovered during our previous study are used to encourage social interaction amongst older adults *(Q1)*. Group A was required to confirm the PAM was an accurate representation of the process used by carers to determine which strategy to apply *(Q2)*. Group B was required in order to ensure the persuasive strategies and PAM were not only used by participants who were previously interviewed. Additionally, all participants were asked to validate the PAM criteria *Q3*. The results of this study are shown in Table 1.

* The research described here is supported by the award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub; award reference: EP/G066051/1.

[1] For more details please see [4].

Table 1. Validation Study Results

Carer Group	Q1	Q2	Q3
A	7/7 (100%)	7/7 (100%)	7/7 (100%)
B	5/5 (100%)	N/A	5/5 (100%)

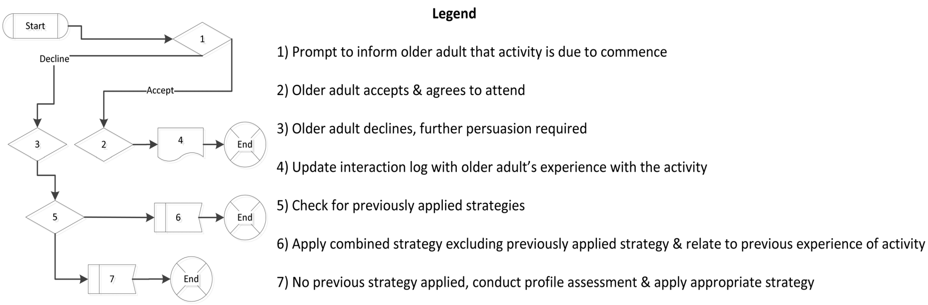


Fig. 1. Dynamic persuasive dialogue model (DPSM) for maintaining social interaction

2.1 Dynamic Persuasive Strategy Model

Participants explained that it was necessary to be aware of previously applied strategies and their outcome in order to prevent dissuading an older adult to participate by repeatedly using a single strategy. Participants explained that where further persuasion was required, strategies would be combined. To address this dynamic aspect in our model, we discussed with carers the concept of an *interaction log*. The *interaction* log would maintain a record of previous strategies applied, the outcome and determine which combination of strategies to apply simultaneously where further persuasion is required.

3 Conclusion

We have validated our findings from our previous work and developed a DPSM. to address the dynamic aspect of the persuasion process. In future work we plan to evaluate the effectiveness of the persuasive strategies, PAM and DPSM.

References

1. Sethi, R., et al.: Telecare: Legal, ethical and socioeconomic factors. In: Biomed. Eng./765: Teleh./766: Assist. Techn. ACTA Press (2012)
2. Sun, H., et al.: Promises and challenges of ambient assisted living systems. In: 6th Int. Conf. on I. T.: New Gen., ITNG 2009, pp. 1201–1207. IEEE (2009)
3. Lee, D., et al.: Participatory and persuasive telehealth. Geront. 58(3), 269–281 (2012)
4. Vargheese, J.P., et al.: Persuasive dialogue for older adults: Promoting and encouraging social interaction. In: CHI 2013 Ext. Abst.: ACM SIGCHI Conf. on Hum. Fact. in Comp. Sys/Proc. ACM (2013)

The University of Edinburgh Head-Motion and Audio Storytelling (UoE - HAS) Dataset

David A. Braude, Hiroshi Shimodaira, and Atef Ben Youssef

Centre for Speech Technology Research,
School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh, EH8 9AB, UK
{d.a.braude,h.shimodaira,abenyou}@inf.ed.ac.uk

Abstract. In this paper we announce the release of a large dataset of storytelling monologue with motion capture for the head and body. Initial tests on the dataset indicate that head motion is more dependant on the speaker than the style of speech.

Keywords: Head Motion, Dataset.

1 Introduction

There are very few datasets that have tracked head and body motion during speech. Those that do exist tend to be short and have very few speakers. To address this lack we are making available a dataset of storytelling monologues that was recorded at the University of Edinburgh. This dataset contains speech and motion capture of the head and upper body and so is suitable for research into virtual avatars or body language. We are also in the process of transcribing the speech to open up avenues of research into utilising linguistic information, and enable the data to be used in speech research.

2 Description of the Dataset

The subjects were 16 UK native English speakers. Nine were female and seven were male. Ahead of the recording session the participants were given five stories, these were classical fairy-tales that they should have been familiar with from childhood.

During the recordings the participants were given the story on a teleprompter (Read speech) and then were asked to retell the story in their own words (Free speech). Previous recordings showed when speakers were asked to choose their own stories for free speech but these stories generally lasted less than two min. They were seated during the recording and they were instructed to tell the story as if to an adult native English speaker.

Five motion capture markers were placed on the chest and the participants wore another four markers on a hat to capture the body and head motion respectively. The motion capture was done with the Natural Point, Optitrack system

Table 1. Lengths of recordings (min:sec)

	Read	Free	Total
Total	371:14	323:22	694:36

Table 2. Mean Cross Entropy distance between different utterances

	intra	inter
Speaker	0.97	3.15
Style	2.86	3.23

using seven V100:R2 cameras at a 100 Hz sampling rate. Audio was captured using a free-standing directional microphone. The audio was captured at 44100 Hz with 32-bit depth and down-sampled to 16-bit .WAV format using Audacity. The audio and motion capture start at the same frame. Table 1 shows the total lengths of recordings available.

To obtain the dataset please visit either the SSPNet Project[1] or CSTR[2] websites where further details about the dataset are provided.

3 Speaker and Scenario Dependency

To determine the similarity of the speakers the Cross Entropy distance was used [1]. A multidimensional Gaussian distribution was used to model the data.

The mean distance between examples of the same type (intra) and examples of different types (inter) is given in Table 2. Style refers to whether it was read or free speech.

From Table 2 it is clear that there are differences between read and free speech and head motion from different speakers and the speaker dependence is higher than the style dependence.

4 Future Work

We are currently in the process of recording a large dataset of dialogues. It will also have a large amount of speakers and long samples from each speaker.

Acknowledgements. This work was supported by EU FP7 SSPNet (grant agreement no. 231287) The financial assistance of the South African National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Reference

1. Helén, M., Virtanen, T.: Audio Query by Example Using Similarity Measures between Probability Density Functions of Features. EURASIP Journal on Audio, Speech, and Music Processing 2010, 1–12 (2010)

[1] http://sspnet.eu/
[2] http://www.cstr.ed.ac.uk/

Incremental, Adaptive and Interruptive Speech Realization for Fluent Conversation with ECAs

Herwin van Welbergen[1], Timo Baumann[2], Stefan Kopp[1], and David Schlangen[3,⋆]

[1] Sociable Agents Group, CITEC, Fac. of Technology, Bielefeld University
[2] Department of Informatics, Natural Language Systems Division, Universität Hamburg
[3] Dialogue Systems Group, Faculty of Linguistics, Bielefeld University

1 Introduction

Human conversations are highly dynamic, responsive interactions. In such interactions, utterances are produced incrementally, subject to on-the-fly adaptation (e.g. speaking louder to keep a challenged turn) and (self) interruptions. While listening, plans for next speaking contributions are constructed, allowing very rapid turn transitions. To enable such fluent interaction in Embodied Conversational Agents (ECAs) we must steer away from the traditional turn-based non-incremental interaction paradigm in which the ECA first fully analyzes user contributions and subsequently fully plans its contribution, which is then executed entirely ballistically (providing no adaptation in nor interruption of ongoing behavior). Recently, several systems have done exactly this and introduce one or more aspects of incrementality, interruptibility, or adaptivity. Their focus is mostly on behavior planning and they introduce a limited set of behavior realization capabilities only where it helps illustrate their flexible planning strategies. Furthermore, many of these systems can be characterized as proof-of-concepts, designed for a single purpose, domain (for example, only generating backchannel feedback) or set of experiments. The focus of our ongoing work is to provide a comprehensive architecture that unifies the fluent *behavior realization* functionality of these more experimental systems and additionally can reproduce other important phenomena that occur in fluent dialog. Our architecture serves 1) as a platform for those experimenting with the 'best' way to deploy fluent behavior realization strategies or those researching social effects of certain deployment strategies and 2) as a building block (specifically the Behavior Realizer) in a ECA architecture that supports fluent human-ECA interaction.

