Yukiko Nakano
Michael Neff
Ana Paiva
Marilyn Walker (Eds.)

LNAI 7502

Intelligent Virtual Agents

12th International Conference, IVA 2012
Santa Cruz, CA, USA, September 2012
Proceedings

Springer

Yukiko Nakano Michael Neff Ana Paiva
Marilyn Walker (Eds.)

Intelligent Virtual Agents

12th International Conference, IVA 2012
Santa Cruz, CA, USA, September, 12-14, 2012
Proceedings

Springer

Volume Editors

Yukiko Nakano
Michael Neff
Ana Paiva
Marilyn Walker

University of California
Baskin School of Engineering
1156 N. High SOE-3
Santa Cruz, CA 95064, USA

E-mails:
y.nakano@st.seikei.ac.jp
neff@cs.ucdavis.edu
paiva.a@gmail.com
maw@soe.ucsc.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-33196-1 e-ISBN 978-3-642-33197-8
DOI 10.1007/978-3-642-33197-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012945668

CR Subject Classification (1998): I.2.11, I.2, H.4-5, D.2, I.4-5, 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 12th International Conference on Intelligent Virtual Agents. IVA is an interdisciplinary annual conference and the main forum for presenting research on modeling, developing, and evaluating intelligent virtual agents with a focus on communicative abilities and social behavior.

This conference represents a field of specialization within computer science, artificial intelligence, and human–machine interaction that aims at creating interactive characters that exhibit human-like qualities and communicate with humans or with each other in a natural way. Intelligent virtual agents should be capable of real-time perception, cognition, and action that allows them to participate in dynamic social environments. Creating these computational models involves the integration of knowledge, methodologies, and theories from a wide range of fields such as sociology, psychology, computer science, artificial intelligence, linguistics, cognitive science, and computer graphics.

IVA was started in 1998 as a workshop on Intelligent Virtual Environments at the European Conference on Artificial Intelligence in Brighton, UK, which was followed by a similar event 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, and Reykjavik, Iceland in 2011.

This year's conference was held in Santa Cruz, California, USA, September 12–14, 2012. It combined a wide range of expertise, from different scientific and artistic disciplines, and highlighted the value of both theoretical and practical work as necessary components to bring intelligent virtual agents to life.

The special topic of IVA 2012 was games and story telling. This topic touches on many aspects of intelligent virtual agent theory and applications. Narrative and story telling is a fundamental aspect of human experience. Telling a coherent compelling narrative involves integration of multimodal presentation functionalities such as speech, gesture, and facial expressions; coherent use of discourse context and appropriate contextual verbal and nonverbal gestures, the ability to portray personality and emotions, and an ability to monitor the audience and their reaction to the story. The talks by the three invited speakers addressed different aspects of essential requirements for IVAs. The talk by Noah Wardrip-Fruin from UCSC discussed different types of characters needed for gaming and narrative applications of IVAs. The talk by Jeremy Bailenson from Stanford discussed expressive gestures and how agents orient to one another by modifying their gestural expression in dialogic contexts. Rolf Pfeifer from Zurich discussed how embodiment affects intelligent agents' perceptions and behavior. One of the sessions at IVA 2012 was dedicated to paper presentations focusing on agents in gaming and story-telling environments.

IVA 2012 received 84 submissions. Out of the 74 long-paper submissions, only 17 were accepted for the long-papers track. Furthermore, there were 31 short papers presented in the single-track paper session, and 18 poster papers were on display.

IVA continues to develop and improve the anonymous reviewing process. This year continued the author rebuttal phase begun with IVA 2011, which led to more informed discussion of the papers. The Senior Program Committee was enlarged this year and given a more active role in reviewer recruitment.

Since 2005, IVA has also hosted the Gathering of Animated Lifelike Agents (GALA), a festival to showcase state-of-the-art agents created by student, academic, or industrial research groups. This year, the GALA event was combined with a demo event where participants were also able to demonstrate their latest results.

This year's IVA also included two workshops. One on "Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction" and one focusing on "Real-Time Conversations with Virtual Agents."

There were many people that contributed their time and talent in order to make IVA possible. First, we would like to thank the members of the Senior Program Committee that took on the great responsibility of making sure that the reviewing for papers in their sections was done on time, in a smooth and professional way, with thoughtful and respectful discussion of submitted work. Also, the Program Committee members dedicated significant time and genuine effort to provide thoughtful paper reviews. The contributions of the SPC and PC were essential to assembling a quality program. We also want to thank our keynote speakers, Jeremy Bailenson from Stanford University, Noah Wardrip-Fruin from the University of California, Santa Cruz, and Rolf Pfeifer from the University of Zurich, for crossing domains and sharing their insights with us. The Center for Games and Playable Media at UCSC helped develop our web presence and conference organization. We would also like to thank Jennifer Bloom at UCSC Conference Services for supporting the conference administration.

Of course, IVA 2012 would not have been possible without all the authors, whose contributions extend beyond the creation of intelligent virtual agents to the creation and support of a vibrant research community, fostering our even deeper passion for this challenging field of research.

September 2012

Marilyn Walker
Michael Neff
Ana Paiva
Yukiko Nakano

Organization

Conference Co-chairs

Marilyn Walker	University of California - Santa Cruz, USA
Michael Neff	University of California - Davis, USA
Ana Paiva	INESC-ID and Instituto Superior Tecnico, Lisbon, Portugal
Yukiko I. Nakano	Seikei University, Japan

GALA/Poster and Demo Chair

Arnav Jhala	University of California - Santa Cruz, USA

Workshop Chair

Jean-Claude Martin	LIMSI-CNRS, France

Senior Program Committee

Elisabeth André	Augsburg University, Germany
Ruth Aylett	Heriot-Watt University, UK
Norm Badler	University of Pennsylvania, USA
Tim Bickmore	Northeastern University, USA
Christina 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
James Lester	North Carolina State, USA
Stacy Marsella	University of Southern California, USA
Jean-Claude Martin	LIMSI-CNRS, France
Catherine Pelachaud	CNRS, TELECOM ParisTech, France
Mark Riedl	Georgia Institute of Technology, USA
Hannes Vilhjalmsson	Reykjavík University, Iceland
Michael Young	North Carolina State, USA

Program Committee

Jan Allbeck
Ivon Arroyo
Ryan Baker
Christian Becker-Asano
Kirsten Bergmann
Kristy Boyer
Hendrik Buschmeier
Angelo Cafaro
Marc Cavazza
Morteza Dehghani
Sidney D'Mello
Jens Edlund
Arjan Egges
Birgit Endrass
Friederike Eyssel
Patrick Gebhard
Marco Gillies
Jonathan Gratch
Alexis Heloir
Rania Hodhod
Ian Horswill
Yvonne Jung
Sinhwa Kang
Jihie Kim
Tomoko Koda
Brigitte Krenn
Michael Kriegel
Arjan Kuijper
Chad Lane
Jina Lee
Brian Magerko
Louis-Philippe Morency
Kasia Muldner
Asad Nazir
Radoslaw Niewiadomski
Santiago Ontanon
Jeff Orkin
Sabine Payr
Christopher Peters
Paolo Petta
Laura Pfeifer
Thies Pfeiffer
Ronald Poppe
Rui Prada
David Pynadath
Stefan Rank
Matthias Rehm
Dennis Reidsma
Charles Rich
Laurel Riek
Albert Rilliard
David Roberts
Mercedes Rodrigo
Jon Rowe
Zsofia Ruttkay
Nicolas Sabouret
Daniel Schulman
Magy Seif El-Nasr
Mei Si
Candy Sidner
Nicolas Szilas
Mariiet Theune
Jim Thomas
David Traum
Ning Wang
Langxuan Yin
Jichen Zhu
Amy Ogan
Astrid von der Putten

Reviewers

Alok Baikadi
Ginevra Castellano
Cathy Ennis
Mohamed Gawish
Joseph Grafsgaard
Eunyoung Ha
Hazael Jones
Jennifer Klatt
Seung Lee
Wookhee Min
Chris Mitchell
Stefan Scherer
Sybren A. Stüvel
Weizi Li

Sponsoring Institutions

Center for Games and Playable Media, UCSC

Table of Contents

IVAs for Learning Environments

Emotion and Personality

Evaluation and Empirical Studies (1)

Multimodal Perception and Expression

Narrative and Interactive Applications

Social Interaction

Authoring and Tools

Evaluation and Empirical Studies (2)

Conceptual Frameworks

Poster Abstracts

Fully Automated Generation of Question-Answer Pairs for Scripted Virtual Instruction

Pascal Kuyten[1], Timothy Bickmore[2], Svetlana Stoyanchev[3], Paul Piwek[4], Helmut Prendinger[5], and Mitsuru Ishizuka[1]

[1] Graduate School of Information Science & Technology
The University of Tokyo, Japan
pascal@mi.ci.i.u-tokyo.ac.jp
ishizuka@i.u-tokyo.ac.jp

[2] College of Computer and Information Science
Northeastern University, Boston, Massachusetts, USA
bickmore@ccs.neu.edu

[3] Spoken Language Processing Group,
Department of Computer Science
Columbia University, New York, USA
sstoyanchev@cs.columbia.edu

[4] NLG Group, Centre for Research in Computing
The Open University, Walton Hall, Milton Keynes, UK
p.piwek@open.ac.uk

[5] National Institute of Informatics
Tokyo, Japan
helmut@nii.ac.jp

Abstract. We introduce a novel approach for automatically generating a virtual instructor from textual input only. Our fully implemented system first analyzes the rhetorical structure of the input text and then creates various question-answer pairs using patterns. These patterns have been derived from correlations found between rhetorical structure of monologue texts and question-answer pairs in the corresponding dialogues. A selection of the candidate pairs is verbalized into a diverse collection of question-answer pairs. Finally the system compiles the collection of question-answer pairs into scripts for a virtual instructor. Our end-to-end system presents questions in pre-fixed order and the agent answers them. Our system was evaluated with a group of twenty-four subjects. The evaluation was conducted using three informed consent documents of clinical trials from the domain of colon cancer. Each of the documents was explained by a virtual instructor using 1) text, 2) text and agent monologue, and 3) text and agent performing question-answering. Results show that an agent explaining an informed consent document did not provide significantly better comprehension scores, but did score higher on satisfaction, compared to two control conditions.

Keywords: Dialogue Generation, Rhetorical Structure Theory, Medical Documents.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 1–14, 2012.

1 Introduction and Motivation

Systems for the automatic generation of dialogue scripts have been used primarily to allow teams of computer-animated dialogue agents to present information to an audience [1-3]. In contrast, we use automatically generated dialogue scripts to drive the conversation between a user and a single virtual agent. Our aim is to evaluate this mode of presentation (following up on [4], which evaluated the use of dialogue script generation for presentation by non-interactive teams of agents).

We propose a system which is capable of creating virtual instruction from textual input only, extending previous work [1], into fully automated generation of agent animation scripts from text. In this section we will use text (as in Table 1) from informed consent documents for clinical trials [23] to illustrate the system. First text is translated into rhetorical structure theory (RST) trees (as in Fig. 1), by annotating discourse relations using high-level discourse analysis. RST trees are then translated into question-answer pairs (as in Table 1), by matching patters on the relations and structure of RST trees. Answers are compiled into an animated virtual instructor, using animation scripts. Users are asked to read the question; click an ask-button; and watch the animation (See Fig. 2 for a screenshot of the virtual instructor answering a question).

The paper is organized as follows. This Section continues with an introduction to the theory of text organization. In Section 2, we describe related work; Section 3 is dedicated to our system design, In Section 4 we discuss some design considerations, in Section 5 we describe our evaluation study. In Section 6 we discuss future work and Section 7 contains the conclusions.

Theory of Text Organization. Text can be segmented into non-overlapping, semantically independent units (EDUs) [11]. Between EDUs rhetorical (discourse) relations describe how the more important part (nucleus) and less important part (satellite) relate (e.g. CONTRAST). Text organization can be represented using rhetorical structure theory (RST) trees (as in **Fig. 1**), leaves in RST trees represent EDUs, arrows in the RST tree point from satellite to nucleus, and arrows are labeled with a discourse relation.

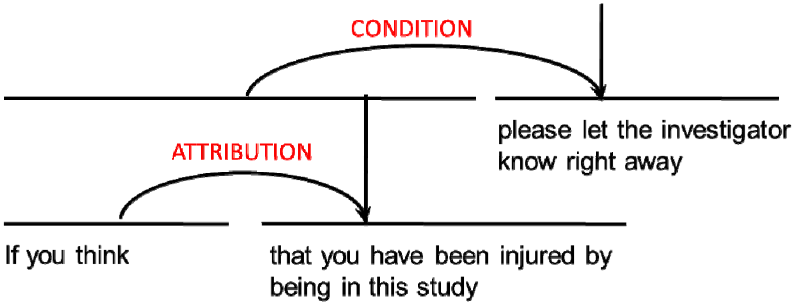

CONDITION
ATTRIBUTION
please let the investigator know right away
If you think
that you have been injured by being in this study

Fig. 1. RST tree, representing the rhetorical structure of text, leaves represent elementary discourse units (EDUs), arrows point from satellite to nucleus, and labels above arrows represents discourse relations

Table 1. Text from an informed consent document for clinical trials [23] and the corresponding question-answer pair generated by our system

Text	Question
If you think that you have been injured by being in this study, please let the investigator know right away.	What if I think that I have been injured by being in this study?
	Answer
	Please let the investigator know right away.

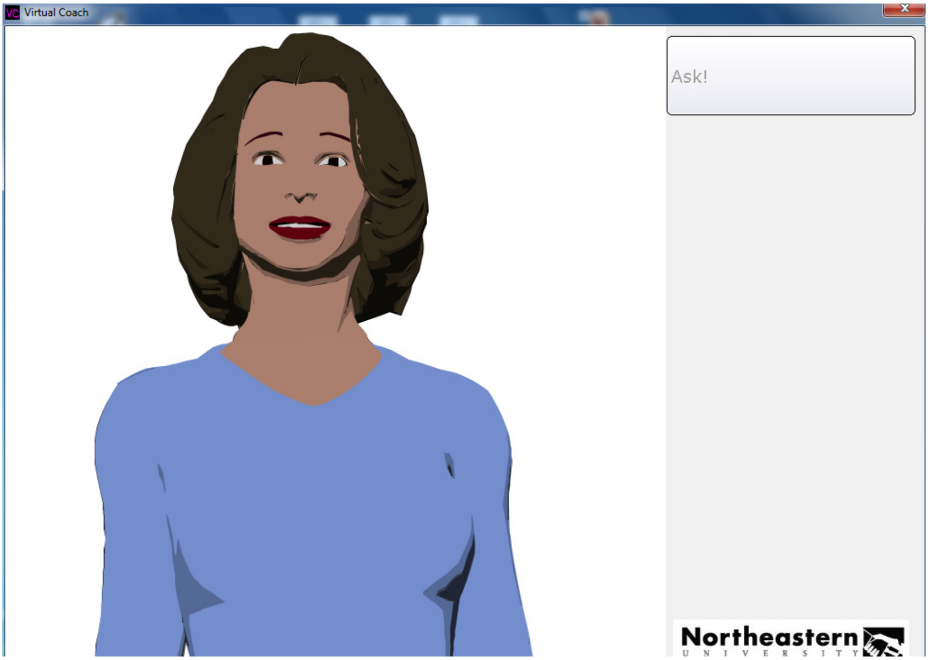


Fig. 2. Screenshot of our virtual instructor answering a question

2 Related Work

The system designed at university of Pennsylvania [5] is similar to our work in that both aim to generate questions from text, using rhetorical analysis. While they use semantic role labeling for analyzing the meaning of the text, our approach is based on support vector machine classifiers for analyzing the discourse structure of the text [6]. When considering question generation at paragraph level the discourse structure of the text becomes important [10].

The aim of the tutor in the project LISTEN is to improve reading comprehension of children [7]. Although both works aim at improving comprehension of the text, their approach is applying semantic role labeling [5] for generating questions instead of discourse analysis and dialogue generation. Further, their generated questions are used as a tool for classification of children self-questioning responses, whereas our generated question-answer pairs are used as input for the virtual instructor.