To this end, we provide behavior planners and human authors of behavior realization strategies with a language for specifying behavior realization plans that allow fluent interaction on an ECA. Our specification language and realizer implementation provides:

⋆ This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster Competition and managed by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the contents of this publication.

1. Incremental construction of behavior.
2. Smooth (prosodic) concatenation of the increments (e.g. successive phrases in speech).
3. Graceful, any-time interruption of ongoing behavior.
4. Mechanisms to adapt certain parameters in ongoing behavior (e.g. pitch, loudness) at fine granularity (at least at syllable level).
5. Mechanisms to preplan behavior that can potentially be executed instantly later on. Multiple execution alternatives can be preplanned concurrently.
6. The realization of apositional beginnings (e.g. uhm) to keep or take the turn without having a plan at hand.

2 Implementation

The requirements for fluent dialog realization are satisfied by combining the flexible BML Realizer AsapRealizer [1] (and its specification language BMLA) with the incremental TTS system Inpro_iSS [2] and by extending both systems with new features. Requirement 1 is partly covered by BMLA: behavior plans can be composed by specifying successor relations between blocks. BMLA does however not provide the specification mechanisms required to postpone ongoing behavior. In addition to allowing one to specify that a BML block should be inserted *after* blocks already in the plan, we therefore contribute a specification mechanism that allow us to plan BML blocks *before* specific other BML blocks in a plan. This is especially useful to plan short contributions (e.g. repairs) *before* some already planned behavior and then continue with the existing plan. Requirement 2) is satisfied by Inpro_iSS's capability to provide smooth prosodic connections between speech increments. To satisfy requirement 3, we combine BMLA's specification mechanisms for interruption with Inpro_iSS capability to interrupt speech at phoneme level. BMLA's parameterization mechanisms are flexible enough to be directly applicable for speech parameterization to satisfy requirement 4. We have implemented adaptable parameters in speech (e.g. loudness, tempo, pitch) using Inpro_iSS's speech adaptation mechanisms. BMLA's preplanning satisfies requirement 5 and fits well with parallel plan construction. To realize requirement 6, we contribute the specification and automatic realization of speech fillers.

We have thus contributed to both the specification and implementation of the realization of incremental, adaptive and interruptive behavior. This contribution is implemented in AsapRealizer, allowing it to support the flexible realization of (synchronized) gesture and speech.

References

1. van Welbergen, H., Reidsma, D., Kopp, S.: An incremental multimodal realizer for behavior co-articulation and coordination. In: Nakano, Y., Neff, M., Paiva, A., Walker, M. (eds.) IVA 2012. LNCS, vol. 7502, pp. 175–188. Springer, Heidelberg (2012)
2. Baumann, T., Schlangen, D.: Inpro_iSS: A component for just-in-time incremental speech synthesis. In: ACL System Demonstrations, pp. 103–108 (2012)

The Effect of Multiple Modalities on the Perception of a Listening Agent

Ronald Poppe, Mark ter Maat, and Dirk Heylen

Human Media Interaction Group, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{r.w.poppe,d.k.j.heylen}@utwente.nl

Listening agents are IVAs that display attentive listening behavior to a human speaker. The research into listening agents has mainly focused on (1) automatically timing listener responses; and (2) investigating the perceptual quality of listening behavior. Both issues have predominantly been addressed in an offline fashion, e.g. based on controlled animations that were rated by human observers. This allows for the systematic investigation of variables such as the quantity, type and timing of listening behaviors. However, there is a trade-off between the control and the realism of the stimuli. The display of head movement and facial expressions makes the animated listening behavior more realistic but hinders the investigation of specific behavior such as the timing of a backchannel.

To migitate these problems, the Switching Wizard of Oz (SWOZ) framework was introduced in [1]. In online speaker-listener dialogs, a human listener and a behavior synthesis algorithm simultaneously generate backchannel timings. The listening agent is animated based on one of the two sources, which is switched at random time intervals. Speakers are asked to press a button whenever they think the behavior is not human-like. As both human and algorithm have the same limited means of expression, these judgements can solely be based on aspects of the behavior such as the quantity and timing of backchannels. In [1], the listening agent only showed head nods. In the current experiment, we investigate the effect of adding facial expressions. Facial expressions such as smiles and frowns are known to function as backchannels as they can be regarded as a signal of understanding and attention.

Experiment Setup

We use an asymmetric version of the SWOZ setting [1]. A human speaker and listener are at different locations and their communication is mediated via the framework. They engage in a conversation where the speaker does the talking and the listener displays backchannel feedback at appropriate times. To this end, the listener is shown the video and audio of the speaker, whereas the listener is represented as a virtual listener to the human speaker. The source of the virtual listener is switched at random times. We ask the human speaker to press a button (the yuck button) everytime the behavior is perceived as unhuman-like.

We use a 2 (role) $\times$ 2 (condition) within-subjects design. In the *static* condition, the listening agent shows only nods, indicated by the human listener by pressing the space bar. Simultaneously, the algorithm from [1] generates backchannel timings. In the face condition, facial expressions are also shown. These are always animated based on those of the human listener using [2]. From the detected activations, we animated those of the mouth and the brows.

The SWOZ framework switches after a random amount of time (between 10 and 50 seconds). The interaction starts with the human listener as the source, and when the speaker presses the yuck button, the source is always set to the human listener. Speakers were informed of this, but were unaware of the duration of the switching interval. Speakers were free to choose any topic, and were provided with some suggestions. Interactions were stopped by the experimenter after approximately five minutes.

Results and Discussion

Ten subjects (five pairs) took part in the experiment. We recorded 111.61 minutes of dialog in the 20 interactions. In 61.56% of the time, the listener agent's head nods originated from the actual listener. The yuck button was pressed 86 times, 57 times (66.28%) while the listening agent's head nods were generated by the algorithm. For the algorithm and human respectively, this amounts to 1.33 and 0.42 yucks per minute. The difference between static and face condition is small: 0.82 and 0.72 yucks per minute, respectively. A repeated measures ANOVA shows a significant effect for source ($F(1) = 5.964, p < .05$), but not for condition ($F(1) = 1.418, p = .26$) or the interaction between the two ($F(1) = 2.303, p = .16$). Closer analysis, however, reveals that the diffence between the two conditions is much larger when the listener is animated by the algoritm: 1.11 and 1.58 yucks per minute for the static and face conditions, respectively. We expect this difference is because observers are more forgiving when the timing of the backchannels is less human-like. We compare the timings of both and consider a displayed backchannel matching if it is produced within a margin of one second of a backchannel produced in the other source. Of all backchannels shown to the speaker, 19.5% of those generated by the human listener, and 10.2% of those of the algorithm match the other source. Indeed the timings of the backchannels produced by the algorithm are less appropriate.

The percentage of matching backchannels for the algorithm in the static and face condition is similar (10.2% and 10.0%). However, the number of yucks is much lower in the latter. Apparently, the additional display of facial expressions causes the speakers to reduce their yuck presses only when the timing is less accurate. This finding is important as adding more modalities might similarly bias the results when performing experiments to analyze the human perception of listening behavior. A;sp, when developing listening agents, this finding might be used to improve the human-likeliness by adding more modalities of expression.

References

1. Poppe, R., ter Maat, M., Heylen, D.: Online backchannel synthesis evaluation with the Switching Wizard of Oz. In: Joint Proceedings IVA Workshops, pp. 75–82 (2012)
2. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: Proceedings ICCV, pp. 1034–1041 (2009)

Who Is the Werewolf?

Mei Yii Lim[1], Ruth Aylett[1], João Dias[2], Birgit Endrass[3], Eva Krumhuber[4], Gert Jan Hofstede[5], and Lynne Hall[6]

[1] Heriot Watt University, Edinburgh
{M.Lim,R.S.Aylett}@hw.ac.uk
[2] INESC-ID, IST, Portugal
joao.dias@gaips.inesc-id.pt
[3] Ausburg University, Augsburg
endrass@hcm-lab.de
[4] Jacob University, Bremen
e.krumhuber@jacobs-university.de
[5] Wageningen University Netherlands
gertjan.hofstede@wur.nl
[6] University of Sunderland, UK
lynne.hall@sunderland.ac.uk

Abstract. This paper describes the virtual drama application MIXER, designed to educate 9-11 year old children in cultural sensitivity through a scenario in which school-age virtual characters play the game Werewolf, while the child user supports a friend character by interacting using an iPad and a novel icon-based language. The use of a Theory of Mind component to support deceitful behaviour in the game is explained and evaluation approach is discussed.

Keywords: cultural understanding, virtual learning environment, intelligent virtual characters, icon-based interaction, theory of mind.

1 Introduction

In an increasingly globalised society, people from quite different cultures mix more than ever before, and cultural misunderstanding is therefore also more widespread, sometimes resulting in serious conflict. Education and training in cultural diversity is an obvious response. The development of cultural sensitivity can be seen as a process of developing acceptance of people belonging to a given out-group into ones own *moral circle* [1], formed by those who follow a common set of moral rules and values and who trust each other [2]. The concept of the moral circle and of in- and out-groups underlies the work discussed in this paper. MIXER (Moderating Interactions for Cross-Cultural Empathic Relationships) application in eCUTE (Education in Cultural Understanding, Technology Enhanced)[1] project uses schools as its basic social unit and the rules of a game as its framework for cultural conflict. Here it draws on the card-game BARNGA [3] which is successfully used for cultural training, making the same analogy between game and social rules within the framework of a social game.

[1] http://ecute.eu/

2 MIXER

MIXER is built up of Unity3D graphics engine, an agent module and the iPad interaction module using an icon based language inspired by Chris Crawford's toy language [4] where actions or feelings are representative icons. The scene is set in a virtual summer camp where two groups of children (IVAs) come together and play *Werewolves*. One of the players, Tom, is new to the summer camp. The rules are explained to him by the first team and as he plays the game, the child user acts as his invisible friend and watches him, responding to his requests for advice on how to react at different stages of the game. The two teams of IVAs play the game with a significant difference in their rules which results in a strong episode of conflict as Tom switches from one team to another. Because of this different rule, Tom is "killed" in the very first round of the game where he would not have been using the first set of rules. The aim is to create reflection on how what appears to be *unfair* behaviour may actually be due to a different set of rules, hence learning that there is more than one way to run a social situation.

The IVAs' mind architecture is based on a cognitive appraisal-based architecture, FAtiMA [5]. This architecture has been extended to support N-level Theory of Mind (ToM) mechanisms to increase the agent's reasoning capability via deception during the game. A ToM process allows for an agent to attribute a mental state to other agents and reason about it. It allows the IVAs playing villagers to have a rationale for their accusations of other IVAs as the werewolf while the IVA that is playing the werewolf to make accusations that divert attention from itself.

Mixer is about to enter a long-term evaluation over a period of weeks. This evaluation will use a set of questionnaires, including the Social Connectedness instrument validated by Yoon and colleagues [6] and composed of the Social Connectedness in Mainstream Society (SCMS) and Social Connectedness in Ethnic Community (SCEC) questionnaires. This instrument asks the participant about their perception of their relationship to other cultures based upon items such as behaviour, knowledge and empathy. It will be applied pre- and post-interaction with MIXER in order to establish if atttitudes have been changed as a result of the experience.

References

1. Singer, P.: The expanding circle. Clarendon, Press Oxford (1981)
2. Hofstede, G.J.: The moral circle in intercultural competence: Trust across cultures, pp. 85–99 (2009)
3. Thiagarajan, S.T., Thiagarajan, R.: Barnga. Intercultural Press (2011)
4. Crawford, C.: (2011), http://www.erasmatazz.com/PeterPan/PeterPan.html (last viewed March 20, 2013)
5. Dias, J., Paiva, A.: Feeling and reasoning: A computational model for emotional characters. In: Bento, C., Cardoso, A., Dias, G. (eds.) EPIA 2005. LNCS (LNAI), vol. 3808, pp. 127–140. Springer, Heidelberg (2005)
6. Yoon, E., Jung, K.R., Lee, R.M., Felix-Mora, M., et al.: Validation of social connectedness in mainstream society and the ethnic community scales. Cultural Diversity & Ethnic Minority Psychology 18(1), 64 (2012)

Evaluating the Impact of ECAs on User Navigation Performance and Perception

Ioannis Doumanis and Adalet Serengul Smith

Middlesex University
The Burroughs Hendon
London NW4 4BT
Id074@live.mdx.ac.uk, s.smith@mdx.ac.uk

Abstract. The paper presents a user study designed to examine the impact of the presence of a multimodal ECA on the users ability to navigate routes of different complexity. The study was conducted in the lab, using high resolution video clips representing two routes in an archaeological attraction. The routes differed both in terms of navigation complexity and length. Participants interacted with an ECA-based system and then with a non-ECA system that provided navigation instructions based on photographs of landmarks. Results indicate that although an ECA does not enhance the participants performance in navigating routes of varying difficulty, participants perceived it as more useful in helping them to decide where to go, than the system without the ECA.

Keywords: Embodied conversational agents, mobile tour guides, navigation systems.

1 Introduction

Embodied Conversational Agents (ECAs) can be a powerful user interface for navigation systems because of their ability to disambiguate navigation instructions using relevant verbal and non-verbal means (e.g., speech augmented by facial expressions and gestures). This multimodal communication can be particularly useful in pedestrian landmark based navigation systems (e.g., Google Maps Navigation) [1], where ECAs can convey photograph information using appropriate verbal and non-verbal behaviours much like humans do. We present a user study that evaluates the impact of the presence of a multimodal ECA on the cognitive accessibility of a pedestrian landmark-based navigation system, providing navigation instructions in a real archaeological attraction. The system consists of a no-ECA version and a version with an ECA capable of augmenting the acoustic navigation instructions with relevant non-verbal behaviours (e.g., gestures to show which way to go). In a within-subjects design 18 participants interacted with both systems each on two routes of variable complexity (i.e.,a simple and a complex route). We evaluated the question of the impact of the presence of the ECA on navigation performance and subjective perception of the cognitive workload required to navigate the routes.