Cloze question generation is based on syntactical analysis [8], and takes a similar approach as our work. Trees are constructed, patterns are matched and questions are generated. Different from our work, questions are generated by identifying definition phrases. A part of these phrases are replaced with answer-blanks. Users are asked to fill in the answer-blanks by choosing from the removed answer phrase and distractors. Whereas our system is aiming at automatically generating virtual instruction, cloze question generation is aiming at helping second and third grade students to learn new vocabulary.

The twins Ada and Grace are two virtual characters guiding visitors at the museum of Science in Boston [9, 24]. While in our system users get questions presented, in their work visitors can ask the twins questions. Questions asked by visitors are mapped to nearest known questions from a knowledge base containing question-answer pairs. Answers belonging to found questions are presented by the twins. Question-answer pairs from this knowledge base are acquired by a question-answer generator called Question Transducer [10]. Question Transducer identifies factual questions from text by matching patters on the syntactical structure of sentences or paragraphs in the text. Unlike the question-answer pairs of the Question Transducer, our question-answer pairs go beyond paragraph boundaries and can cover larger spans of text (up to the entire text).

A prototype which aims at providing authors of medical texts feedback about their writing style links two systems G-DEE and Greta using XSLT transformations [25]. G-DEE is a document analysis tool capable of automatically detecting importance of recommendation in clinical guidelines uses shallow natural language processing techniques. And Greta, an agent platform supporting detailed non-verbal expressions linked to a TTS.

3 System Design

Our system (illustrated by Fig. 3) generates RST trees from text using high-level discourse analysis. Based on this analysis, question-answer pairs are generated, by translating the RST tree into coherent dialogue. Question-answer pairs are then translated into an agent scripting language. In the final step, scripts are compiled into a run-time agent system (See Fig. 2 for a screenshot of our system).

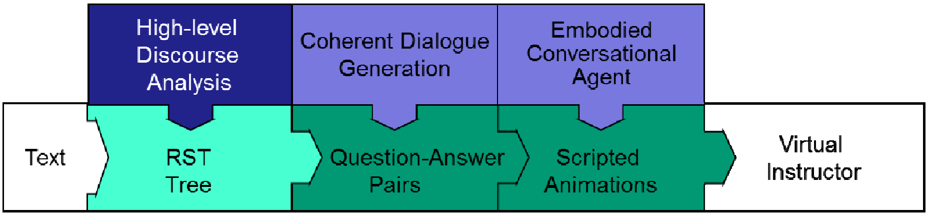

High-level Discourse Analysis
Coherent Dialogue Generation
Embodied Conversational Agent
Text
RST Tree
Question-Answer Pairs
Scripted Animations
Virtual Instructor

Fig. 3. Setup of the system which generates a virtual instructor based on text, fully automated

Data between each module is sequenced using XML-files. Besides some minor annotation of the input text, the overall process is fully automated. Text is annotated for guidance of EDU segmentation during the high-level discourse analysis. Annotation of bulleted lists is manual; annotation of sentence- and paragraph-boundaries is scripted.

High-Level Discourse Analysis. RST trees are generated by the system using a high-level discourse analyzer, called HILDA [6]. The discourse analyzer first segments text into EDUs. Then, (typically) binary discourse relations are identified between EDUs. HILDA is using three classifiers: 1) for EDU segmentation, 2) for discourse labeling and 3) for RST tree construction. HILDA first segments text into EDUs (illustrated by Fig. 4), and then constructs an RST tree (illustrated by Fig. 1). RST trees are constructed in an iterative process: in each step the two most likely adjacent RST sub-trees or EDUs are merged into a new RST sub-tree and labeled with the most likely discourse relation (illustrated by Fig. 5).

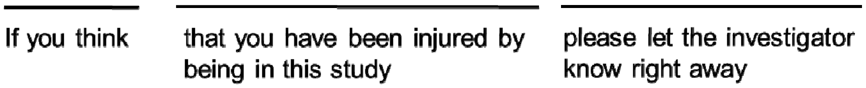

If you think
that you have been injured by being in this study
please let the investigator know right away

Fig. 4. HILDA segments text into EDUs

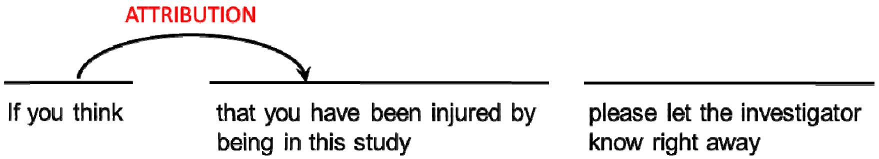

ATTRIBUTION
If you think
that you have been injured by being in this study
please let the investigator know right away

Fig. 5. HILDA merges the most likely adjacent RST sub-trees or EDUs into a new RST sub-tree with the most likely label

Coherent Dialogue Generation. For mapping from RST structure to a dialogue script we use the approach developed in the CODA project [12]. In CODA, a parallel corpus of annotated monologues and dialogues was constructed, where the dialogues express the same information as the aligned monologues. From this, a mapping was inferred from RST structures in monologue to the dialogue act sequences in dialogue. These mapping are used by the CODA system to map an RST tree (such as the one in Fig. 1) to a sequence of dialogue acts (as in Table 1). The input for the CODA system is a sequence of one-level RST trees. It maps this to alternative (ranked) sequences of dialogue acts, and verbalizes the top-ranked sequence. The final output is an XML representation of a dialogue act sequence (usually consisting mostly of question-answer pairs).

Embodied Conversational Agent. The user interface for explaining the document to users was based on an embodied conversational agent system developed for health counseling [13]. In this system, dialogue between a single agent and a user is scripted using a custom hierarchical transition network-based scripting language. Agent nonverbal conversational behavior is generated using BEAT [14], and includes beat (baton) hand gestures and eyebrow raises for emphasis, gaze away behavior for signaling turn-taking, and posture shifts to mark topic boundaries, synchronized with synthesized speech. User input is obtained via multiple choice selections of utterances. The system automatically translates XML representation of question-answer pairs into the agent scripting language for compilation into the run-time system.

4 Design Considerations

Question-answer pairs of our system go beyond paragraph boundaries and can cover larger spans of text (up to the entire text). HILDA generates a single RST tree for the entire text, CODA then maps at various depths discourse relations in this RST tree to a sequence of dialogue acts. If CODA maps a discourse relation at the root of an RST tree, then the question-answer pairs of these dialogue acts cross paragraph boundaries.

A previously conducted case study indicated structural differences between RST trees generated by HILDA and RST trees used for deriving the rule-base of CODA [15]. Some tail EDUs of sentences were merged with the heads of adjacent sentences, causing misalignments in the RST tree. Some discourse relations in the rule-base of CODA were not identified by HILDA. Changes were made to the initial design and configuration of HILDA and CODA, in order to reduce these differences.

Table 2. Question-answer pairs generated by HILDA

Misaligned question-answer pair, based on the traditional implementation of HILDA	
Question What if I think that I have been injured by being in this study?	**Answer** Please let the investigator know right away. If your part in this study takes place at Bohemia Medical Center.
Aligned question answer-pairs, based on the proposed implementation of HILDA	
Questions	**Answers**
What if I think that I have been injured by being in this study?	Please let the investigator know right away.
What if my part in this study takes place at Bohemia Medical Center?	You can get treatment for the injury at Bohemia Medical Center.

Effect of RST Structure on Questions-Answer Pairs. One of the classifiers of HILDA responsible for the structure of RST trees has been trained with features considering RST sub-trees with a maximum span of three EDUs [6]. Because some sentences in text are segmented into more than three EDUs, we expect some of the structural differences identified [15], are caused by the span limitations of the classifier. Take for example an EDU continuing the text (of Table 1): "If your part in this study takes place at Bohemia Medical Center", here HILDA has several options to construct an RST tree.

Traditionally HILDA merges the last EDU of the first sentence with the first EDU of the second sentence (illustrated by Fig. 6). Alternatively HILDA could merge EDUs of the first sentence with EDUs of the second sentence (illustrated by Fig. 7). CODA generates different question-answer pairs based on the two RST trees (listed in Table 2), where the RST tree of the traditional version induces a misalignment. In order to prevent such misalignment we propose a two phase discourse analysis by first merging EDUs within sentences and afterwards merging RST sub-trees.

Effect of Discourse Relations on Patterns of CODA's Rule Base. Not all discourse relations of CODA's rule base can be identified by HILDA. In order to increase the number of rules which CODA can match on RST trees generated by HILDA, we created new rules for CODA's rule base. When all subclasses of a superclass were listed in the rule-base, the superclass was added as well. For example the rule Explain_Init-Complex-InfoReq_Explain matches, among others, on Elaboration-Additional and Elaboration-Obj-Attribute. We extended this with Elaboration, which can be identified by HILDA.

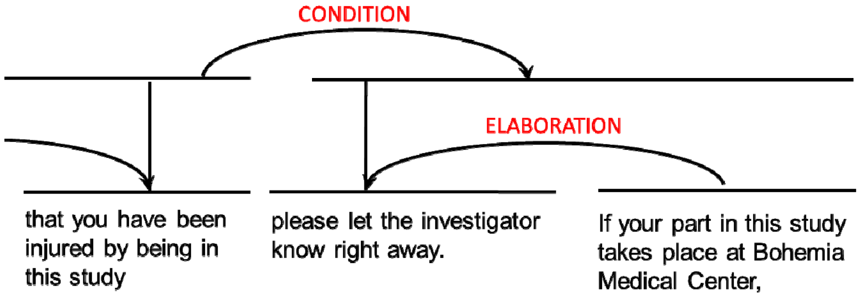

CONDITION
ELABORATION
that you have been injured by being in this study
please let the investigator know right away.
If your part in this study takes place at Bohemia Medical Center,

Fig. 6. Merging the last EDU of the first sentence with the first EDU of the second sentence

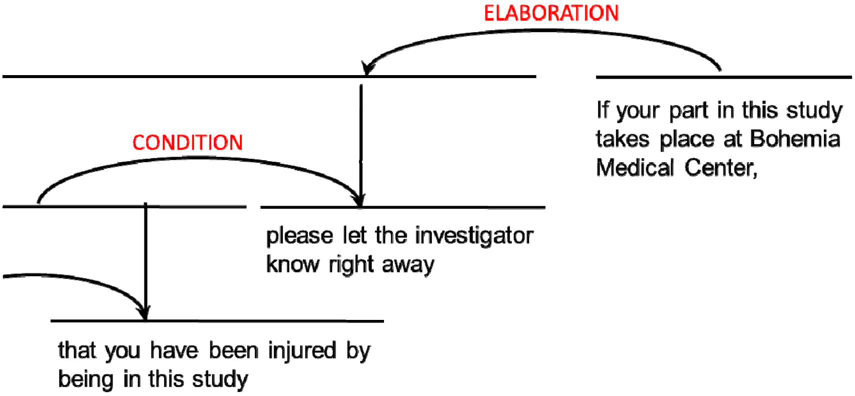

ELABORATION
If your part in this study takes place at Bohemia Medical Center,
CONDITION
please let the investigator know right away
that you have been injured by being in this study

Fig. 7. Merging EDUs of the first sentence with EDUs of the second sentence

5 Evaluation

We conducted an evaluation study to test the effectiveness of our agent-based question-asking system at augmenting explanations of complex text documents. We hypothesized that if a user conducts a question-asking dialogue with an agent about a text, in addition to reading the text, that they will be more cognitively engaged in the material, understand more about it, and be more satisfied with the experience, compared to simply reading the text by itself.

To test this hypothesis, we conducted a 3-arm, counterbalanced, within-subjects experimental study, comparing the question-asking agent (QA) to reading the text (TEXT) and, thirdly a control condition in which the agent read the text (READ), intended to control for exposure time with the agent and hearing the document contents through multiple modalities (text and speech).

The task domain for the experiment is the explanation of research informed consent documents for colonoscopy clinical trials. This domain was selected because the documents contain a wide range of medical and legal terms, facts and concepts that provide a good test for an automated explanation system. Administration of informed consent for clinical trials is often completed without ensuring that participants understand all the terms of the consent agreement, resulting in many potential research subjects signing consent forms that they do not understand [16-18]. In addition, there has been prior work demonstrating some success at having virtual agents explain clinical trial informed consent documents [19, 20]. Colonoscopy is an important area to address: colon cancer is the second leading cause of cancer-related deaths (60,000 deaths each year in the US), and colon screenings have been proven to reduce colon cancer deaths up to 90%, yet compliance with medical recommendations for colonoscopy and other screening is very low. We created three research informed consent documents for this study by taking descriptions of colonoscopy clinical trials [23], adding standard language about research informed consent (from [21] and other sources), and ensuring that the length and complexity was approximately the same across all three.

Our primary hypotheses for the study are:

H1: Users will understand more about documents in the QA condition compared to the TEXT and READ conditions.

H2: Users will be most satisfied with the informed consent process in the QA condition compared to the TEXT and READ conditions.

Measures. Comprehension was assessed by a closed-book knowledge test, consisting of three YES/NO questions (e.g., “Will you be able to choose which of the four bowel preparation medications you will use?”), and three multiple choice questions (e.g., “What risk is associated with ingestion of iodinated oral contrast?”) for each document. Satisfaction was assessed using several single-item, scale response self-report questions, based on the Brief Informed Consent Evaluation Protocol (BICEP) [17], including likelihood to sign the consent document, overall satisfaction with the consent process, and perceived pressure to sign the consent document (Table 3). We also asked single-item scale response questions about satisfaction with the agent, desire to continue working with the agent, and the amount of information provided (from "too little" to "too much").

Table 3. Scale Self Report Measures Used

Measure	Question	Anchor 1	Anchor 2
Satisfaction with Agent	How satisfied are you with the instructor?	Not at all	Very satisfied
Desire to Continue with Agent	How much would you like to continue working with the instructor?	Not at all	Very much
Satisfaction with Experience	How satisfied were you?	Extremely unsatisfied	Extremely satisfied
Amount of Information Provided	How much information did you get?	Too little	Too much
Pressure to Sign	How much pressure did you feel?	No pressure	Extreme pressure
Likely to Sign	How likely would you have been to sign it?	Extremely unlikely	Extremely likely

Participants. A convenience sample of twenty-four subjects, 29% female, aged 28-36, participated in the study. Participants were mostly students (58%), well educated (all had some college), and had high levels of computer literacy (58% described themselves as being "experts").

Experimental Protocol. Verbal informed consent was obtained from study participants, after which they completed a brief demographic questionnaire and were randomized into one of six study sequences defining the order of conditions. We randomized the order in which the study conditions were experienced by each participant while holding the order of presentation of the three documents constant, to counterbalance both order effects and the effects of any particular informed consent document. Participants next completed three rounds of document explanation and filling out comprehension and satisfaction questionnaires. Finally, a semi-structured interview was held with them about their experience and preferences among the three conditions, before they were paid and dismissed.

The study was conducted on a standard desktop computer using a mouse and keyboard for input, and all questionnaires were administered via web forms on the same computer. The entire study was conducted within the Embodied Conversational Agent application interface described in Section 3.3. All agent utterances were accompanied by conversational nonverbal behavior generated using BEAT [22].

In the TEXT condition, the agent walked on the screen and said "Hi, I am Karen. I am going to explain an informed consent document to you for a clinical trial." After the user clicked "OK, let's get started!", the first page of the document filled the screen, and the user was allowed to read it until they clicked a "I'm through reading this." button, at which point the next page of the document was displayed. When the last page of the document had been read, a message was displayed on the screen informing the participant that the session was over.

The READ condition was identical to TEXT, except that after each page of the document was displayed, the agent re-appeared and read the page to the participant in an uninterruptable monologue.

The QA condition was also identical to TEXT, except that after each page of the document was displayed, the agent re-appeared and engaged the user in a question-and-answer dialogue, as generated by the system described in Section 3. Question-and-answer pairs were delivered in sequence. For each, the question was displayed in text on the screen and the user could push an “Ask!” button, after which the agent re-appeared and delivered the answer.

Quantitative Results. We conducted repeated-measures ANOVAs for all self-report measures, knowledge test scores, and session duration, in SPSS. Table 1 shows descriptive statistics for the outcome measures.

Table 4. Study Results (mean and (SD))

	TEXT	**READ**	**QA**	**p**
Session Duration (seconds)	505 (251)	1081 (249)	1011 (247)	<.001
Comprehension	77% (21%)	69% (28%)	76% (22%)	n.s.
Satisfaction with Agent	3.83 (1.88)	3.96 (1.69)	4.35 (1.75)	n.s.
Desire to Continue with Agent	3.70 (1.82)	3.73 (1.83)	4.30 (1.94)	n.s.
Satisfaction with Experience	4.09 (1.47)	3.83 (1.83)	4.39 (1.92)	n.s.
Amount of Information Provided	4.35 (1.27)	3.38 (1.34)	3.96 (1.07)	.07
Pressure to Sign	2.35 (1.23)	2.52 (1.65)	2.22 (1.28)	n.s.
Likely to Sign	4.13 (1.58)	3.78 (1.70)	4.22 (1.81)	n.s.

There was a significant effect of condition on session duration, with TEXT taking significantly less time F(2,22)=31.7, p<.001. There was a trend for participants to rate the TEXT condition as providing too much information compared to either the READ or QA conditions, F(2,44)=2.80, p=.07. No other significant differences were found, although the various satisfaction measures were all trending with QA rated more highly than the other conditions.

The only significant order effect found was that the Amount of Information Provided was rated as increasing session by session, F(2,42)=3.32, p<.05. This could be due to actual differences in the information content of the three documents, or effects of user fatigue. There were a few significant differences by gender, with females more satisfied with the agent compared to males.

Qualitative Results. Semi-structured interviews with 23 participants were transcribed and coded for common themes. When asked for their overall impressions, several participants volunteered that they liked the concept of an agent explaining a document:

- “It’s easier to remember if presented as conversation.”
- “The avatar helped me to concentrate.”
- “A great way to remember.”

Although several others felt uncomfortable with the system:

- "I'm uncomfortable to be explained by an animated character/avatar. I would prefer a human being."
- "I prefer to read instead of listening."

Many participants also volunteered that they liked the question-asking interaction:

- "By asking questions, I am able to get info you need without unnecessary information."
- "The questions did help."
- "Enjoyed question answering, although weird."

Many participants had suggestions for improving the question-asking interaction with the agent. One of the most common suggestions was to make it more interactive, by allowing users to select their questions from a menu:

- "There is no additional values of having an animated character/avatar if there is not much interactivity with avatar, because the avatar was just reading/repeating the content which I read already and the questions were preset."
- "I want to decide my own rhythm."

Other suggestions included grouping the questions by relevance, and displaying or including a question when giving the answer, to provide better context.

When asked which of the three explanation methods was most informative, 15 (65%) expressed a clear preference for the QA condition, with an additional 2 indicating a tie between QA and READ:

- "The [QA] conversation is easier to remember due to highlights."
- "It had interaction, which I liked."
- "Because she was answering questions instead of just reading."
- "[QA] had the best presentation, and [READ] had the best content."

When asked which of the three methods they would like to actually use for informed consent, 12 participants expressed a preference for QA, and an additional three said that either QA or another method would be alright.

- "When I read it I may not know what is important. The questions highlight to me what is important."
- "I felt less pressure to understand the document."
- "I got a little bit lost with the questions, some I would not ask, therefore I lost track."

Discussion. We found generally positive acceptance of the question-asking system by participants. We did not find any support for H1, regarding improved document comprehension with QA compared to the control conditions. However, we did find partial support for H2, with a majority of participants interviewed stating that the QA condition was the most informative compared to the other presentation methods, and all quantitative satisfaction measures trending with QA as the most preferred (although the differences are not significant). This result mirrors results from a previous study in which an agent explaining an informed consent document did not provide significantly better comprehension scores, but did score higher on satisfaction, compared to two control conditions [19].

Lack of support for H1 could be attributed to several factors. Participants were mostly students, who are used to receive information from documents. Further, we presented all questions in pre-fixed order, which may have led to lower engagement. The lower comprehension results for the READ condition, when compared to the TEXT condition, could be attributed to the quality of the TTS or presentation style, e.g. while reading the text to our participants we did not provide any subtitles.

6 Future Work

More Specific Relations. Not all discourse relations of CODA's rule-base can be identified by HILDA, and not all discourse relations identified by HILDA exist in CODA's rule-base. Therefor we are planning to study whether it is possible to improve HILDA's performance for specific domains, in particular for CODA's domain. We are planning to study whether we could improve the output of the overall system, when HILDA has been trained to identify a subset of CODA's discourse relations. Improved output could be measured in terms of more diverse question-answer pairs or increased quality of the question-answer pairs.

Variable Length of Question-Answer Pairs. Question-answer pairs generated by our system vary in length, because CODA maps discourse relations at various depths in the RST tree. When discourse relations are matched at the root of RST trees, the generated question-answer pairs will be long. And when discourse relations are matched near the leaves of the RST tree, the generated question-answer pairs will be short. Besides, when RST trees are unbalanced, questions will differ in length from answers. A couple participants of our evaluation study noted the variable sizes of the question-answer pairs and stated they preferred shorter answers. The preferable length of the question and answer may depend on the application of the system. Therefore, we are planning to study whether it is possible to guide the RST tree construction of HILDA, as well as the pattern matching of CODA in order to generate questions and answers of specific length.

Improved Comprehension and Satisfaction. There are several aspects we can explore to improve comprehension and satisfaction of our users. In a future evaluation study, we could investigate different presentation styles. We could allow users to select questions of their interest and let them intervene during the answering. We could also present the question-answer pairs as a dialogue between two agents. Finally, we could use a different TTS and presentation style of answering questions. We could highlight important aspects of the document while answering a question or add subtitles when the instructor is answering a question.

7 Conclusions

We introduced a novel approach for automatically generating a virtual instructor from textual input only. We described the system design, consisting of high-level discourse analysis, coherent dialogue generation and embodied conversational agent scripting. Furthermore, we discussed some design considerations in order to reduce structural differences found in a previous case study. Finally we conducted an evaluation study to test the effectiveness of our agent-based question-asking system at augmenting explanations of complex text documents. Results show that an agent explaining an informed consent document did not provide significantly better comprehension scores, but did score higher on satisfaction, compared to two control conditions.

Acknowledgements. This work was initiated while Dr. Svetlana Stoyanchev was a visitor of NII and was extended while Pascal Kuyten was a visitor of OU. The experiment was conducted while Dr. Tim Bickmore was a visitor of NII. This work was supported, in part, by a grant from the US National Cancer Institute (NIH grant R01 CA158219), by the NII Grant Challenge Project and the Japanese Government Scholarship program (文部科学省奨学金).

References

1. Piwek, P., Hernault, H., Prendinger, H., Ishizuka, M.: T2D: Generating Dialogues Between Virtual Agents Automatically from Text. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 161–174. Springer, Heidelberg (2007)
2. André, E., Rist, T., van Mulken, S., Klesen, M., Baldes, S.: The Automated Design of Believable Dialogues for Animated Presentation Teams. In: Embodied Conversational Agents. MIT Press, Cambridge (2000)
3. Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schroeder, M., Baumann, S.: Fully Generated Scripted Dialogue for Embodied Agents. Artificial Intelligence Journal 172(10), 1219–1244 (2008)
4. Stoyanchev, S., Piwek, P., Prendinger, H.: Comparing Modes of Information Presentation: Text versus ECA and Single versus Two ECAs. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 377–383. Springer, Heidelberg (2011)
5. Mannem, P., Prasad, R.P., Joshi, A.: Question Generation from Paragraphs at UPenn. In: Proc. of QG 2010: The Third Workshop on Question Generation, Pitsburg, PA, pp. 84–91 (2010)
6. Hernault, H., Prendinger, H.: duVerle, D. and Ishizuka, M. HILDA: A Discourse Parser Using Support Vector Machine Classification. Dialogue and Discourse 1(3), 1–33 (2010)
7. Chen, W., Mostow, J., Aist, G.: Using Automatic Question Generation to Evaluate Questions Generated by Children. In: Question Generation: Papers from the 2011 AAAI Fall Symposium (2011)
8. Gates, D., Aist, G., Mostow, J., McKeown, M., Bey, J.: How to Generate Cloze Questions from Definitions: A Syntactic Approach. In: Question Generation: Papers from the 2011 AAAI Fall Symposium (2011)

9. Nouri, E., Artstein, R., Leuski, A., Traum, D.: Augmenting Conversational Characters with Generated Question-Answer Pairs. In: Question Generation: Papers from the 2011 AAAI Fall Symposium (2011)
10. Heilman, M.: Automatic Factual Question Generation from Text. Ph.D. Dissertation, Carnegie Mellon University, CMU-LTI-11-004CMU-LTI-09-013 (2011)
11. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3), 243–281 (1988)
12. Piwek, P., Stoyanchev, S.: Data-oriented monologue-to-dialogue generation. In: Proc. Of 14th Annual Meeting of ACL, Short Papers, Portland, Oregon, pp. 242–247 (2011)
13. Bickmore, T., Picard, R.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer Human Interaction 12, 293–327 (2005)
14. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT. In: Conference BEAT: The Behavior Expression Animation Toolkit, pp. 477–486 (2001)
15. Kuyten, P., Hernault, H., Prendinger, H., Ishizuka, M.: Evaluating HILDA in the CODA Project: A Case Study in Question Generation Using Automatic Discourse Analysis. In: Question Generation: Papers from the 2011 AAAI Fall Symposium (2011)
16. Wogalter, M.S., Howe, J.E., Sifuentes, A.H., Luginbuhl, J.: On the adequacy of legal documents: factors that influence informed consent. Ergonomics 42, 593–613 (1999)
17. Sugarman, J., Lavori, P.W., Boeger, M., Cain, C., Edson, R., Morrison, V., Yeh, S.S.: Evaluating the quality of informed consent. Clinical Trials 2, 34 (2005)
18. Sudore, R., Landefeld, C., Williams, B., Barnes, D., Lindquist, K., Schillinger, D.: *Use of a modified informed consent process among vulnerable patients: a descriptive study. J. Gen. Intern. Med. 21, 867–873 (2006)
19. Bickmore, T., Pfeifer, L., Paasche-Orlow, M.: Using Computer Agents to Explain Medical Documents to Patients with Low Health Literacy. Patient Education and Counseling 75, 315–320 (2009)
20. Fernando, R.: Automated Explanation of Research Informed Consent by Embodied Conversational Agents. College of Computer and Information Science. MS. Northeastern University, Boston (2009)
21. `http://ccr.coriell.org/Sections/Support/NIGMS/Model.aspx?PgId=216`
22. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: Conference BEAT: The Behavior Expression Animation Toolkit, pp. 477–486 (2001)
23. `http://nci.nih.gov/clinicaltrials`
24. Yao, X., Tosch, E., Chen, G., Nouri, E., Artstein, R., Leuski, A., Sagae, K., Traum, D.: Creating conversational characters using question generation tools. Dialogue and Discourse 3(2), 125–146 (2012)
25. Georg, G., Cavazza, M., Pelachaud, C.: Visualizing the Importance of Medical Recommendations with Conversational Agents. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 380–393. Springer, Heidelberg (2008), doi:10.1007/978-3-540-85483-8_39

The Virtual Apprentice

Weizi Li and Jan M. Allbeck

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

Abstract. Over the past couple of decades, virtual humans have been attracting more and more attention. Many applications including, video games, movies, and various training and tutoring systems have benefited from work in this area. While the visual quality of virtual agents has improved dramatically, their intelligence and socialization still needs improvement. In this paper, we present work towards endowing agents with social roles and exploiting Explanation-Based Learning (EBL) to enable them to acquire additional, contextual behaviors from other agents. These virtual humans are capable of learning and applying role related actions from multiple agents and only adopt behaviors that have been explained to them, meaning that their definition of a role may be a subset from one or more agents. This results in emergent behaviors in heterogeneous populations.

Keywords: Virtual Humans, Social Roles, Explanation-Based Learning.

1 Introduction

Virtual humans have increasingly attracted attention over the last decade. Many applications including games, movies, urban and transportation planning systems, and training and tutoring simulators are prospering due in part to this burgeoning technology. Observing the potential, researchers from various disciplines have invested tremendous effort to improve these artificial lives' visual quality and life-like behaviors. While many impressive strides have been made, we believe virtual humans can be further elevated in two respects: incorporation of social roles and inclusion of learning and evolving abilities.

We argue that it's important to include social roles into virtual humans for several reasons. First, roles can better organize agent behaviors and demonstrate agent internal attributes such as goals and duties. In the real world, we usually have multiple roles, and our behaviors, goals and obligations are heavily associated with each of them. Thus, for virtual agents, roles are an ideal tool for governing various behaviors and their incentives. Secondly, for longer duration simulations in which agents are continuously learning and evolving in order to perform long-term tasks, roles are needed to improve consistency, believability,

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 15–27, 2012.

and reasonableness. Third, certain domain-specific simulations, for example military and medical scenarios, can be only realized with the presence of social roles such as commanders, soldiers, civilians, doctors and patients. In the virtual human and animation research community, though roles have been assigned to characters in many applications and various planning algorithms have been developed to control agent behaviors, more studies are needed to show how roles could be adopted and the interrelationship between roles and behaviors.

The ability to learn has been successfully applied to software agents, robots and virtual characters, particularly in applications for communicating with real humans. However, learning between virtual characters still needs further exploration. For example, we can imagine Non-Player Characters (NPCs) in a game learning strategies for combating player actions from each other, creating a reasonable and increasingly challenging evolution of game play. Furthermore, the behavior selection mechanisms for virtual characters are often hand crafted and static. The knowledge base of the agents is assumed to be complete, resulting in agents that lack the ability to demonstrate emergent behaviors. This limits their use in the current and future applications. Virtual humans capable of developing contextual behaviors will facilitate longer, more compelling simulations and games.

The purpose of this work is trying to partially fill the gap and advance behavioral animation by endowing virtual humans with social roles and the ability to learn. Specifically, we address role adopting phenomenon via learning by observation and explanation. We exploit an Explanation-Based Learning (EBL) mechanism, allowing the agent, through their observation and other agents' explanations, to acquire new knowledge and concepts based on prior knowledge, and eventually be capable of adopting new roles. In order to better actualize our idea, we also introduce semantics to our virtual world by organizing all objects and their features into an ontology. Additionally, for illustrating the effectiveness of our approach, we describe an example in which the resulting agents demonstrate more reasonable interactions and behaviors more consistent with their environment and roles.

2 Related Work

In order to create believable agent behaviors, numerous efforts have invested in developing sophisticated behavior selection mechanisms and simulating an agent's decision-making process. Some researchers have explored computational models such as decision-networks [33] and fuzzy logic [15]. Others have demonstrated the use of various social-psychology factors such as in [24,19] or the BDI architecture [26]. In addition, a great deal of work has addressed this problem using cognitive approaches, for example [12,23,13]. Similar to ours, there exist several works that exploit semantics to facilitate agent behaviors. For example, [11] annotates the virtual environment with information to support navigation of the agents and their interactions with the objects. Chang et al [5] and Kao et al [17] integrate semantics into the agent planning and reasoning processes.

Though generating significant results, for the most part, these research efforts assume the agent knowledge base is complete. In other words, the number of agent behaviors is fixed. Thus agents are not only prohibited from learning and evolving, but also cannot demonstrate emergent behaviors.