[1] http://www.google.co.uk/intl/en_uk/mobile/navigation/

2 Results and Discussion

Performance Measures: Results show a significant interaction of time to complete a route for the type of ECA and the order of task (simple vs. complex route and vice versa) ($F(1, 32) = 11.940$; $p < .01$). A further analysis of the interaction showed that the variation of order significantly influenced the time performance of participants using the system without the ECA ($F(1, 32) = 13.213$; $p < .01$), but not the system with the ECA. One possible explanation is that participants focused their attention more on the additional modalities (other than speech) provided by the two systems (i.e., text or gestures), to disambiguate the navigation instructions.On the system without the ECA, participants commented that if they would fail to read the text they would become unsure what step to take. The gestures used in the system with the ECA, did not have such a negative impact on their ability to navigate the assigned routes.

Usefulness: Results show a significant effect of order of task ($F(1, 32) = 7.040$; $p < .05$).Participants thought that the system with the ECA was more useful in assisting the navigation of the complex route than in the navigation of the simple route. They had the opposite views for the system without the ECA. One explanation is that the landmark pictures, were enough for the effective navigation of the simple route and the presence of the ECA was not deemed as necessary. Then, the system without the ECA was thought as less useful in the complex route, than in the simple route because the system read the textural instructions very fast for the participants to read.

Perception of the Cognitive Workload: We report the following significant findings: a) An effect of the type of ECA ($F(1, 32) = 5.434$; $p < .05$) on the difficulty to make sense of the navigation instructions. Participants using the system without the ECA thought that the navigation instructions were significantly more difficult than the participants using the system with the ECA. The ECA augmented the acoustical information with relevant gestures to show the participants where to go. The combination of relevant body gestures with speech made it easier for the participants to make sense of the instructions given by the ECA (and to take better navigation decisions) than the system without the ECA that used only voice and text.b) An effect of the order of task ($F(1, 32) = 6.516$; $p < .05$) and an interaction between the type of ECA and the type of route ($F(1, 32) = 6.527$; $p < .05$) on the clarity of the presentation of the navigation instructions. A further analysis of the interaction identified that participants viewed the system with the ECA as a significantly more consistent method of presenting navigation instructions of variable difficulty than the system without the ECA. This consistency can be attributed to the gestures used by the ECA in both types of routes. The gestures were well synchronized with the speech and designed in a way to augment the instructions provided. On the other hand, the text used by the system without the ECA though synchronized with the speech, made it difficult for participants to read the instructions (especially when an instruction was too long).

Modelling Social Power Intelligent Agents

Gonçalo Pereira, Rui Prada, and Pedro A. Santos

INESC-ID and Instituto Superior Técnico and Technical University of Lisbon
Avenida Professor Cavaco Silva, Porto Salvo, Portugal
goncalo.pereira@gaips.inesc-id.pt, rui.prada@ist.utl.pt,
pasantos@math.ist.utl.pt
http://gaips.inesc-id.pt/

Abstract. Social power, regardless of its pervasiveness and acknowledged impact in a multitude of human social processes remains little explored in IVAs. As such, to address this gap in social intelligence and consequently virtual agent believability, we briefly introduce the core processes for a cognitive architecture for social power intelligent agents.

Keywords: social power, cognitive architecture, social intelligence.

1 Introduction

Social power is one of the most pervasive concepts in human societies due to its function as a *social heuristic* [1] for decision making. It combines diverse (and in themselves) complex decision influencing social concepts such as formal/informal norms, resource/action dependencies or social status[2]. Given its impact in a in a multitude of social processes[2] and interactions we argue it is fundamental to endow social intelligent virtual agents with such cognitive structures.

2 Cognitive Architecture for Social Power Intelligence

To create an agent architecture that can endow agents with power awareness and the ability to generate behaviors based on plans including power strategies our main inspiration was Raven's Power Interaction Model [1]. It defines the main stages underlying the cognitive process of a social power episode associated with an influence attempt, from both an influencing and influenced perspective.

From an influencing agent's perspective, its influence cognitive process is triggered by a motivation to influence (e.g. satisfy need or meet a role requirement). Next an assessment of its powers over the target of the influence is performed and an interaction planning is executed taking into account all the situational factors and possible effects. When the targeted agent detects the influence attempt it decides on its compliance based on the powers that the influencing agent has over it and expected effects of complying with the required action.

Based on the cognitive stages identified in [1] we conceptualized an agent architecture for a social power intelligent agents which integrate three fundamental core social power processes in a typical agent architecture. Next we will briefly introduce the conceptualized core social power processes.

2.1 Power Situational Analysis

This process identifies and quantifies the (believed) social power forces relevant to a given interaction. To do so the agent must not only detect powers directly mapped from the social power underlying factors (e.g. rewards, liking relations, expertise relations), but also their interdependencies (e.g. norms and coercions/punishments) and relation to beliefs regarding other agents' beliefs.

2.2 Power Effects Assessment

This process identifies the effects (or outcomes) of a current (or possible) social power interaction. Some inspirations to model these effects come from French and Raven[3]. For example, when reward power is used referent power increases (related to identification based on liking or attraction). In the case of coercions the effect is opposite.

2.3 Power Interaction Planner

The purposed of this process is to perform planning for social power-based interactions. This capability enables an agent to reason about possible influence situations and choose its best option to influence other agents by integrating all its knowledge about social powers, their effects and its own power utilization preferences. An agent can influence others in many different manners, for example it might simply ask another agent to do something or it might ask that same thing while emphasizing a legitimate reason for it. These modality considerations are what we call power strategies and are used to give emphasis on one or several bases of power.

3 Summary

In this work we briefly introduce theoretical background to inspire social power modeling in IVAs. This interdisciplinary link is the basis for the development of a IVA cognitive architecture that can endow agents with reasoning and interaction capabilities in social contexts with an ecology of power.

Acknowledgments. Work supported by FCT(INESC-ID multi annual funding) through the PIDDAC Program PEst-OE/EEI/LA0021/2013 and FCT scholarship SFRH/BD/66663/2009.

References

1. Raven, B.: A power/interaction model of interpersonal influence: French and raven thirty years later. Journal of Social Behavior & Personality 7(2), 217–244 (1992)
2. Castelfranchi, C.: The micro-macro constitution of power. Protosociology 18-19, 208–268 (2003)
3. French Jr., J., Raven, B.: The bases of social power. In: Studies in Social Power, pp. 150–167. Univer. Michigan, Oxford (1959)

Evaluating the Accessibility of ECA-based Information Presentation Systems

Ioannis Doumanis and Adalet Serengul Smith

Middlesex University
The Burroughs Hendon
London NW4 4BT
Id074@live.mdx.ac.uk, s.smith@mdx.ac.uk

Abstract. The paper presents a new technique for evaluating the accessibility of cultural heritage content presented by Embodied Conversational Agents (ECAs). The technique combines data from face expression analysis, eye tracking and retention tests to measure the accessibility of the content. In order to validate this technique two tour guide applications were created with an embodied conversational agent (ECA) that presents cultural content about a real-tourist attraction. The agent simulates two attention-grabbing mechanisms, one humorous and another serious to attract the users' attention. In particular, it gives the illusion of analysing the users' attention to the presentations and requests his/her attention back when it has been deviated. A formal study was conducted that compared two versions of the application in the lab. Results indicate that the proposed technique actually works. The data collected from the face expression analysis and eye-tracking helped to explain particularly good and bad performances in retention tests. In terms of the study results, strong quantitative and qualitative evidence was found that an ECA should not attract more attention to itself than necessary, to avoid becoming a distraction from the flow of the content. Then, it was found that the ECA had an inverse effect on the retention performance of participants with different gender and their use on computer interfaces is not a good idea for elderly users.