While few above mentioned works incorporate social roles, a virtual train station with pedestrian performing different roles is described in [28]. Grimaldo et al [14] simulate a virtual university bar with agents acting in two roles: waiter and customer. Pelechano et al. include roles among other factors to simulate an evacuation scenario [24]. In our previous work [18], we have also simulated social roles and explored the idea of role switching. However, we did not include an ability to learn and evolve individual definitions of roles. Another group of work includes using social roles to communicate with real humans (e.g. [16,32]) in serving training and tutoring scenarios [29]. Although these works have assigned agents social roles, the agents are still lacking an ability to learn and few utilize roles as a tool to organize agent internal attributes and behaviors.

In contrast, there are several works that have endowed their virtual characters with a learning ability. For example, Blumberg et al [3] integrated learning activity to a synthetic character. Orkin et al designed a restaurant game [22] in which they collect behavior and dialog from real human players and later apply them to virtual characters to enhance a gaming experience. The idea of learning from observation in [8] is similar to ours, but their approach addresses interactions between software agents and real world experts. Still other work addresses learning activity entirely within a virtual or simulated environment. Cohen et al [6] built a learning baby with sensorimotor interactions in a simulated environment and Conde et al [7] use reinforcement learning to allow characters to find their path to goal locations. However, the learning activities in these works are not among virtual characters but between characters and the environment. To summarize, having virtual humans learn from real humans often requires a lot of effort from the human participants. Also, enabling virtual characters to learn from each other in addition to the environment will further enhance their behaviors. A virtual human population that evolves more autonomously is both more natural and less demanding of simulation authors.

3 Semantic Virtual World

Semantics can be very helpful when constructing operable virtual worlds. They can be used to better organize knowledge of the environment such as objects and their features. They can also facilitate agent-object and agent-agent interactions by endowing agents with the ability to retrieve corresponding features and information efficiently. In addition, by separating environment characteristics from agent stories, object entities become scenario-independent and can be easily applied to other environments without massive modifications.

We have adopted an ontology to hold virtual world semantics. The ontology consists of hierarchical classes, properties, and relations between instances of the classes. An instance is the child of at least one class and is described by properties which link it to various values, including numbers, strings, and other instances and classes. All objects and their features are stored and updated in the ontology. The object features include geometry, color, status, and spatial information. A small part of our ontology is shown in Fig. 1. Later in the paper we will show how the ontology greatly assists agents in reasoning about the virtual world and learning new knowledge and concepts.

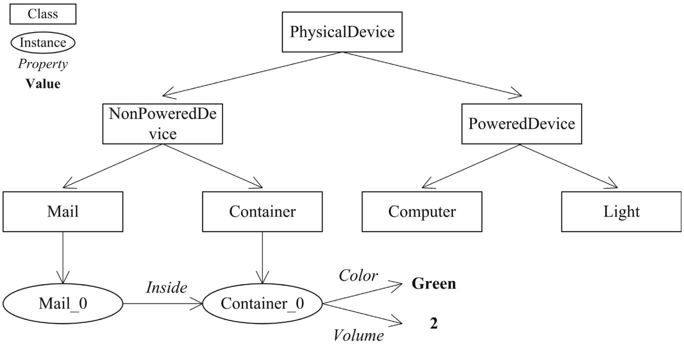


Fig. 1. Partial ontology of object entities

4 Intelligent Social Agents

In this section, we will first provide a definition of social roles extracted from socio-psychology literature and then explain in more detail our learning strategy.

4.1 Social Roles

According to [1], a role is *the rights, obligations, and expected behavior patterns associated with a particular social status*. In Stark's textbook, *Sociology*, he indicates that roles can be achieved or assigned by someone else [30]. Furthermore, they can be semi-permanent, such as having an occupation, or they can be transitory, such as being a patient. Ellenson's work [10] points out that each person could play a number of roles, or in other terms, engage in a *role set*. With these definitions and descriptions, and also by taking into account discussions from other social-psychology work [2,20], we conclude that roles are patterns of behaviors for given situations or circumstances. They can be achieved, assigned and abandoned, having various durations and are often associated with social relationships. Furthermore, multiple roles can be possessed by an individual at the same time. Given this summation, agent will switch from one role to another by performing characteristic behaviors of the latter role. For example, a *Trainee* could switch to *Administrator* by performing its feature behaviors such as *Post flyers* and *Organize professor mail* which could be freely defined by users. For more extensive discussion about role switching phenomenon, we refer readers to our previous paper [18].

What's more, people can have their own definition and expectations of a single role. For example, being a professor to some may include both *research* and *teaching*, while for others conducting either *research* or *teaching* solely is considered enough to have that role. Because each role consists of several characteristic behaviors, obligations, and duties and each individual can have their own definition and expectations for a certain role, our society is colorful and diverse. To create more virtual human heterogeneity, we also allow our agents to have different definitions of various roles.

Lastly, roles can be influenced and constrained by many factors, such as biology or genetics. For instance, a female is unlikely to take on the role of father, and some athletes and musicians seem genetically predisposed to excel at those roles. This implies that certain roles, or more specifically, certain behaviors have physical and intellectual prerequisites. In this work, we assume that agents attain a set of prerequisites for performing elementary actions such as talking, nodding, carrying, picking up, and also possess the ability to learn.

4.2 Learning Strategy

Nearly since the birth of the computer, various learning methods have been developed and used to solve real world problems effectively and efficiently. While this powerful tool, learning, has been successfully utilized in building software agents, robots and more, its usage in simulating interactions between autonomous virtual agents residing in virtual worlds has not been fully explored. Currently many applications using virtual humans, such as video games and training simulators, adopt scripted behaviors and lots of "if - then" rules. This not only inhibits the virtual characters ability to learn and evolve, it also makes the configuration of simulations laborious, since for each different scenario, a mound of extra rules need to be designed and included. In this work, we are trying to partially resolve this problem by equipping our virtual agents with an ability to learn and allowing them to learn from each other. To proceed, we would like to point out that our goal of incorporating a learning mechanism differs from more conventional applications. Traditionally, as we have mentioned, learning methods are used to solve certain tasks more efficiently and accurately. Here, since we are simulating virtual humans and they are expected to behave, reason and learn like real humans, our goal in including a learning method is to generate more reasonable simulations and enhance behavioral animation. This distinction also explains why we adopt a specific learning method rather than just copying knowledge from one agent to another.

The learning phenomenon of real humans is very complex and still under discussion in terms of its exact form and process. Nevertheless, it is widely believed that it involves certain approaches such as explanation-based learning, analogical learning, instance-based learning and reinforcement learning. Also we know that under certain conditions a particular learning method is favored over the others. Given this problem is extremely sophisticated, we are not trying to simulate all aspects of real human learning activity but concentrate on learning by observation and explanation. To achieve this, we have adopted Explanation-Based Learning (EBL). EBL is an analytical learning method. Based on prior knowledge, observation, explanation, and expanded information provided by training examples, new knowledge and concepts can be learned [9,21,31,27]. While we acknowledge not all skills and concepts can be acquired through observation and explanation, in many cases we do obtain knowledge in this fashion. For example, imagine you are a trainee who is going to work in an office environment. At the beginning, you probably need to learn various duties from observation and your supervisor's explanation. In other social settings, for instance, traveling to a different country, when you are learning the local culture and manners, most likely the learning method is also observation and explanation. For acquiring this kind of knowledge and concepts, other learning methods seem less plausible. To be specific, we do not have abundant examples needed for inductive learning methods such as decision tree learning and neural networks or possess many similar examples we can compare with in order to carry out instance-based learning or face situations fulfilled with probabilities that Bayesian networks could manage or attain direct and/or indirect feedback as a training source for reinforcement learning. With these considerations and after taking several other learning methods into account, we find EBL is the most plausible and effective approach.

In general, EBL includes the following components (for a more thorough discussion, we refer readers to [9,21]): *Goal Concept*, a target concept with a set of relevant features; *Training Example*, a typical positive example of a concept to be learned; *Domain Theory*, prior knowledge which can be used to analyze or explain why the training example could satisfy the goal concept; And finally a *Learned Rule*. As one may notice, one of the keys to this approach is prior knowledge assignment. We need to determine what kind of knowledge should be given to our agents in order to achieve generality and scenario-independence. To address this, we have found some psychology studies showing that babies are born with physical and spatial reasoning [4] and language acquiring abilities [25]. Even though there are no conclusions about which abilities are innate, given that we are simulating normal intellectual and physical level adult-like agents, we believe it is reasonable to give them at least following three categories of base knowledge while still preserving the generality:

- Color: $Red, Green, Blue, Yellow, Cyan \ldots$
- Spatial Relationship: $Inside, Outside, Above, Below, \ldots$
- Common Object Type (lowest level of our ontology class): $Mail, Container, Computer, \ldots$

Finally, we need to mention that one premise that needs to be met in order to successfully perform EBL is that all prior knowledge has to be correct. This premise is to ensure that all further inferences drawn would also be correct. However, we believe in virtual humans simulation this criteria can be loosen since it is reasonable and natural for one to have false knowledge and later draw

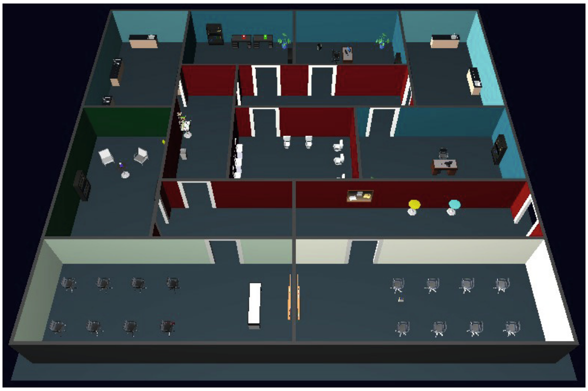


Fig. 2. A school environment

false inferences. Actually this might accentuate the imperfect nature of human behaviors in our virtual humans.

5 Implementation and Example

In this section, we will detail our implementation of the learning and role adopting process through an extended example. Most commonly when using EBL, agent learning is through observation and explanation of a series of actions performed by real world experts. Since we are aiming to create a purely autonomous world, we adapt the term observation and explanation to indicate such behaviors occurring between virtual humans. One trainee can observe other agents' actions and these agents can explain their current action series. In this fashion, the trainee learn new knowledge and concepts.

As an example scenario, we have created the school environment shown in Fig. 2. This example includes three agents. One takes the role of *Trainee*, while other two become an *Administrator* and a *Housekeeper*. The goal of this simulation is to teach the trainee several duties associated with being an administrator and a housekeeper such that he will eventually be capable of adopting these two roles.

In this particular example, we form the duties of an *Administrator* as *Organize professor mail*, *Post flyers* and *Fill paper for office equipments* while a *Housekeeper*'s duties include *Check classroom* and *Water plant*. In order to successfully carry out these duties, the trainee must first learn several new concepts. For example, the place for storing a professor's mail needs to be known when performing *Organize professor mail*. The condition of a plant needs to be considered before performing *Water plant*. Here we use the former case to illustrate the learning process. Assume in this scenario we have two professors, namely *ProfA* and *ProfB*, and the trainee is learning from his trainer, the current administrator, where *ProfA*'s mail is. In this context, we put the learning task in EBL form as following:

- Goal Concept: $MailToProfA(x)$
- Training Example: A positive example, $MailToProfA(Obj1)$
  $Inside(Obj1, Office\_1)$
  $Inside(Obj1, Obj2)$
  $Type(Obj1, Mail)$
  $Type(Obj2, Container)$
  $Color(Obj1, White)$
  $Color(Obj2, Red)$
  ...
- Domain Theory:
  $MailToProfA(x) \leftarrow Location(x, Office\_1) \wedge Inside(x, y) \wedge Type(x, Mail) \wedge Type(y, Container) \wedge Color(y, Red)$
  $Location(x, Office\_1) \leftarrow Inside(x, Office\_1)$
  ...
- Learned Rule:
  $MailToProfA(x) \leftarrow Inside(x, Office\_1) \wedge Inside(x, y) \wedge Type(x, Mail) \wedge Type(y, Container) \wedge Color(y, Red)$

The final learned rule states "*Mail* x is for *ProfA* if x is inside *Office_1* and also inside y which is a *Container* and has the color *Red*". With this newly learned knowledge, the trainee can perform behaviors such as "Transfer *ProfA*'s mail to his container" and "Retrieve *ProfA*'s mail from his container".

The detailed implementation of this learning task is as following. First of all, we have two actions *Observe* and *Explain* associated with the trainee and administrator, respectively. Then, we put the positive example and the domain theory into a database and assign the positive example with a boolean value initially set to 0 indicating the current training status. Once the administrator starts to *Explain* and the trainee starts to *Observe*, an underlying recursive algorithm will begin. In each iteration, the algorithm will attempt to match and prove each concept in the domain theory by using facts in the positive example. The whole procedure will continue until the goal concept is proved. To be specific, for above example, when the procedure begins, the algorithm will first try to prove the concept $MailToProfA(Obj1)$. However, this will fail since the concept $Location(x, Office\_1)$ is not in prior knowledge (i.e. Color, Spatial Relationship and Common Object Type). Then the algorithm will continue to try to prove next concept $Location(x, Office\_1)$ and this time it will succeed because the fact $Inside(x, Office\_1)$ of the positive example is in the prior knowledge. After $Location(x, Office\_1)$ has been proved, the goal concept $MailToProfA(Obj1)$ will also be proved and the whole procedure will complete. At this point, the training status of the positive example switches from 0 to 1 indicating the *Explain* process and also the *Observe* process of the positive example are over. When these two processes are finished, the last step, "generalization" will begin. This step essentially replaces instances in proved positive example with variables. In general, since all the instances are organized in an ontology as we mentioned in Section 3, this replacement is simply climbing the object hierarchy. However, this procedure is subject to one piece of prior knowledge which is the Common Object Type. In the beginning of "generalization", the proved positive example would have following form: $MailToProfA(Obj1) \leftarrow Inside(Obj1, Office\_1) \wedge Inside(Obj1, Obj2) \wedge Type(Obj1, File) \wedge Type(Obj2, Container) \wedge Color(Obj2, Red)$, here since the $Obj1$ and $Obj2$ both have a type specified, *Mail* and *Container*, these two instances can climb object hierarchy only to their types. In contrast, if an instance in some rules does not have a type specified then it can climb the object hierarchy all the way to "PhysicalDevice" according to Fig. 1. With this, the whole learning process is considered complete and the knowledge has been added to the knowledge base of the trainee. From this example, we can also see an advantage of EBL, which is it filters non-relevant object features when forming the final learned rule, such as $Color(Obj1, White)$. This is similar to real world cases, where an object can have multiple features, but we only need to know some features for some tasks. Other concepts can be learned in a similar vein, for simplicity we only list the final learned rules:

- $MailToProfB(x) \leftarrow Inside(x, Office\_1) \wedge Inside(x, y) \wedge Type(x, Mail) \wedge Type(y, Container) \wedge Color(y, Green)$
- $ReadyForTransfer(x) \leftarrow Inside(x, Hallway\_1) \wedge Above(x, y) \wedge Type(y, Table) \wedge Color(y, Cyan)$
- $ReadyForPost(x) \leftarrow Inside(x, Hallway\_1) \wedge Above(x, y) \wedge Type(y, Table) \wedge Color(y, Yellow)$
- $ItemLeftInClassroom(x) \leftarrow Inside(x, Classroom) \wedge \neg Type(x, WhiteBoard) \wedge \neg Type(x, LectureDesk) \wedge \neg Type(x, StudentChair)$
- $NeedWatering(x) \leftarrow Type(x, Plant) \wedge Color(x, Yellow)$

Our approach is convenient and efficient. While this example had only a few rules, more can be added to the system just by providing positive examples and the corresponding domain theories in the database. Although the underlying algorithm is recursive, processing only applies to single examples and their domain theory which is succinct and independent. Given this, including more rules, tasks, objects, and roles does not dramatically increase the computational cost.