Keywords: Embodied conversational agents, human-centered computing, mobile tour guides, eye tracking.

1 Introduction

To date, little research in the ECA community has been conducted using advanced techniques for usability research like eye tracking, let alone, using a technique that combines data from eye tracking, with data from face expression capturing. The human face is one of the strongest indicators of a human's cognitive state and hence how humans perceive stimuli (information, images, etc.). A technique that combines data from face expression recognition and eye-tracking can augment any traditional techniques for accessibility evaluation (e.g., questionnaires, retention tests, etc.). For example, with careful logging one can see which part of the content provided by the ECA system is more confusing, which part requires the users to think more intensively, etc. In addition, eye-tracking data can reveal where the user was looking when a particular expression occurred (e.g., confusion). In order to validate this new technique for accessibility evaluation of ECA-based information systems, in this study, it was decided to explore the space of ECA attributes such as manipulating these to attract attention. This mechanism is an important attribute that presenters must have to effectively gain the attention of their audience back, if that has been lost. Although various strategies are possible, a version of ECA using either serious or humorous attention-grabbing messages was developed for this purpose. An ECA with an automated attention-grabbing mechanism would be very difficult to control as some participants would follow what the ECA is saying, while others not. For this reason, it was decided to simulate this feature so that all participants would experience an ECA that uses either one of the two attention-grabbing strategies (serious or humorous). The hypothesis was that the ECA that uses an attention grabbing mechanism enhances the participants' ability to retain information from presentations in an archaeological attraction. The group that participated in this study was composed of thirteen participants. We conducted a short pilot study with one of the female participants, to test the approach and calibrate the equipment properly. The remaining twelve participants were assigned equally to the experimental conditions at random. The age range of the group of females was varied to investigate possible age effects.

2 Results and Discussion

Retention Performance: An ANOVA statistical test revealed a significant interaction of the type of ECA and the participant's gender ($F(1, 20) = 5.845$; $p < .05$). The interaction was further analysed using simple main effects analysis. It showed that the variation of ECA influenced the retention performance of the female participants ($F(1, 20) = 7.509$; $p < .05$) but not the retention performance of the male participants. The female participants scored significantly higher when they experienced the presentations with the non-attention-grabbing ECA than with the attention-grabbing ECA. The male participants scored better with the attention-grabbing ECA, than with the non-attention grabbing ECA. An explanation can be found in the facial expression data [1]. In the video files, it was observed that the female participants were annoyed by the repeated requests by the ECA for attention. This feeling most likely led them to lose focus on the content of the presentation, which in turn resulted in lower retention performances with the attention-grabbing ECA. No signs of irritation at the male participants when the attention-grabbing messages occurred were observed.

Face Recording: A camera attached to the desktop computer recorded the participant's face from a straight angle. It was observed (a full analysis of the face expression data can be found in [1]) that the attention-grabbing ECA evoked more facial expressions than the non-attention grabbing ECA. It was also observed that the female young participants displayed more facial expressions than the male ones and the older female participants.

Eye Tracking – Heat Maps and Gaze Trails: A series of heat maps were produced and analysed to illustrate differences between the two agent conditions per individual participant, per gender and for each of the two groups of participants. We report the following significant findings: a) Participants looked more at the ECA's face that its body features. b) The background images attracted more attention, than the ECA itself. c) The attention grabbing ECA was more effective in diverting the participants' attention to the background images than the non-attention grabbing ECA. d) Elderly female participants paid too much attention to the presentations with the attention-grabbing ECA, more than all the other participants. Gaze trails provided more quantifiable data on the look zone (i.e., the background or the ECA) and for how long. We conducted a series of ANOVAs that showed the following: A significant main effect of the type of ECA on the total time ($F(1,188) = 17.661$; $p < .001$). Participants paid significantly more attention to the system with the attention-grabbing ECA, than to the system with the non-attention grabbing ECA. A significant effect of the look zone (i.e., ECA or background) on the total time (($F(1,188) = 159.840$; $p < .001$) and the number of fixations ($F(1,188) = 79.994$; $p < .001$). Participants, regardless of the type of ECA they used, paid more attention to the background images, than the ECA itself. However, the attention-grabbing ECA was more effective in directing the participants' attention to the background, than the non-attention grabbing ECA.

Correlation (Facial Expressions, Gaze Trails, Retention Tests): We performed a correlation between the facial expressions, gaze trails and retention tests to help explain particularly good or bad performances in the retention tests. From both groups of participants, we chose one example of particularly good and bad performance in the retention tests. The retention samples selected for the attention grabbing conditions reflect both attention-grabbing strategies (humorous and serious). Then, based on the correlated data we attempted to explain the outcome. Beginning with the bad performances, the selected participants did not remember much (avg. = 3.5%) when they watched the presentations with the attention-grabbing ECA (humorous or serious). A review of the gaze trail data for the participants reveal that they spent significantly more time looking at the attention-grabbing ECA and the background ($F(1, 28) = 5.436$; $p < .05$) than the non-attention grabbing ECA. Furthermore, their face recordings reveal that at each presentation until the interruption message occurred the participants had either neutral/blank or attentive facial expressions, which shows that they were attentive to the presentations. However, when the ECA requested for attention, we noticed the following: a) the first time the ECA requested attention, participants were either surprised or curious, possibly because they did not expect the ECA to observe their behaviour and b) the initial emotion degraded gradually in every presentation when the ECA asked for attention. In fact, one of the two participants got annoyed at the second presentation when the ECA asked for her attention again. This was most likely because she was already paying attention to the presentation. It is obvious that the repeated interruptions diverted participants from the flow of the presentations, which in turn distracted them from keeping the content in mind. With regards to good performances, participants remembered a moderate amount (avg. = 34.5%) of information from the presentations. A careful review of the gaze trail data for the two participants reveals the following: The attention-grabbing ECA attracted significantly more attention to the background ($F(1, 28) = 5.939$; $p < .05$) than the non-attention grabbing ECA. Then, their face recordings reveal that until the interruption message of the attention-grabbing ECA, participants had a neutral/blank face, which shows their attentiveness to the content of the presentations. Nonetheless, in contrast to the other two sets of participants, their facial expression during the interruption messages were more constrained (e.g., a slight smile, or even neutral/blank). This most likely means that those participants were not distracted by the attention-grabbing messages and they were able to keep their focus on the presentations. In addition, the patterns observed in the participants with the bad performances when the ECA requests for attention were not observed in these participants. Apart from the slight reactions, the attention-grabbing messages did not have any effect. This provides a reasonable explanation for their good performances in the retention tests. Based on this discussion, although with caution, we argue that our proposed technique works. The technique combines data from face expression analysis, eye-tracking and retention tests to provide a high-quality alternative to the more expensive and more unpleasant method of measuring the user's brain activity [2].

References

1. Doumanis, I.: Evaluating Humanoid Embodied Conversational Agents in Mobile Guide Applications. Thesis (PhD). Middlesex University (2013)
2. Simple Usability Inc., Emotion Response Analysis through EEG technique (2013), `http://www.simpleusability.com/` (accessed February 19, 2013)

Finding the Timings for a Guide Agent to Intervene User-User Conversation to Provide Information Actively

Shochi Otogi, Hung-Hsuan Huang*, Ryo Hotta, and Kyoji Kawagoe

College of Information Science & Engineering, Ritsumeikan University, Japan
hhhuang@acm.org

1 Introduction

As the advance of embodied conversational agent (ECA) technologies, there are more and more real-world deployed applications of ECA. Various kinds of kiosk information systems are used in public places, such as shopping malls, museums, and visitor centers. The typical situation in which such systems are used is that a group of people stand in front of the kiosk and operate it in order to retrieve the information they request while talking with one another. Therefore, in order to implement an ECA that can serve as an information kiosk in public places, multi-party conversation functionality for simultaneous interaction with multiple users is indispensable.