In addition, as we discussed in Section 4.1, every individual can have their own definition of a specific role, which contributes variety and diversity to our society. Here, we apply this feature for the purpose of generating heterogeneous virtual humans. To provide an illustration, in our example, *Trainee* will learn all duties of being an *Administrator* except *Fill paper for office equipment*. Alternatively, the trainee can learn partial duties from multiple administrators, but not an entire set from either. This will result in the trainee having a definition of *Administrator* that is consistent with, but different from others in the world. An agent's definition of a role can evolve over time as more tasks are observed and explained. The actions of the agent while in that role will then also evolve to correspond with the changing role definition. Both the learning and performing procedures are demonstrating in Fig. 3 and Fig. 4, respectively.

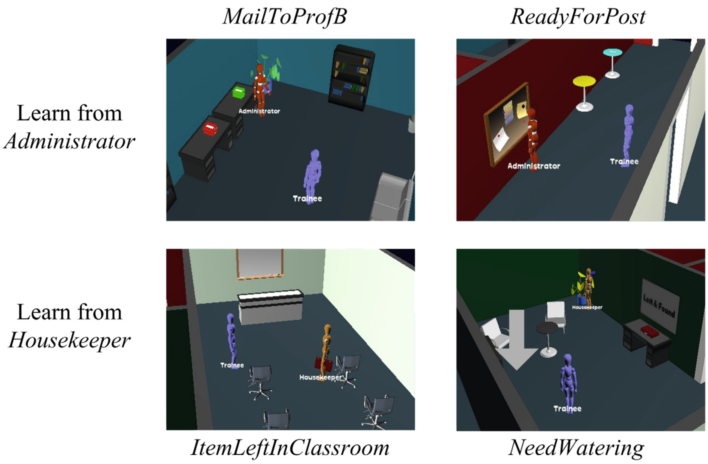


Fig. 3. Trainee learned concepts: *MailToProfB*, *ReadyForPost*, *ItemLeftInClassroom* and *NeedWatering*

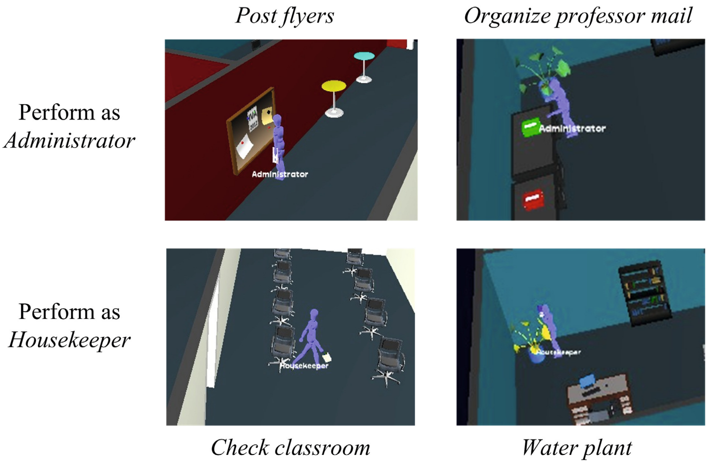


Fig. 4. Trainee adopts the role *Administrator* and *Housekeeper* and is performing the corresponding duties

6 Conclusion and Future Work

Our ultimate goal is to be able to simulate populations of virtual humans over extended periods of time with reasonable behaviors that are appropriate to the context and evolve as the agents gain knowledge and experience just as they do in the real world. In this paper we have presented a method that uses roles and Explanation-Based Learning (EBL) to organize the agent behaviors and enable emergent behaviors. Furthermore, our approach is designed such that the learning is general and scenario independent. Agents could learn definitions of roles in one scenario and apply them in completely different scenarios. While the behaviors would be contextually reasonable and fitting in the new scenarios, they may not be exactly what the author has in mind for them. Often methods that increase the autonomy of the agents also decrease control over them. With our method, the learned rules for a role are stored in a database and could simply be deleted if they are not desired for new scenarios. One could also imagine simple interactive supervised learning techniques to eliminate undesired behaviors. Alternative learning techniques might also be used to enhance agent behaviors in other situations. For example, could an established learning technique be used to create emerging interpersonal relationships?

Because agents can learn partial definitions of roles and from multiple agents, heterogeneous populations evolve. Unfortunately, this means that conflicts can also arise. What if an agent is being taught conflicting behaviors for a role? One housekeeper explains that items left in classrooms should be put in lost and found, while another housekeeper explains that that can be kept and taken home. Which explanation should be used? Certainly a person's own individual differences including morals would have an impact, but another consideration is the status level of those involved. One agent might out rank the other. The status relationships between agents can be stored in a hierarchy and referenced to decide such conflicts. Such a hierarchy might also be used to distribute tasks when there are multiple agents with the same role. In addition, there exists certain complex task which consists of several sub-tasks. For learning and performing this kind of task, another database field could be added to indicate the learning and performing order of sub-tasks. Also, since the knowledge base of each agent is separated from other's, agents can even have different orders to learn and perform a complex task.

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. Webster's College Dictionary. Random House (1991)
2. Biddle, B.J.: Role Theory: Concepts and Research. Krieger Pub. Co. (1979)
3. Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M.P., Tomlinson, B.: Integrated learning for interactive synthetic characters. In: Proceedings of the 2002 ACM SIGGRAPH Conference, pp. 417–426. ACM (2002)
4. Carey, S., Spelke, E.: Domain-specfic knowledge and conceptual change. In: Hirschfeld, L.A., Gelman, S.A. (eds.) Mapping the Mind. Cambridge University Press (1994)
5. Chang, P., Chien, Y.-H., Kao, E., Soo, V.-W.: A Knowledge-Based Scenario Framework to Support Intelligent Planning Characters. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 134–145. Springer, Heidelberg (2005)

6. Cohen, P.R., Atkin, M.S., Oates, T., Beal, C.R.: Neo: learning conceptual knowledge by sensorimotor interaction with an environment. In: Proceedings of the First International Conference on Autonomous Agents, AGENTS 1997, pp. 170–177 (1997)
7. Conde, T., Thalmann, D.: Learnable behavioural model for autonomous virtual agents: low-level learning. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2006, pp. 89–96 (2006)
8. Costa, P., Botelho, L.: Learning by observation in software agents. In: Proceedings of the 4th International Conference on Agents and Artificial Intelligence, ICAART (2012)
9. Dejong, G., Mooney, R.: Explanation-based learning: An alternative view. Machine Learning 1, 145–176 (1986)
10. Ellenson, A.: Human Relations, 2nd edn. Prentice Hall College Div. (1982)
11. Farenc, N., Boulic, R., Thalmann, D.: An informed environment dedicated to the simulation of virtual humans in urban context. Computer Graphics Forum 18(3), 309–318 (1999)
12. Funge, J., Tu, X., Terzopoulos, D.: Cognitive modeling: Knowledge, reasoning and planning for intelligent characters. In: Proceedings of the 1999 ACM SIGGRAPH Conference, SIGGRAPH 1999, pp. 29–38 (1999)
13. Goertzel, B., Pitt, J., Wigmore, J., Geisweiller, N., Cai, Z., Lian, R., Huang, D., Yu, G.: Cognitive synergy between procedural and declarative learning in the control of animated and robotic agents using the opencogprime agi architecture. In: Proceedings of the 25th AAAI National Conference on Artificial Intelligence, AAAI 2011. AAAI Press (2011)
14. Grimaldo, F., Lozano, M., Barber, F., Vigueras, G.: Simulating socially intelligent agents in semantic virtual environments. Knowl. Eng. Rev. 23(4), 369–388 (2008)
15. Ji, Y., Massanari, R.M., Ager, J., Yen, J., Miller, R.E., Ying, H.: A fuzzy logic-based computational recognition-primed decision model. Inf. Sci. 177(20), 4338–4353 (2007)
16. 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)
17. Kao, E., Chang, P., Chien, Y.-H., Soo, V.-W.: Using Ontology to Establish Social Context and Support Social Reasoning. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 344–357. Springer, Heidelberg (2005)
18. Li, W., Allbeck, J.M.: Populations with Purpose. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 132–143. Springer, Heidelberg (2011)
19. Luo, L., Zhou, S., Cai, W., Low, M.Y.H., Tian, F., Wang, Y., Xiao, X., Chen, D.: Agent-based human behavior modeling for crowd simulation. Computer Animation and Virtual Worlds 19(3-4), 271–281 (2008)
20. McGinnies, E.: Perspectives on Social Behavior. Gardner Press, Inc. (1994)
21. Mitchell, T.M., Keller, R.M., Kedar-Cabelli, S.T.: Explanation-based generalization: A unifying view. Machine Learning 1, 47–80 (1986)
22. Orkin, J., Roy, D.: Automatic learning and generation of social behavior from collective human gameplay. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009, pp. 385–392 (2009)
23. Paris, S., Donikian, S.: Activity-driven populace: a cognitive approach to crowd simulation. IEEE Comput. Graph. Appl. 29(4), 34–43 (2009)

24. 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)
25. Pinker, S.: The Language Instinct. HarperCollins (1995)
26. Rao, A.S., Georgeff, M.P.: Modeling rational agents within a bdi-architecture (1991)
27. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edn. Prentice-Hall (2009)
28. Shao, W., Terzopoulos, D.: Autonomous pedestrians. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 19–28 (2005)
29. Sklar, E., Richards, D.: The use of agents in human learning systems. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2006, pp. 767–774 (2006)
30. Stark, R.: Sociology. Thomson Wadsworth Publishing (2006)
31. Tom, M.: Machine Learning. McGraw-Hill (1997)
32. 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)
33. Yu, Q., Terzopoulos, D.: A decision network framework for the behavioral animation of virtual humans. In: Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA), pp. 119–128 (2007)

The City of Uruk: Teaching Ancient History in a Virtual World

Anton Bogdanovych[1], Kiran Ijaz[2], and Simeon Simoff[1]

[1] School of Computing, Engineering and Mathematics, University of Western Sydney
{a.bogdanovych,s.simoff}@uws.edu.au
[2] Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia
kijaz@it.uts.edu.au

Abstract. In this paper we show how 3D Virtual Worlds can be utilised for teaching ancient history. Our goal is to build an accurate replica of one of humanity's first cities in a 3D Virtual World and provide history students with facilities to explore the virtual city and learn about its past in the simulated 3D environment. Unlike the majority of similar historical reconstructions, an important feature of our approach is having virtual agents that are capable of simulating everyday life of ancient inhabitants, which includes common tasks like eating, sleeping, working and communicating with one another. In order to offer educational value the agents act as autonomous tutors and are capable of sensing the students through their avatars and interact with them both in terms of performing joint actions and through verbal communication. We show how such virtual environments can be built, explain the technology behind its artificial intelligence controlled population and highlight the corresponding educational benefits. To validate the impact of using the 3D environment and virtual agents in history education we conducted a case study that confirmed the beneficial educational aspects of our approach.

1 Introduction

Virtual worlds have become a significant phenomenon in many areas of research, but education is one of the most prominent disciplines where virtual worlds found many applications. A number of studies confirm the importance of using Virtual Worlds for education [1], [2], [3] advocating the use of this technology in a wide range of scenarios, from training customs officers to teaching science to primary school students. One of the key benefits of virtual worlds that was exploited in prior works is the possibility of closely replicating an existing physical environment, so that participants conduct their learning activities in familiar settings and as a part of their learning, also study a remote physical location. But another significant benefit is that virtual worlds make it possible to immerse learners into a realistic reconstruction of a physical location that no longer exists. Such possibility opens new horizons for teaching subjects like ancient history.

Many promising works related to history-oriented virtual worlds focus on reconstructing ancient buildings, objects and even entire cities that are partially or completely destroyed at present. For example, "Rome Reborn" [4] represents

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 28–35, 2012.

a historically accurate reconstruction of a large part of ancient Rome. Visitors of historical simulations similar to "Rome Reborn", are normally able to browse through 3D models of historically significant objects and inspect them from different angles and proximity. They can also select a particular object (i.e. Roman Colosseum) and closely explore its architectural details.

While the majority of history simulations are limited to recreating buildings and artefacts, some pioneering research also explores simulating human life, where reconstructed 3D environments are populated with virtual agents that behave similar to the ancient citizens that used to occupy the given reconstructed site. Most such work, however, employs the so-called "virtual crowds" [5]. Such crowds normally consist of a large number of avatars dressed as local citizens of the reconstructed site. The state of the art in using agent crowds in historical simulations is outlined in [6] where a virtual City of Pompeii is populated with a large number of avatars that walk around the city avoiding collisions. In this work the avatars are simply moving around and are not involved into historically authentic interactions, but simply serve as "walking decorations".

Another important simulation employing virtual agents in historical reconstructions is the "Forbidden City" project [7]. The project simulates the one square-kilometre palace in ancient China grounds called The Virtual Forbidden City. Similar to other historical simulations a significant effort has been put into a realistic recreation of the architecture of the city, while a much smaller effort has been spared on the development of virtual agents. The agents in the Forbidden City are supplied with very limited "intelligence". Their actions are highly scripted and their ability to interact with the users is limited to scripted monologues. The number of available agents is also quite low and the majority of those act purely as guides rather than as virtual inhabitants of the city.

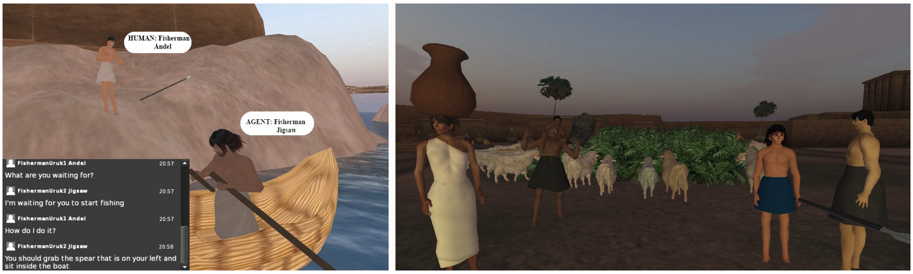
The overview of existing work showed that virtual agents are rarely used in historical simulations and in those few cases when they are employed - their application is very limited. Furthermore, there exists little evidence that such agents contribute to the quality of learning in the domain of ancient history. Thus, in our project we investigate the use of historical 3D reconstructions, the role of believable conversational agents in such reconstructions and investigate the contribution of virtual agents to student learning. We have developed a historically accurate reconstruction of one of humanity's first cities (Uruk) and populated the city with virtual agents that simulate daily life of ancient citizens of this city and are capable of complex interactions with their environment, as well as able to engage into complex multimodal interactions with students.

2 The City of Uruk

As a case study we have created a virtual reality simulation of the ancient city of Uruk in the virtual world of Second Life [8], based on the results of archaeological excavations and under supervision of subject matter experts. The objective of this simulation is to teach history students everyday life of ancient Uruk citizens.

To populate the city with virtual citizens (agents), we designed a number of scenarios that we obtained after detailed discussions with subject matter experts and history consultants. As the result of these discussions we identified roles the agents play, scenes they participate in, interaction protocols and social norms. We followed the methodology described in [9] to structure the knowledge received from the experts and transform it into formalisations suitable for developing the underlying multiagent system. The Virtual Institutions technology [10] was then used to build the normative system, produce agent code skeletons and enable agents automatically comply with social norms and being able to engage into interactions with other agents and humans, while adhering to the social norms.

The agents in the Uruk simulation represent a slice of Uruk society among which are fishermen families, priest, king and a number of workers (i.e. pot maker, spear maker). The agents can sense changes in the environment state, which result in them updating their beliefs accordingly. They are supplied with a number of internal goals and plans to reach those goals. The current implementation features fishermen families where men's daily routines include sleeping, eating, fishing and chatting. The females do house work, sleep, eat bring water from the well and go to markets. The king agent walks around his palace and invites students to ask him about his ruling strategies. The priest agent conducts a prayer in the temple, accepts gifts and explores the city. Other agents represent various workers: pot makers, spear makers, etc. Those workers produce goods, exchange goods with one another, attend the prayer and in their spare time explore the city, provide information to students and simulate social interactions with other agents. Fig. 1 illustrates agents "living" in virtual Uruk[1].



FishermanUruk1 Andel 20:57
What are you waiting for?

FishermanUruk2 Jigsaw 20:57
I'm waiting for you to start fishing

FishermanUruk1 Andel 20:57
How do I do it?

FishermanUruk2 Jigsaw 20:58
You should grab the spear that is on your left and sit inside the boat

Fig. 1. Student interacting with an agent; Group of agents in the city centre

2.1 Implementing Virtual Agents

Our agent architecture is based on the BDI (Belief Desire Intention) model and functions similarly to Jack Intelligent Agents platform [11].

We have created our own agent library (VIAgents) for developing agents that follow our formal model. It provides them with communication facilities, planning and goal oriented behaviour. Each agent has a number of beliefs that map to a state of the virtual world and the institutional state, a number of goals the agent wants to achieve and the number of plans the agent can perform in order to achieve these goals. As the result of sensing the changes in the environment or events received from the institutional infrastructure or other agents, the

[1] A demo video is available at: http://www.youtube.com/watch?v=15BaDjd7f1c

agent may establish new goals or drop the existing goals and will start or stop corresponding plans accordingly. The BDI architecture is also extended with a learning mechanism that enables the agent to learn by example from a human expert. This part of the agent architecture is outlined in [12].

Interactions with the Environment. There are two parts of the environment the agents are involved in: the visual part rendered by Second Life and the normative part supported by Virtual Institutions technology. The agents interact with the virtual world by sending commands to the Second Life server through `libopenmetaverse` library [13] . Through these commands they are capable of moving around the virtual world, perform certain animated behaviours, perform commands on the objects in the environment and interact with other agents or humans. The normative part enforces the social norms on all the Second Life participants. Agents interact with this part by sending and receiving illocutions (text messages) to the AMELI system [14].

In the visual part the agents are capable of sensing the changes of the environment state, including the movement of the objects or avatars, new objects or avatars, actions performed by all participants, etc. In the normative part the agents can send and receive the institutional illocutions and sense the state changes resulted by these illocutions. The state of the visual part of the environment represents parameters like the time of the day, positions and transformations of the objects and agents in the virtual world. The institutional state corresponds to the set of currently active scenes and the state of each scene.

Object Use. To convincingly simulate daily life of the ancient people it was important to enable agents to use objects in the environment (i.e. take a spear, jump on a boat and go fishing). We implemented a fully-fledged object use library for the history simulation context, which includes a set of classes allowing agents to identify an object in the virtual world, attach it to the default attachment point, play a certain animation (e.g. rowing) associated with a given object, wear an object that is a piece of clothing, detach the piece of closing, drop an object to the ground and detach the object and hide it in the avatars inventory.

Goal Generation and Planning. Each of our agents follows a classical BDI model [11], where agents' actions are shaped by an individual agent's beliefs about the virtual environment, goals the agent must achieve and plans that represent a method of achieving certain goals. The Virtual Institutions technology [10] that is employed for encoding the normative layer of the virtual environment provides facilities for dynamic generation of goals for every agent when a certain institutional action is required to be performed. Some of the internal agent goal triggers are manually embedded into the agent code.

The changes in the virtual world (that happen as the result of actions of students or other agents) might result agents updating their goals. In order to achieve their goals they can use either static or dynamic planning. Static plans are instructions prescribed by an agent programmer and can't be changed at runtime. In the case of dynamic planning an agent can sense its current state in the environment and can react to environment changes re-evaluating its current plan. Rather than having a complete recipe provided for every situation the agent can encounter - the agent is simply given the list of possible actions and has to find a way of combining those to reach its goals. For making this task realistic, each action is supplied with its execution preconditions and postconditions. The preconditions define the state, scene and objects that the agent must have to execute the given action. The postconditions determine how those attributes will change after performing the action. Pre- and post-conditions are included as environment annotations and can be updated at run time.

Conversational Ability. All agents can chat with human visitors on a number of preprogrammed educational topics through the instant message and chat mechanisms provided by Second Life. In order to participate in a conversation our agents employ the ALICE chat engine based on the AIML language [15]. Each agent uses a number of AIML files that represent what can be seen as a common sense database. Additionally to this database every agent is supplied by personalised AIML files that reflect on its personality and the data relevant for its role within the virtual society. While conversing with students agents are capable of talking about their current state, the goal they pursue, reasons for pursuing this goal and give explanation about surrounding objects via the environment-, self-, and interaction-awareness model [16].

3 Validation of Learning Effectiveness

In order to test the learning effectiveness of using virtual worlds and virtual agents in history education, in our study we compared two different ways of learning history with two different sample groups. The first group, which we call the “Traditional Group”, was advised to read a history text describing the facts about the city of Uruk (3000 B.C.) and its inhabitants. Participants in the second group (the “Virtual Group”) were asked to visit the virtual Uruk to have an interactive learning experience about the same facts as the text based group. All study participants were undergraduate university students with no previous knowledge about Uruk. The two groups were aiming for the same learning objectives and contents, and were also able to access supervision if they required. It was assumed that any variation in student learning outcomes and appeal to students could be attributed to the group type factor. Based on our research objective of evaluating the learning of history and culture in 3D Virtual Worlds in comparison to text based learning, we have the following hypotheses:

- **H1:** Students in the virtual group will have significantly better learning outcomes in comparison to the students from the traditional group.
- **H2:** Student evaluations will show whether learning history has more appeal to the virtual group students than students in the traditional group.

Research Instruments. For testing the above hypotheses we aimed to collect quantitative and qualitative data about the usefulness of using virtual worlds and conversational agents in teaching ancient history. Therefore, we developed a questionnaire that contained two parts: the pre-test part collected the users’ demographics and their background knowledge about Uruk. The second part was a post-test questionnaire aiming to test the knowledge gained about Uruk in each study groups. The pre-test questionnaire consisted of six multiple choice questions about the participant's demographical information and one pre-coded (open ended) question to test previous knowledge of the historical city. The post-test part of the questionnaire aimed at examining the knowledge gained in these user groups (an open-ended questionnaire for participants' feedback). Only students that showed having no previous knowledge about Uruk and ancient Mesopotamia were selected to continue after the pre-test.

The traditional group was provided with a text describing the city of Uruk, specifically it included people, buildings, climate of the city, historical significance, food, agriculture, trade and inventions made in that era. This text was designed with the help of subject matter experts. This text description comprised the same facts on which we developed our virtual city of Uruk.

The objective of the post-test questionnaire was to measure the student's knowledge of historical concepts presented in two different ways. The post-test questions fall under four major aspects to test the knowledge: climate and buildings, people/food/animals, agriculture and trade, Uruk inventions. These categories were based on the Uruk's prototype we developed in the Virtual World.

Participants. After the initial pre-test screening we selected 40 undergraduate students from the University of Technology, Sydney with no previous knowledge about Uruk and ancient Mesopotamia. All the participants were briefed about the study at the start of their respective sessions, by the researcher or their teacher. These students were then randomly assigned to one of the study groups, the traditional group or the virtual group. At least one researcher was present in each study session. The traditional group was assisted by their teacher on behalf of the researcher. A brief description of the study was given for 5 minutes followed by a class session which lasted about 30-40 minutes for two respective groups. The time required to fill out the post-test questionnaire was limited to 15 minutes and additionally 5 minutes were given to provide the post-test feedback.

3.1 Study Findings

The study showed a clear difference in performance gained between the traditional and virtual group students. The metric to measure any performance variations between the two groups was based on the marks achieved in the post-test exam. This performance comparison is illustrated in Table 1.

Overall, virtual group students outperformed the traditional group, with an average mark of 60.96%, while the students in the traditional group achieved an average score of 41.05%. The minimum performance achieved by traditional group students was 23% in comparison to the 40% minimum marks gained by the virtual group students. Similarly, the highest marks achieved in the traditional group were 66%, which was quite low compared to 88% achieved by the virtual group students. Moreover, these results show that virtual group performance stayed high for more students than in the traditional group.

Table 1. Student's Performance

Comparison: Traditional vs Virtual Group			
Traditional Group	Students(%)	Marks Obtained(%)	Weighted Average(%).
	15	23 - 25	
	35	32 - 39	41.05
	35	40 - 48	
	15	58 - 66	
Virtual Group	Students(%)	Marks Obtained(%)	Weighted Average(%)
	20	40 - 49	
	35	51 - 60	60.96
	25	61 - 69	
	20	77 - 88	

4 Conclusion

We presented a 3D simulation of the city of Uruk and showed how it was used for teaching ancient history and conducted an experimental evaluation of learning effectiveness of virtual Uruk. A group of students learning from a text document were compared to a group learning through interacting with virtual agents in the virtual world. The quality of learning was evaluated through conducting a written exam with students in each study group. The study outlined better performance achieved by the virtual group over the traditional text reading group. The study also showed that the virtual group was more engaged and willing to spend more time on learning.

References

1. Loyalist College: Virtual World Simulation Training Prepares Real Guards on the US-Canadian Border (2009), http://secondlifegrid.net.s3.amazonaws.com/docs/Second_Life_Case_Loyalist_EN.pdf
2. Gibson, D.: Games and Simulations in Online Learning: Research and Development Frameworks. Information Resources Press, Arlington (2007)
3. The Open University in Second Life: The open universitys place for us: Providing geographically dispersed students & faculty a place to meet and learn together. Linden Lab: Case Studies, May 11 (2009)
4. Guidi, G., Frischer, B., Russo, M., Spinetti, A., Carosso, L., Micoli, L.L.: Three-dimensional acquisition of large and detailed cultural heritage objects. Machine Vision Applications 17(6), 349–360 (2006)
5. Gutierrez, D., Frischer, B., Cerezo, E., Gomez, A., Seron, F.: AI and virtual crowds: Populating the Colosseum. Journal of Cultural Heritage 8(2), 176–185 (2007)
6. Mam, J., Haegler, S., Yersin, B., Mller, P., Thalmann, D., Van Gool, L.: Populating Ancient Pompeii with Crowds of Virtual Romans. In: 8th International Symposium on Virtual Reality, Archeology and Cultural Heritage - VAST (2007)
7. Palace Museum and IBM: The Forbidden City: Beyond Space and Time (2009), http://www.beyondspaceandtime.org
8. Linden Lab: Second Life (2012), http://secondlife.com

9. Bogdanovych, A., Rodríguez, J.A., Simoff, S., Cohen, A., Sierra, C.: Developing Virtual Heritage Applications as Normative Multiagent Systems. In: Gleizes, M.-P., Gomez-Sanz, J.J. (eds.) AOSE 2009. LNCS, vol. 6038, pp. 140–154. Springer, Heidelberg (2011)
10. Bogdanovych, A.: Virtual Institutions. PhD thesis, UTS, Sydney, Australia (2007)
11. Howden, N., Ronnquist, R., Hodgson, A., Lucas, A.: JACK intelligent agents - summary of an agent infrastructure. In: Proceedings of the 5th ACM International Conference on Autonomous Agents (2001)
12. Bogdanovych, A., Simoff, S., Esteva, M.: Training believable agents in 3D electronic business environments using recursive-arc graphs. In: IC-Soft 2008, pp. 339–345. INSTICC (2008)
13. Openmetaverse library for Second Life (2012), http://lib.openmetaverse.org
14. Esteva, M., Rosell, B., Rodriguez-Aguilar, J.A., Arcos, J.L.: AMELI: An Agent-Based Middleware for Electronic Institutions. In: Proceedings of AAMAS 2004, vol. 01, pp. 236–243. IEEE Computer Society, Los Alamitos (2004)
15. Wallace, R.: The elements of AIML style. ALICE AI Foundation (2004)
16. Ijaz, K., Bogdanovych, A., Simoff, S.: Enhancing the Believability of Embodied Conversational Agents through Environment-, Self- and Interaction-Awareness. In: Reynolds, M. (ed.) Australasian Computer Science Conference (ACSC 2011). CRPIT, vol. 113. ACS, Perth (2011)

An Analysis of the Dialogic Complexities in Designing a Question/Answering Based Conversational Agent for Preschoolers

Anuj Tewari, Ingrid Liu, Carrie Cai, and John Canny

Dept. of Electrical Engineering and Computer Science, University of California, Berkeley, USA
{anuj,jfc}@eecs.berkeley.edu, ingy@berkeley.edu, cjcai@mit.edu

Abstract. Parents are well aware that pre-school children are incessantly inquisitive, and the high ratio of questions to statements suggests that questions are a primary method utilized by children for language acquisition, cognitive development, and formulating knowledge structures. Question-asking is furthermore a comfortable medium for a child to stay engaged in natural discourse and the activity at hand. To take advantage of the naturalness and learning benefits of question-answer exchanges, there could be intelligent agents that can engage a child in activities while setting children in the mood to ask meaningful, information-seeking questions. There are currently multiple intelligent agents that can interact with older children and adults to promote literacy or teach topics in specific domains. This paper thus focuses on the complexities of designing an intelligent agent for younger children, by collecting and analyzing data and categorizing children's questions, which are often ill-formed.

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

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 child's later educational and life opportunities [5],[3],[4]. Hart and Risley [2] report a factor of two difference in the working vocabularies of high vs. low-SES (Socio-Economic Status) 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 pre-school stage. "The richness of nouns, modifiers, and past-tense verbs in their parents' utterances, their parents' high propensity to ask yes/no questions, especially auxiliary-fronted yes/no questions; and their parents' low propensity to initiate and use imperatives and prohibitions were more strongly predictive of the children's

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 36–45, 2012.

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. As noted in the above studies, the greatest impact on child literacy will come from intervention at pre-school ages.

While it is becoming increasingly clear that conversations and language interactions serve as an important tool in the child's cognitive process, a growing body of research is also suggesting that pre-school children are voracious inquisitors. One recent study found that preschoolers ask approximately 80 questions/hour [1] which constitutes more than one-fourth of their utterances. These questions are an essential part of language development: they provide primary experience with question construction, statement construction, explanation construction, complex tenses etc. The child question-asker is primed for an answer. Unlike other forms of interaction (reading, games) no external influence is needed to garner the child's interest or build motivation. The questions reflect the child's current state of knowledge and should take them just beyond it. In other words, child-initiated questions are naturally in the child's Zone of Proximal Development (ZPD). Question-asking, not surprisingly, goes beyond literacy and is an integral part of children's cognitive development [1].

It is safe to assume that parents are the primary teachers for preschool children, but many interventions directed at parents reproduce the gap. Educational interventions for children involving parents appear to be dependent on the parent's educational level, so literacy differences persist across generations. For instance, dialogic reading (defined later) interventions involving high-SES parents were far more effective than with low-SES parents [17]. Children evidently need some form of linguistic engagement for many hours a week, with a language-able partner who can engage with them in age-appropriate language-learning activities. Since research in early child development suggests that for pre-school children question-answering serves as a frequent and heavily-utilized medium of synchronizing mental models with adult-like understanding of the world, this linguistic engagement can come in form of interactive question-answering systems. Since children spend a significant amount of time playing alone, or out of home, there might be instances when they don't find an adult around to answer their questions. There might also be times when the adult doesn't have sufficient information at hand to answer a child's question. This explains the need for expert interactive systems that can work as engaging question-answering agents. However, before any type of technology push, we want to establish a theoretical framework in which such interventions can be based. Therefore, this paper outlines the dialogic complexities involved in designing a Q/A system for preschoolers, by analysis of transcripts from the CHILDES database [14].

2 Related Works and Background

Child development research has shown that children rapidly acquire knowledge of new words starting at 18 months of age. According to Jean Piaget's theory of development, it is during the pre-operational period (ages 2-7) during which children become able to represent ideas through language and mental imagery [18]. Vocabulary size more than doubles between 18-21 months and again between 21-24 months of age, and a typical child understands at least 10,000 words by first grade. These patterns suggest a high propensity for children to acquire vocabulary at a very young age, and that preschool age is likely an appropriate time to engage children in language learning [21]. Moreover, scaffolded linguistic interactions with adults significantly advance children's learning. For example, toddlers whose mothers follow their attention by labeling objects of joint attention tend to have larger vocabularies later on [21]. Adult grammar provides semantic clues that aid children in deciphering the meaning of words, and social cues also help children develop competency through the corrective feedback that adults give when children use words incorrectly.