The conversation style of most contemporary ECA systems are either agent-initiative (e.g. the agent always asks questions and the user answers) or user-initiative (e.g. the user always ask questions and the agent answers). However, the natural conversation occurred between human and human is usually mixed-initiative, i.e. both of the agent and the user may take initiative during the conversation. In this study, we are proceeding a project that aims to build an ECA capable of mixed-initiative conversation with multiple users in a typical application, information providing for users' collaborative decision making. The agent needs to reason what information to provide from the users' conversation even their demands are not clearly described. The agent also needs to identify the timings when the user may be interested in the information being provided without making them feel disturbed.

2 Interaction Corpus Collecting WOZ Experiment

To collect the video corpus for an analysis of the situations when the agent can probably intervene to do active information providing, a WOZ experiment on three collaborative decision making tasks was conducted. Pairs of experiment participants were instructed to interact with a life-size virtual agent on a screen, and to retrieve information in order to make a collaborative decision regarding given tasks. The conversation experiment was conducted with the following presumptions: (1) The users want to collaboratively make a decision from multiple candidates with the help from the agent who is knowledgeable about that task domain. (2) The users have a rough image of what they want, but they do have no idea about particular candidates in advance. (3) The users discuss on their own and acquire new information from the agent. (4) The conversation ends when the users made the final decision.

* Corresponding author.

A total of 12 pairs of college students were recruited as the participants for the experiment, all of whom were native Japanese speakers. The students came from various departments ranging from economics, life science to engineering with average age, 19.2. Of the all 12 pairs, eight were male pairs and four were female pairs. Each pair was instructed to complete three decision-making tasks: travel planning, lecture registration, and part-time job hunting. These tasks were chosen because the student participants are supposed to be familiar with these issues. In order to stimulate more active discussion, the participants were instructed to make rankings on the final decided three choices. All participant pairs were assigned to take the three sessions in various orders to cancel order effects. One student who major in computer science was recruited to operate the WOZ agent. He was chosen due to his familiarity with operating a GUI-based WOZ application, which ensured that there would be smooth interaction. The operator was asked to practice on the WOZ user interface for two hour prior to the experiment to further ensure that the agent's response time was quick enough. All the sentences that the agent could speak during the experiment were listed in a menu where relevant sentences were grouped for the WOZ operator to select from. There was also a text field that allowed the operator to type arbitrary utterances, in case these were needed.

From the data corpus, it is found that eight situations are possible for the agent to intervene user-user conversation. They are defined as the follows: *Stagnation:* the user-user conversation has gotten into stagnation and has difficulty to proceed more. *Additional support:* additional support seem to be required toward the last utterance from the agent. (1) Not understand: the users could not understand what the agent just said. (2) Question: the users had a new question to ask on what the agent just said. (3) Reaction: the users showed some reactions in their movements on what the agent just said. (4) Not hear: the users did not hear what the agent just said. *Question composition:* the users are discussing with each other and are composing a question that will issued to the agent. *Reminding:* the agent can remind the users when they are trying to recall something. (1) Forgot information: the users has forgotten some information that was provided by the agent. (2) Forgot utterance: the users has forgotten something what they said by themselves. Totally there were 149 instances found in the 12 pairs of experiment corpus. The average number of the intervention timing for each pair is 12.4 times, therefore, the frequency is about once per minute.

3 Conclusions and Future Works

This paper presents a work in finding the timings for the agent to intervene user-user conversation to provide active support in decision-making tasks. A WOZ experiment was conducted for collecting human interaction data. From the collected data corpus, eight kinds of timings were found probably allowing the agent to do intervention in the target task. For future works, we would like to enlarge the corpus and would like to apply machine learning techniques to develop the method for identifying these timings from nonverbal cues, face direction, body posture, and speech status. Next, we would like to introduce the mechanism of context management and understanding in the future. Finally, we would like to incorporate the intervention timing estimation feature into a real-world system.

Gestural Adaptation in Extravert-introvert Pairs and Implications for IVAs

Jackson Tolins[1], Kris Liu[1], Yingying Wang[2], Jean E. Fox Tree[1], Marilyn Walker[1], and Michael Neff[2]

[1] University of California, Santa Cruz
{jtolins,kyliu,foxtree}@ucsc.edu, maw@soe.ucsc.edu
[2] University of California, Davis
{yiwang,mpneff}@ucdavis.edu

Abstract. We compared nonverbal expressive behavior across matched and mismatched extravert/introvert pairs. We found that participants' gestures changed over time, adapting to the gesture style of their partner. Results will be used as the basis for the implementation of adaptable personality expression in interactive virtual agents.

Keywords: personality, gesture, non-verbal expressive behavior.

1 Gestural Manifestations of Personality in Dyads

Studies on expressive behaviors and extraversion/introversion present an extraverted individual as likely to have more animated, more frequent, and more expansive gestures, an expressive face, and wide or frequent smiles [2]. While some research has been conducted on which aspects of a particular gesture are repeated across entrainment [1], no previous study has attempted to explore how the gestures produced by a person with one personality type influence those of a person with a similar or different personality type.

We compared the behavior of an extravert-extravert dyad to an extravert-introvert dyad. The participants engaged in a loosely structured conversation from which data was collected both through audiovisual recording and motion capture suits. Motion capture was performed with a Vicon optical motion capture system consisting of 12 4-megapixel cameras, hung 9 feet above the ground on rails around the perimeter of the motion capture studio. Three participants were recruited through newspaper advertisement and were asked to complete an online pretest personality survey. Participants were chosen who scored at least .8 standard deviations above or below the mean on the personality profile's extraversion scale.

2 Results Show Stylistic Adaption of Gesture, Which Has Implications for IVAs

Gestures were transcribed using a three-tiered system that captured both temporal and spatial dimensions of gestures. Results follow:

Rate: The gesture rate of the introvert was the highest, and also the most stable over time of the three conversationalists. This indicates less effort to adapt to her interlocutor. When interacting with the introvert, the extravert increased in gesture rate over time, moving closer to the rate of her conversational partner. The matched extravert-extravert pairing presented a similar pattern, with one partner moving towards the rate of the other, but in this case it was in the opposite direction, with one extravert at a low rate throughout and the other reducing their rate to match by the end.

Broadness: The introvert started with narrower gestures and became broader while the extravert did the reverse, to the point of having narrower gestures than the introvert. This pattern was not seen in the extravert pair, where both participants shifted to larger gestures over the course of the conversation.

Elbows Out: The matched extravert pair moved together towards more open arm positions. Contrastively, the introvert remained stable while the extravert reduced the expansiveness of their arms while gesturing to more closely match that of the introvert, who kept a more closed-arm position.

Outwardness: All participants began the interaction with gestures that stayed relatively close to the body. Over the course of the conversation the matched extravert pair displayed movement towards each others style, but the mismatched introvert-extravert pair reduced the outwardness of gestures.

These preliminary data suggest that agents may need to adapt to their interlocutor and that this adaptation may be dependent on the personality the agent is trying to portray. Implications for agent design include that (1) agents must be able to sense their interlocutor's movements in order to respond correctly and interactively adapt their own expressive behavior, (2) agents need to change their behavior over time to model addressee adaptation, (3) agents may need to be assigned a personality type, as different personalities appear to adapt differently, and (4) agents may need to be programmed in advance with information about their addressees.

Acknowledgements. This research was supported by NSF Grant IIS-1115742. We thank the research assistants: Jasmine Norman, Nick Russell, Anna Wartan, and Ruben Sanchez.

References

1. Bergmann, K., Kopp, S.: Gestural alignment in natural dialogue. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society (2012)
2. La France, B.H., Heisel, A.D., Beatty, M.J.: Is there empirical evidence for a nonverbal profile of extraversion?: a meta-analysis and critique of the literature. Communication Monographs 71(1), 28–48 (2004)

Web-Enabled 3D Talking Avatars Based on WebGL and HTML5

Jonas Beskow and Kalin Stefanov

KTH Speech, Music and Hearing
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
{beskow,kalins}@kth.se

Abstract. We describe a system for plugin-free deployment of 3D talking characters on the web. The system employs the WebGL capabilites of modern web browsers in order to produce real-time animation of speech movements, in synchrony with text-to-speech synthesis, played back using HTML5 audio functionalty. The implementation is divided into a client and a server part, where the server delivers the audio waveform and the animation tracks for lip synchronisation, and the client takes care of audio playback and rendering of the avatar in the browser.

Keywords: talking avatar, webGL, html5, text-to-speech.

1 Introduction

The web as a platform for intelligent virtual agents is increasing in popularity. Traditionally, many web-based agents have employed text-based interfaces and been constrained in terms of graphical expression by browser limitations (e.g. animated GIF:s or 2D graphics), or have required the use of 3:rd party browser plug-ins. Recent web standard developments have made it possible to build advanced browser-based applications and interfaces involving streaming audio and hardware-accelerated 3D graphics without resorting to plug-ins. In this paper we describe how we leverage these recent browser advances in order to bring real-time animated talking 3D avatars with high fidelity lip synchronisation to the web.

2 Talking Avatars

People are highly sensitive to lip movements during perception of speech, and incongruent visual and auditory information is known to result in reduced intelligibility or sometimes the entire percept being altered [4]. We have been developing 3D animated avatars driven from text or speech [1] or motion capture [2] that have been shown to increase audiovisual intelligiblity of speech.

In the system presented here, we use face models generated using *FaceGen Modeller* software. The standard FaceGen models have a blendshape based facial parameterization consisting of 41 shapes, corresponding to different key poses, e.g. articulatory positions. Animation is produced by dynamically assigning weights to the different blendshapes on a frame by frame basis.

3 Client-server Architecture

The implementation is divided into a server part and a client part. The *LipSpeaker* is the client part, which runs in any WebGL-enabled web browser, and is responsible for actually rendering of the avatar to the browser window. The client is written in JavaScript, and 3D rendering is done using an intermediate highlevel graphics API [3]. Certain idle animations (eye blinks etc) are produced locally in the client. Speech animation, on the other hand, is produced server side by the *LipService* (see below). Speech animation is audio-timed in order to ensure AV synchronisation. For this to work, it is essential that accurate audio playback timing information is available. We use HTML5 audio functionality, which in our experiments has provided suficcient accuracy.

The *LipService* is the server part of the implementation. Its first function is to act as the middleware between the client and a text-to-speech service. In our current implementation we use CereProc CereVoice Cloud TTS, but in principle any TTS system that provides metadata regarding phoneme and timing information would work. LipService retrieves the speech data from the TTS server and returns an audio URL to the client.

The other function of *LipService* is to generate speech animation tracks for the avatar. This is done using a rule based system taking co-articulation and non-verbal facial movements into account [1] that takes phoneme and timing metadata as input and produces animation tracks as outout. These are sent back to the client in the form of a JSON object.

Total system latency, from synthesis request to start of animation playback is on the order of 1-2 seconds.

4 Notes

The WebGL based avatar was developed with support from the European Education, Audiovisual and Culture Executive Agency (EACEA) for use in the *LipRead* project. The avatar can be tried online at

`http://www.speech.kth.se/~kalins/projects/lipread/avatar.html`

References

1. Al Moubayed, S., Beskow, J., Granström, B.: Auditory-visual prominence: From intelligibilitty to behavior. Journal on Multimodal User Interfaces 3(4), 299–311 (2010)
2. Alexanderson, S., Beskow, J.: Animated lombard speech: Motion capture, facial animation and visual intelligibility of speech produced in adverse conditions. Computer Speech & Language (in press, March 2013)
3. Mr. Doob. Three.js (2013), `https://github.com/mrdoob/three.js` (Online; accessed April 21, 2013)
4. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264(5588), 746–748 (1976)

Deus-Ex - Interactive Machinima Movies for Cognitive Behavioral Therapy

Pablo Paredes, Anuj Tewari, and John Canny

Electrical Engineering and Computer Science,
University of California, Berkeley
{paredes,anuj,jfc}@cs.berkeley.edu

Abstract. This poster presents a theoretical framework for the use of interactive machinima (machine + cinema) as an adaptable means to deliver Cognitive Behavioral Therapy (CBT) to large audiences. The expected gains are to improve engagement, likeability and adherence through cinematographic fun and increase efficacy by providing a balance of interactivity and narrative. Theoretical foundations are aggregated in a interactive narrative agent that adapts to changes to a user mental model based on personality, cognitive and emotional traits of users with depression and a therapy model based on CBT therapy types and progression.

1 Introduction

The World Health Organization reported that mental disorders are a major financial burden for the worlds economy. They are currently the third most costly health problem in terms of disability-adjusted life-years around the world, and the largest in the US and Canada. They represent 10% of the global disease burden. Fewer than 25% of depressed patients receive the necessary treatment and an even smaller percentage of at-risk groups have access to adequate preventive care.

Cognitive Behavioral Therapy (CBT) is an effective approach for a variety of mental health and chronic diseases. CBT principles have been integrated into several computerized formats as both an adjunct to face-to-face therapy and as a stand-alone treatment. Computerized CBT (CCBT) is a fully automated implementation of CBT where patients follow a complete treatment through interactive texts and figures.

In this paper we introduce the theoretical framework to model mental health situations to create adaptive interactive machinima (machine + cinema) movies. Figure 1 shows printed material of a common CBT manual distributed to patients (in this case related to interpersonal relationship problems), and a machinima interactive decision tree created to enhance/replace this material.

2 Interactive Movie Elements

Adding interaction elements to a flat machinima movie should increase engagement and learning of the content. To guarantee those goals the interactive elements should

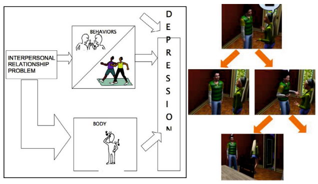


Fig. 1. Original CBT manual extract vs. machinima video screenshot

adapt to the changes in the user mental model, such as core beliefs, cognitive distortions, personality distortions and affective states. Interactive elements should also adapt to the therapy progression state and types. Figure 2 shows a summary of the three key elements of a machinima interactive movie: User Mental Model, Therapy Model and Machinima Movie Agent. These three elements interplay to generate an engaging, immersive experience that can be adapted to context, personal situations, therapy progression and even to be shared with family members.