According to psychologist Lev Vygotsky, such interactions are not merely external forces that provoke internal change in an individual, but rather integral to the very mechanism of cognitive development [23]. Because childhood word learning both increases rapidly at an early age and demands support from adult modeling, it is valuable to examine ways in which adult-child interactions at the preschool age can be modeled through software interfaces. It is common knowledge that young children ask a considerable number of questions, but to correlate children's inherent motivation to develop theories about the world with their question asking, the amount, content, and responses to adult's answers have been analyzed. In a longitudinal study of transcripts involving four children, ages 2.5-4, 71% of the questions were information-seeking questions, and of these, 75% were fact-seeking and 25% were explanation seeking questions [7]. Noninformation-seeking questions ranged from seeking attention, clarification, action, permission, play, towards a child or animal, or were unknown [7].

Based on questions with young children, such as asking the children for sentence completions, Piaget concluded that young children had very primitive notions of causality under 5 or 6 years old [18]. However, recent works are re-examining Piaget's claims. Shultz performed an experiment, where children of ages 3, 5, 7, and 9, were shown three pairs of two objects, where one object was the cause of an effect, and asked to identify the object which created the effect. Children of all ages were able to correctly link the causes and effects using attributes of the source or result. Hood and Bloom [9] find that children make causal statements and responses to causal questions by adults from at least age 24 months, and by 30 months, they can ask causal questions productively. Furthermore, these causal questions are oftentimes more sophisticated than one word questions such as "why" and "how" that are meaningful in the context of specific domains such as natural phenomena, biological phenomena, physical mechanisms, motivation/behavior, and cultural conventions. In a study by Callanan and Oakes [7], parents of children ages 3, 4, and 5 were asked to record forms for children's questions, with special focus on causal questions for two weeks. At age 3, 20% of "why" questions were simply "why?" at age 4, 10% were "why?" and at age 5, 4% were "why?". This demonstrates that age plays a major role in the kind of questions that children ask. Frazier et al. [8] performed a laboratory experiment where investigators engaged children in conversation about a set of unusual toys and alternated between providing explanatory versus non-explanatory answers to the children's questions. Shultz's experiments provide evidence that children can judge causality by using their knowledge of object attributes, or by generative transmission, rather than on attributes such as spatial or temporal contiguity [20].

Many recent research papers have focused on categories of children's questions through manual coding. These studies generally perform diary studies, or perform laboratory experiments observing the question/answering dynamic for young children. Questions can be coded along several dimensions: information-seeking versus non information-seeking, response desired, content, response type, and information given in the response [1]. The response desired can be a fact or explanation if the question is information-seeking, or it can be attention, clarification, etc. if the question is non information seeking [1]. Content can range from the label, appearance, property, etc. of the questions' subjects [1]. Callanan and Oakes also derived the statistics of the ratio of causal question types, the situations in which they emerge, and their content through a diary study of 30 preschool children [7]. Frazier et al. derived the statistics of parent's responses to children's causal questions and children's responses to different responses by their parents by examining longitudinal studies from CHILDES and through laboratory experiments with 42 preschool children [8].

3 Children's Questions in Various Activities

There are several components requiring research from various fields that are necessary to construct any technology that promotes question asking by pre-school aged children. The conversation dynamics between children and adults have their own structures and processes, with complex rules of turn-taking. In this domain, we are primarily interested in how to best encourage a child to ask meaningful, instructional questions while keeping them engaged. To anticipate and correctly answer the questions that children may ask, it is necessary to properly identify and group the questions with the type of response needed. Determining the types and levels of engagement children have during specific activities in their daily lives will guide us in designing technology that promotes their question asking.

3.1 Materials

The focus of our question categorization is to investigate how engagement differs with interaction. For our analysis, we had to choose between labeling dialogs of spontaneous child play or dialogs of children with controlled play activities in a laboratory setting. For spontaneous child play, the dialogs would have to be coded for the activity type, and there would be variation within groups of interaction types, such as the type of toys a child had access to in a game of pretend. Furthermore, the transcripts of child and parent interaction lack any details regarding the surrounding objects, simultaneous events, and other extraneous circumstances, making them difficult to code for interaction type and difficult to annotate for disturbances. For controlled play activity, there are always pitfalls related to the naturalness of interaction in an unfamiliar setting with new objects. The observer's paradox is an additional concern, which affects both child and parent, since cameras and investigators easily distract the children, while parents are concerned with their appearance as guardians [11].

After preliminary analysis of both types of datasets within the CHILDES database, the spontaneous child play was determined to be very difficult to annotate in a consistent manner, and a laboratory study of adult with child interactions was chosen. The

Gleason database includes transcripts of 24 different children- 12 boys and 12 girls-in various activities with their father and mother separately. In the lab, the child and parent engaged in three activities: playing with a toy auto, reading a picture book, and playing store (also referred to as "pretend"). The parents were encouraged to divide the time evenly among the activities, and the activity order and parent order were randomized [16]. Since there are laboratory studies of the child with the Mother and Father separately and the studies are spaced out, we only analyze the transcripts of children who are between the ages 3-4 for both visits. This left 6 children, ages 3;1.04[1], 3;7.01, 3;2.21, 3;2.12, 3;2.03, and 3;7 during the Father's visit and ages 3;0.20, 3;6.07, 3;2.02, 3;3.16, 3;2.21, and 3;7.25 during the Mother's visit.

3.2 Procedure

Since we are interested in building an interactive interface for addressing children's questions, we code the questions in the Gleason study across various dimensions of question types. The first dimension chosen was **questions of causality**. The causal categories were chosen from the Callanan and Oakes [7] study as a comprehensive overview of children's causal questions, and no other causal question types were found during coding. The second category was the **response type expected**. If a child's question is in a causal category, then the question requires an explanation. If a child's question does not fit in a causal category, but is still information seeking, then the response needed is a fact. This includes clarification of a previous statement, confirmation that a belief or answer is correct, or any other question that seeks information. If the child asks a question for attention, to direct the conversation to a different topic, to direct attention of the adult to an object, to request something, or to signify interest or impatience, then the question is non-information-seeking. Of the information seeking questions (fact-seeking and explanation-seeking), the question can be directed completely towards the activity at hand and provide the child with no new information of the world. These questions are labeled "within scope". Questions that are "outside scope" can still be about the current toys the child is interacting with; however, it should add to the child's knowledge base of object names, properties, or mechanisms in the world. Lastly, the adult prompts many of the questions that children ask. To engage a child, adults will often ask the child a question. When the child repeats the question, the question is not the result of the child's inherent interest, but of the adult's mode of interaction, and is thus coded.

3.3 Discussion

The Gleason dataset was relatively simple to divide into the three activities, since the parents were encouraged to split time between the activities evenly across their half hour in the lab. There was a section at the end where the investigators holding the study gave the children a gift: this section was not included in the category analyses. There were also instances where the children noticed a camera in the room and conversed about the

[1] Ages are represented with year;month.day, where day is optional. For example, 1;5.10 is a child that is 1 year, 5 months, and 10 days old.

camera: this section was also removed in category analyses. Lastly, there was overlap between activities. When fully engaged in an activity, the child would stay focused on the task at hand; however, between activities, it would take several turns of persuasion by the parent to continue with the next task. A new task is considered started when the parent or child suggests the activity, and there is no further debate after that line of starting the new activity.

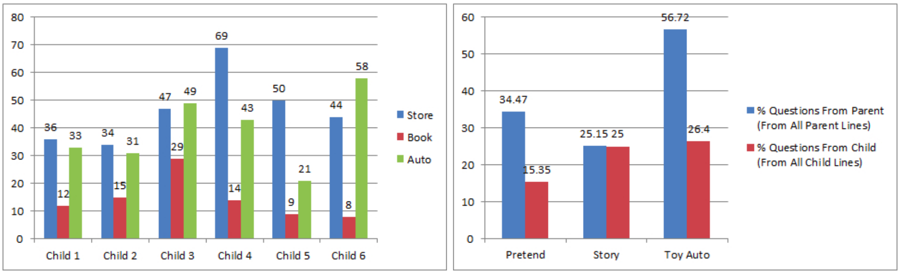


Fig. 1. Left: Ratio of conversational turns by parents and children percentages in various activities. **Right:** Percentage of questions by parents and children in various activities. Y-axis denotes percentages in both graphs.

Play Encourages Questioning: The type of role children take based on activity can be inferred by the ratio of child statements to questions and the number of child turns to adult, as they vary greatly between activities. Figure 1 presents these ratios. For example, in the game of pretend, children took up relatively more turns in the conversional exchange, but asked fewer questions relative to their increased speech. From the transcripts, it is clear that children are more interested in the role-playing aspect of pretend, than asking questions about familiar objects. Thus, they direct the conversation towards the make-believe world they wish to enact, while asking questions only when they are uncertain what a toy prop is. In contrast, children on average took a more passive role in interpreting the picture book. In general, the parents made up the story for the child, and most children took the role of listener, with varying degrees of participation in story-making.

When constructing the toy auto, parents tended to take a more verbally active, tutorial role, answering questions, giving suggestions in both statements and questions-and giving and asking for additional information about the different components of the car. This is reflected in the relatively high ratio of questions to statements by parents. These numbers hold across all children, and Figure 2 presents the percentage of child-initiated questions asked per child per activity, out of all questions asked in the activity by the child and adult. From this figure, it is more apparent that in the story activity, parents ask many more questions than children. Overall, children were more active when playing store or playing with the auto. When reading the story with their parents, however, the amount of child-initiated questions varied greatly. It should be noted that this cannot be seen as a result of just the child. There are two sides to the conversation, and the variance in child-initiated questions may also be due to the specific dynamics between the parent and child. The percentage of child-initiated questions is presented in Figure 2.

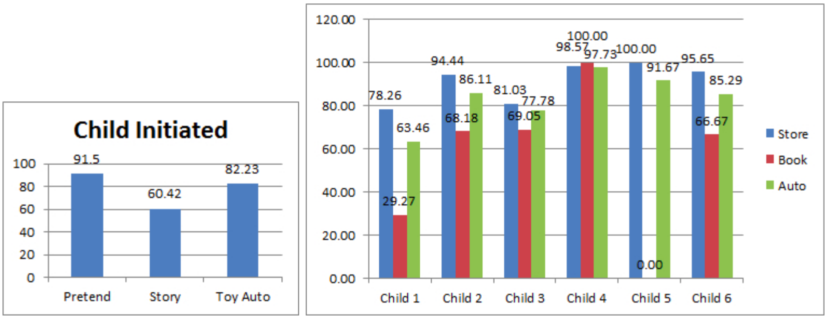


Fig. 2. Left: Percentage of questions initiated by children across activities out of all questions asked by children in an activity. **Right:** Percentage of questions initiated by each child across activities out of all questions asked by a child in an activity.

Children's Questions Are Grounded in Scope: For our study, a question is considered "outside scope" if a response would either help categorize an object, describe properties of an object, or explain how an event, such as how to connect an engine to a car, would take place. In this case, unlike Chouinard's definition [1] of information-seeking questions, "outside scope" can include clarifications of what an adult said. At the same time "outside scope" is also much more conservative. Fact-seeking questions such as "where is it?" are considered within scope because they only pertain to the activity at hand.

Since most of the questions children asked while reading the picture book involved what explicitly was transpiring on the pages, children could ask few "outside scope" questions during this activity. On the other hand, since the toy auto activity was focused primarily on constructing the auto, the child could ask many questions on the mechanisms, pieces, and properties of the car. There were quite a few "outside scope" questions in the game of pretend as well, since even though children were involved in role-playing, the children were unfamiliar with many props and asked for their labels. As a note, there were few "outside scope" questions during reading, most likely resulting from the fictional nature of the story and the listener-role adopted by the child. If the book were non-fiction, the results might be very different. The percentages of child questions that are outside scope are presented in Table 1.

Table 1. Percentage of questions that are outside scope across activities

Activity	Outside scope questions
Pretend	28%
Story	13%
Toy Auto	40%

Negatively Phrased Statements Lead to Followup Questioning: It was observed that the number of "why?" questions increase following a negatively phrased statement by an adult. Since we are interested in promoting meaningful questions from children, we examined the ratio of "why" questions following negative phrased statements versus other question types.

Since children ask questions to solidify their understanding of the world, it should be expected that questions contradicting children's current beliefs should prompt them to ask more "why" questions. In the Callanan and Oakes [7] study, this was coded for by checking for "why" questions which contain negative words and phrases, such as "not" and "can't", and they found that the proportion of "why" questions with negative words and phrases to be low overall. As we are interested in ways to promote meaningful question asking in children, we decided to approach this from a different angle. Instead of looking at "why" with negative words or phrases, we look at the number of "why" questions as a result of an adult making a statement in a negative way. The percent of "why" statements following adults that use any of the words "not", "no", "neither", or "never", including contractions, was compared with the percent of other questions following the negative words. For this study, we use all free-interaction studies for the age range 3-4 ([6], [10], [14], [19], [22]). The results are available in Table 2.

Table 2. Number and percentage of why questions and all questions following a negatively phrased statement

	#	%
"Why" Questions Following Negative Statement	175	14.31(from "why" questions)
Total Why Questions	1223	
Questions Following Negative Statement	799	6.41 (from all questions)
Total Questions	12474	

Without further analysis, we can only make hypotheses for the greater percentage of "why" questions following negative statements. Children could, as mentioned above, be asking "why" questions because their expectations of the world were violated. Other possibilities include conversation formalities, greater comfort with the language structure of "why", increasing the likelihood of being granted permission to do something that was originally forbidden, etc. In conversation formalities the child may ask "why", as a way to express interest, which is an important for maintaining conversational discourse. There are many studies on children repeating statements by parents, so there is also support for children using "why", because repeating a commonly used phrase can be related to repeating a previously said line. Asking "why" to persuade adults for granting permission is a probable hypothesis as children often ask "why" when denied permission; however, more often than not parents will still deny the child's request after being asked "why" which reduces the plausibility of children asking "why" to be given permission. Regardless, the significantly higher percentage of "why" questions following negative statements suggests that using negative statements in a conversation can be a useful technique for prompting children to ask "why".

4 Conclusion

Overall, we have shown that children tend to take on a more conversational role and ask questions more frequently when involved in play-like activity. In storytelling activities, they tend to take a more passive role while the parents take the lead. Moreover, children tend not to ask "outside scope" questions; their questions most often remain grounded in the activity at hand. In particular, ambiguity in the activity, such as in a storybook reading, tend to prompt "why" questions whereas activities requiring joint attention on a constructive task are more likely to prompt "how" questions. Negative responses from an adult may also be more likely to elicit questions from children; however, care should be taken to ensure that the negative responses are not harsh, but rather meant to encourage inquisitiveness and metacognition. Such findings suggest that conversational Q/A agents may be more effective in keeping the child verbally engaged if they have built-in activities that require joint attention on a creative task or puzzle, as well as mild negative responses which may pique a child's interest in asking further questions.