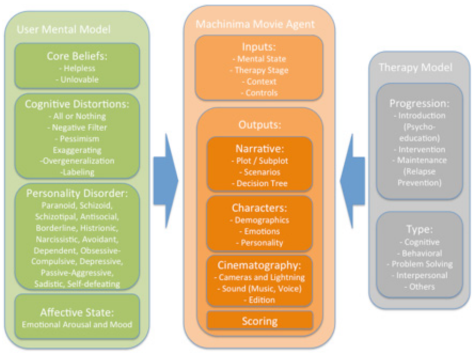


Fig. 2. Machinima Movie Agent as it relates with User and Therapy Models

2.1 User Mental Model

Psychopathology over decades has been addressing contradictory psychological traits by further separating or consolidating traits as more evidence becomes available. Based on interviews with clinical psychologist researchers, we have collected a group of theoretical foundations that would help define an effective user mental model for CBT-based machinima movies for depression treatment. The main elements of this module are Core Beliefs, Cognitive Distortions, Personality Disorder associations and Affective State

2.2 Therapy Model

Once a patient has engaged in therapy, the type of content and amount of interaction can change depending of the stages of the therapy and the type of therapy being administered. The progression of therapy and its type determines the features affecting the model.

2.3 Machinima Movie Agent

In order to ensure that mental health state is properly modeled by the machinima movie, it is important to create a simple agent that is capable of producing meaningful outputs with limited inputs and controls. A series of Inputs, affected by the therapy and user mental model affect a series of key outputs (narrative, characters, cinematography and scoring) that help define an interactive machinima plot with basic gaming elements.

Author Index

Adam, Carole 434
Allbeck, Jan M. 1, 349
André, Elisabeth 444
Andrist, Sean 249
Antunes, Rui Filipe 382
Artières, Thierry 217
Aylett, Ruth 472

Balint, Tim 1, 349
Bartie, Phil 440
Batrinca, Ligia 116
Baumann, Timo 468
Baur, Tobias 444
Beale, Russell 436
Becker-Asano, Christian 406
Ben Youssef, Atef 285, 466
Bergmann, Kirsten 139, 203
Bersoult, Anne 462
Beskow, Jonas 486
Bevacqua, Elisabetta 430
Bickmore, Timothy 68
Bída, Michal 458
Braude, David A. 285, 466
Brom, Cyril 15, 458
Brown, Timothy 129

Canny, John 129, 442, 488
Cassol, Vinícius Jurinic 448
Cavedon, Lawrence 434
Chanoni, Émilie 462
Clausen-Bruun, Mette 446
Coon, William 425
Courgeon, Matthieu 406

Dalmas, Tiphaine 440
Damian, Ionut 444
de Kok, Iwan 168
De Loor, Pierre 430
Demulier, Virginie 430
Dennis, Matt 464
Dias, João 472
Dignum, Virginia 358
Dimas, Joana 460
Ding, Yu 217
Doumanis, Ioannis 474, 478

Ebata, Ryuji 438
Ek, Tobias 446
Endrass, Birgit 472

Flach, Laura Mattos 448
Förger, Klaus 239

Gebhard, Patrick 444
Gemrot, Jakub 458
Gillies, Marco 419
Gleicher, Michael 249
Gratch, Jonathan 275, 339, 368

Haake, Magnus 446
Hall, Lynne 472
Hartholt, Arno 368
Havlíček, Jan 458
Hawes, Nick 436
Heylen, Dirk 168, 470
Higashino, Hiroshi 423
Hoffmann, Laura 452
Hofstede, Gert Jan 325, 472
Honkela, Timo 239
Hotta, Ryo 482
Huang, Hung-Hsuan 482
Huang, Lixing 275

Jack, Brian 68
Janarthanam, Srinivasan 440
Jiang, Chen 421
Johnson, Andrew E. 159
Johnson, Graham 106
Jonker, Catholijn M. 358

Kadlec, Rudolf 15, 458
Kahl, Sebastian 203
Kang, Yilin 29
Kätsyri, Jari 450
Kawagoe, Kyoji 482
Khooshabeh, Peter 339
Klatt, Jennifer 452, 455
Kleinsmith, Andrea 419

Ko, Peiyi 442
Koda, Tomoko 423
Kopp, Stefan 79, 203, 295, 468
Kramer, Marcel 79
Krämer, Nicole C. 149, 452, 455
Krumhuber, Eva 472
Kühne, Vivien 149
Kulms, Philipp 295

LaFond, Cynthia M. 159
Lee, Sangyoon 159
Leigh, Jason 159
Lemon, Oliver 440
Leuski, Anton 368
Leymarie, Frederic Fol 382
Lhommet, Margaux 303
Li, Weizi (Philip) 1
Lim, Mei Yii 472
Lisetti, Christine 92
Liu, Kris 396, 484
Liu, Xingkun 440
Lopes, Phil 460
Luo, Changwei 428
Luo, Pengcheng 229

Macedonia, Manuela 139
Mackaness, William 440
Mäkäräinen, Meeri 450
Marsella, Stacy C. 44, 303, 368
Marson, Fernando Pinho 448
Martin, Jean-Claude 406
Mascarenhas, Samuel 325
Masthoff, Judith 464
McGregor, Ewan 417
Morency, Louis-Philippe 116, 189, 368
Musse, Soraia Raupp 448
Mutlu, Bilge 249

Nebel, Bernhard 406
Nédélec, Alexis 430
Neff, Michael 180, 229, 396, 484
Ng-Thow-Hing, Victor 229

Ochs, Magalie 263
Oren, Nir 464
Otogi, Shochi 482
Ozkan, Derya 189

Paasche-Orlow, Michael 68
Paiva, Ana 58, 325
Paredes, Pablo 488
Pauchet, Alexandre 462
Payne, Jeunese 106
Pelachaud, Catherine 217, 263
Pereira, Gonçalo 313, 476
Pitsch, Karola 79
Poppe, Ronald 470
Porayska-Pomsta, Kaśka 444
Prada, Rui 313, 325, 460, 476
Preto, Guida 460
Pynadath, David V. 44

Ragni, Marco 406
Ravenet, Brian 263
Renambot, Luc 159
Ribeiro, Tiago 58
Rich, Charles 425
Rishe, Napthali 92
Robertson, Paul 106
Rosenthal-von der Pütten, Astrid Marieke 149, 452

Samperi, Katrina 436
Santos, Pedro A. 313, 460, 476
Scherer, Stefan 116
Schlangen, David 468
Şerban, Ovidiu 462
Shapiro, Ari 116, 368
Shimodaira, Hiroshi 285, 466
Sidner, Candace L. 425
Sloan, Robin J.S. 417
Smith, Adalet Serengul 474, 478
Sripada, Somayajulu 464
Stahl, Philip 406
Stanković, Igor 430
Stefanov, Kalin 486
Stratou, Giota 116, 368
Sumi, Kaoru 438
Szymkowiak, Andrea 106

Takala, Tapio 239, 450
Tan, Ah-Hwee 29
ter Maat, Mark 470
Tewari, Anuj 129, 442, 488
Tolins, Jackson 396, 484
Traum, David 368
Tree, Jean E. Fox 396, 484

Vala, Marco 58
van der Zwaan, Janneke M. 358

van Welbergen, Herwin 432, 468
Vardoulakis, Laura 68
Vargheese, John Paul 464
Vincent, Catherine 159

Walker, Marilyn 396, 484
Wang, Ning 44
Wang, Yingying 180, 484
Wang, Yuqiong 339
Wang, Zengfu 421, 428
Webber, Bonnie 440

Yaghoubzadeh, Ramin 79
Yasavur, Ugan 92
Yu, Jun 421, 428