References

1. Chouinard, M.M.: Children's questions: A mechanism for cognitive development. Monographs of the Society for Research in Child Development, 72 (1, Serial No. 286) (2007)
2. Hart, B., Risley, T.: Meaningful Differences in the Everyday Experience of Young American Children. Paul H. Brookes (1995)
3. NFCL: National Family Literacy Organization, main site http://www.famlit.org
4. NIH: Clear Communication: An NIH Health Literacy Initiative, http://www.nih.gov/clearcommunication/healthliteracy.htm
5. NELP: National Early Literacy Panel, Developing Early Literacy, National Institute for Literacy, NIFL (2008), http://www.nifl.gov
6. Brown, R.: A First Language: The Early Stages. Harvard University Press, Cambridge (1973)
7. Callanan, M.A., Oakes, L.M.: Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. Cognitive Development (1992)
8. Frazier, B.N., Gelman, S.A., Wellman, H.M.: Preschoolers Search for Explanatory Information Within AdultChild Conversation. Child Development 80(6), 1592–1611 (2009)
9. Hood, L., Bloom, L., Brainerd, C.J.: What, when, and how about why: A longitudinal study of early expressions of causality. Monographs of the Society for Research in Child Development (1979)
10. Kuczaj, S.: The acquisition of regular and irregular past tense forms. Journal of Verbal Learning and Verbal Behavior (1977)
11. Labov, W.: Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia (1972)
12. MacWhinney, B., Snow, C.: The child language data exchange system. Journal of Child Language (1985)
13. MacWhinney, B., Snow, C.: The child language data exchange system: An update. Journal of Child Language (1990)
14. MacWhinney, B.: The CHILDES project: Tools for analyzing talk, 3rd edn. Lawrence Erlbaum Associates, Mahwah (2000)
15. MacWhinney, B., Snow, C.: The Child Language Data Exchange System: An update. Journal of Child Language 17, 457–472 (1990)
16. Masur, E., Gleason, J.B.: Parent-child interaction and the acquisition of lexical information during play. Developmental Psychology (1980)
17. 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)
18. Piaget, J.: Judgment and reasoning in the child. Routledge and Kegan Paul, London (1969)
19. Sachs, J.: Talking about the there and then: The emergence of displaced reference in parent-child discourse, vol. 4. Lawrence Erlbaum Associates, Hillsdale (1983)

20. Shultz, T.R., Mendelson, R.: The use of covariation as a principle of causal analysis. Child Development (1975)
21. Siegler, R.S., Alibali, M.W.: Children's Thinking, 4th edn. Prentice Hall (June 2004)
22. Warren-Leubecker, J.N., Bohannon, A.: Intonation patterns in child-directed speech: Mother-father speech. Child Development (1984)
23. Vygotsky, L., Hanfmann, E., Vakar, G.: Thought and Language. Studies in Communication. MIT Press (1962)

Building Autonomous Social Partners for Autistic Children

Sara Bernardini[1], Kaska Porayska-Pomsta[1], Tim J. Smith[2], and Katerina Avramides[1]

[1] London Knowledge Lab., Institute of Education, University of London,
23-29 Emerald Street, London WC1N 3QS, United Kingdom
{S.Bernardini,K.Porayska-Pomsta,K.Avramides}@ioe.ac.uk
[2] Psychological Sciences, Birkbeck, University of London,
Malet Street, London WC1E 7HX, United Kingdom
tj.smith@bbk.ac.uk

Abstract. We present the design and implementation of an autonomous virtual agent that acts as a credible social partner for children with Autism Spectrum Conditions and supports them in acquiring social communication skills. The agent's design is based on principles of best autism practice and input from users. Initial experimental results on the efficacy of the agent show encouraging tendencies for a number of children.

Keywords: Virtual Social Partners, Pedagogical Agents, Autonomy.

1 Introduction

This paper presents an autism practice-based approach to designing and implementing an autonomous virtual agent that can act credibly both as a peer and as a tutor to support young children with Autism Spectrum Conditions (ASCs) in developing social communication skills. The design of the agent is based on participatory design workshops with practitioners and children as well as the SCERTS framework [7] - an established educational intervention approach aimed to support social communication (SC) and emotional regulation (ER) of children with ASCs through appropriately designed transactional support (TS). Our pedagogical agent is implemented in a virtual environment called ECHOES intended for real-world use in schools as part of their everyday activities.

Autism is a spectrum of neuro-developmental conditions that affects three main areas ("triad of impairments" [1]): (i) *communication*: problems with verbal and non-verbal language; (ii) *social interaction*: problems with recognising and understanding other people's emotions and with expressing their own emotions; and (iii) *patterns of restricted or repetitive behaviours*: problems with adapting to novel environments. We focus on enhancing the *social communication* competence of children with ASCs because this is the domain with which they typically have the most difficulty [6] and because recent studies indicate that individuals with ASCs and their caregivers consider support for social communication as the most desirable feature of technology-enhanced intervention [8].

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 46–52, 2012.

The paper is organised as follows: Section 2 discusses why children with ASCs may benefit from virtual agents; Section 3 presents the SCERTS model, which provides the theoretical foundation of our agent's design; Section 4 describes the design and the implementation of the ECHOES agent. Sections 5 and 6, respectively, report the evaluation of the agent and offer our conclusions.

2 Agent Technology for Autism

Children with ASCs have an affinity with technology and are motivated by computer-based training [8]. Software programs are predictable and structured environments that can accommodate the children's need for organisational support and their preference for routine behaviours. The anxiety linked with social interaction can be mitigated by the use of artificial peers which are tireless, consistent and positive towards the child regardless of the child's behaviours. An appropriately designed artificial peer can meet individual children's needs and allow them to exercise the same skill in different scenarios, from structured situations to gradually more unpredictable contexts, thereby increasing the chances of transferring the learned skills from the virtual to the real world [11,2].

Despite a recent growing interest in the potential of artificial agents, both virtual and physically embodied, for human-computer interaction the efforts have focused primarily on agents with little or no *autonomy*, with the exception of the Thinking Head project [4], which focuses on developing a talking head that teaches social skills through its ability of realistically portraying facial expressions, and the virtual peers, Baldi and Timo [2], which are 3-D computer-animated talking heads for language and speech training. Autonomous agents carry a significant potential for autism intervention for children, because they can compliment the intensive one-on-one support that the children need, by allowing human practitioners to focus on the most complex aspects of face-to-face interventions, while managing any repetitive tasks and on-demand access.

The approach presented in this paper focuses on the development of a *fully autonomous* agent, i.e. an agent that is able to decide independently how to act best in order to achieve a set of high-level goals that have been delegated to it. In keeping with the classic agent theory of Wooldridge and Jennings [13], in addition to autonomy, the ECHOES agent is equipped with: (i) *pro-activeness*; (ii) *reactivity*; and (iii) *social ability.* The agent's pro-activeness is important to maintaining the child's attentional focus and to foster motivation. Reactivity is fundamental to adapting the support to the children's changing needs as well as cognitive and affective states, while social ability – to maximising the chances of the child to experience a sense of self-efficacy in communicating with the agent.

3 Pedagogical Underpinnings for the ECHOES Agent

In order to identify the social communication skills that a virtual agent needs to possess to act as a credible social partner to children with ASCs and to support their social competencies, we draw from SCERTS [6], a comprehensive approach to social communication assessment and intervention in autism.

SCERTS identifies the particular skills that are essential for successful social communication and which we argue are also necessary for an *ideal* virtual agent to possess in order to act as a credible social partner to children with ASCs. These skills are encapsulated in three overarching domains: (i) *Social Communication*: spontaneous and functional communication, emotional expression, and secure and trusting relationships with others; (ii) *Emotional Regulation*: the ability to maintain a well-regulated emotional state to cope with stress and to be available for learning and interacting; (iii) *Transactional Support*: support to help caregivers respond to the childs needs and interests, adapt the environment, and provide tools to enhance learning. SCERTS breaks down each domain into a number of essential constitutive components. Then, for each component, it provides a detailed description of the education objectives to be achieved, the strategies for intervention and the assessment criteria. We build on this operationalisation of social communication for designing the agent's behaviour.

The interaction between the child and the agent is structured around twelve different *learning activities* and is facilitated by a large multitouch LCD display. The learning activities focus on social communication and, in particular, on the two sub-components of social communication that have been identified by SCERTS as the most challenging for ASCs children: (i) *Joint Attention:* ability to coordinate and share attention and emotions, express intentions, and engage in reciprocal social interactions by initiating/responding to bids for interaction; and (ii) *Symbol Use:* understanding of meaning expressed through conventional gestures and words and ability to use nonverbal means to share intentions.

4 The Design and Implementation of the ECHOES Agent

We designed an artificial social partner, called Andy, that can act credibly both as a peer and as a tutor to ASCs children based on (a) the SCERTS model and (b) input from thirty ASCs practitioners, who participated in two workshops organised over the lifespan of the ECHOES project and involving storyboarding tools, group discussions and individual interviews. We now describe the main design choices that we have made to create Andy.

Agent's Intelligence: Among the various domain-independent agent architectures that have been proposed for building agents, FAtiMA [3] is ideally suited to fulfil the design requirements of our agent, because it combines the kind of *reactive* and *cognitive* capabilities needed to implement an autonomous, proactive and reactive agent with the *socio-emotional* competence that we envisaged for our agent in this context. The cognitive layer of FAtiMA is based on artificial intelligence *planning* techniques [9], while the emotional model is derived from the OCC theory of emotions [5] and the appraisal theory [10]. A FAtiMA agent is characterised by: (1) a set of internal goals; (2) a set of action strategies to achieve these goals; and (3) an affective system composed of emotional reaction rules, action tendencies, emotional thresholds and emotion decay rates. The two main mechanisms controlling a FAtiMA agent are *appraisal* and *coping*. The agent experiences one or more of the 22 emotions of the OCC model, based on its appraisal of the current external events and its subjective tendencies to experience certain emotions instead of others. The agent deals with these emotions by applying problem-focused or emotion-focused coping strategies. Both the appraisal and the coping work at two different levels: the *reactive* level, which affects the short-term horizon of the agent's behaviour, and the *deliberative* level, which relates to the agent's goal-oriented behaviour.

In ECHOES , each learning activity has an associated FAtiMA agent model. All these models share the same specification of the agent's affective system, because we want the agent to maintain the same personality from session to session in order to establish a trusting relationship with the child. Andy is a positive, motivating and supportive character that tends to be happy and does not get frustrated easily. We obtained such behaviour by manipulating Andy's emotional reaction rules as well as the emotional thresholds and decay rates of the OCC emotions available in FAtiMA. We control Andy's facial expressions and gestures through the specification of Andy's action tendencies. For example, the agent smiles when it is happy and opens its mouth when it is surprised. While Andy's personality does not change between activities, the goals that the ECHOES agent actively tries to pursue and its action strategies are specified for each learning activity based on: (i) the high-level *pedagogical goals* on which the activity focuses and (ii) the specific *narrative content* of the activity itself.

Repertoire of the Agent's Behaviours: Since the focus of our environment is on supporting children's social communication, the agent's actions are either concrete demonstrations of the related skills or actions performed to trigger the child to practice those skills. Specifically, we define the joint attention and symbolic use in terms of three component skills: (i) *Responding* to bids for interaction; (ii) *Initiating* bids for interaction; and (iii) Engaging in *turn taking*. Our agent is able to perform these skills in three different ways: (a) Verbally by using simple language or key words; (b) Non-verbally through gaze and gestures such as pointing at an object or touching an object; and (c) By combining verbal and non-verbal behaviours. Initiating a non-verbal bid for interaction by the agent involves Andy looking at the child, then looking at and indicating an object and then looking back at the child. Our agent is able to make requests, to greet the child by name, to comment on events happening in the garden and to use exploratory actions on objects. This variety of behaviours makes the interaction dynamic enough to keep the child engaged and to foster generalisation, while retaining a degree of predictability that is essential to supporting the childs attentional focus. The practitioners emphasised the importance of providing the children with *positive feedback* in order to reduce the children's anxiety of social interactions and to help them experience a sense of self-efficacy. Andy always provides the child with positive feedback, especially if the child follows the agent's bids for interaction correctly. If the child does not perform the required action, the agent first waits for the child to do things at their own pace and then intervenes by demonstrating the action. Andy is a *responsive* agent. Its responsiveness ranges from physical reactions to actions performed on its body (e.g., Andy laughs if the child tickles it), to the ability to respond to the child's changing needs and cognitive and emotional states. Ideally, the agent should attune its emotional tone to that of the child in order to keep emotional engagement with the child. Such a sophisticated level of reactiveness requires that the agent is able to assess in real-time the current cognitive and emotional state of the user. Our current system includes a simple *user model*, which evaluates the child's current state based on the real-time information from the touch system.

Agent's Physical Appearance and Communication: Following SCERTS' emphasis on creating opportunities for children with ASCs to play with other children and develop positive relationships with them [6], we chose a *child-like* physical appearance for the agent. Based on input from the practitioners, we gave our agent a *cartoonish* look and animated it based on observations of children's cartoons in order to motivate the children.

Our agent uses (a fragment of) the *Makaton* language [12] to facilitate communication with the child. Makaton is a language programme of signs and symbols to support spoken language and is used with speech. Since auditory information can be challenging for children with ASCs, we kept the *spoken language* as simple as possible. The agent can perform a good range of positive *facial expressions*: it can smile, laugh, look surprised, happy, and excited. These expressions are implemented by careful use of movements in the agent's lips, eyes and eyebrows. They are usually accompanied by *body gestures* consistent with emotion showed through the face in order to reinforce the message.

5 Experimental Results

To assess the impact of Andy and ECHOES on social communication in children with ASCs, a large scale multi-site intervention study was conducted. The system was deployed in five schools in the UK. 19 children with ASCs participated in the study, during which they played with ECHOES for 10 to 20 minutes, several times a week over an eight week period. The SCERTS Assessment Protocol (SAP) [7] was modified into a finer-grained coding scheme that could be applied to videos of children's interactions with Andy. The modified SAP coding scheme contains 16 main behavioural categories. Due to space limitations, we will focus only on two main social behaviours, which are severely impaired in children with ASCs: responding to and initiating bids for interaction. Fifteen minute periods during which the children interacted with Andy were identified for analysis from the beginning, middle and end of the intervention period. Each video was blind-coded by a coder trained in the modified SAP coding scheme. Video annotations were applied in "Elan" and moderated by a second coder.

Child's Response to Agent: The mean probability that a child responded to the practitioner's bids for interaction during the table-top pre-test was 0.62 (SD=0.19) and 0.71 (SD=0.14) after the intervention (slight increase was not significant t(14)=-1.644, p=.123). Across the three ECHOES sessions the probability of responding to Andy's initiations decreased slightly: beginning = 0.57 (SD=0.22), middle = 0.53 (SD 0.25), end = 0.49 (SD 0.24); although this decrease was not significant. However, the probability of responding to the human practitioner's bids for interaction increased marginally across the three sessions: beginning = 0.74 (SD=0.21), middle = 0.75 (SD 0.21), end = 0.81 (SD 0.21), and neared significance: beginning vs. end, t(18)=-2.017, p=.059. This increase may suggest a comfort with the ECHOES environment that elicited a level of responsiveness in the child not observed during the table-top activity. As Andy is a critical part of the ECHOES environment we can assume his presence contributed to this improvement.

Child's Initiations to Agent: A more advanced social behaviour rarely observed in ASCs children is the initiation of social interactions. The frequency with which the child initiated an interaction during the table-top pre-test was low, 8.63 times (SD 7.94) and did not change by the post table-top session, 7.84 times (SD 10.0): t(18)=.456, p=.654, n.s. Although there was no improvement in the real-world scenario, initiations to Andy numerically increased across the three ECHOES sessions from 4.98 (SD 8.05) to 6.68 (SD 7.68) and 9.58 (SD 13.67). Unfortunately, this group difference did not reach significance (t(18)=-1.699, p=.106, n.s.) even though eight children increased their number of initiations to Andy, seven produced the same number and only four decreased. This suggests that the heterogeneity in our ASCs population may obscure a group increase. For a number of children Andy appears to be eliciting a large increase in spontaneous initiations. This is strikingly obvious when examining videos of the ECHOES sessions. For example, one child who showed no initial interest in Andy spontaneously waved and said "Hi Andy!" when the agent walked on the screen in a later session. Such behaviours were extremely surprising to teachers within the school who believed the child in question to be non-communicative.

6 Conclusions and Future Work

In this paper, we presented our approach to designing and implementing an autonomous pedagogical agent for supporting social communication skills of children with ASCs based on principled intervention guidelines, recommendations from autism practitioners and children themselves. We undertook a major evaluation of the ECHOES system involving a significant number of children. A preliminary analysis of child behaviours in relation to Andy shows that, whilst no statistically significant differences in social communication have been observed across all children, some children benefited from their exposure to Andy and the ECHOES environment. Andy's reciprocal interactions with the children appear to elicit spontaneous social behaviours. With improved versions of the system, especially focusing on a more comprehensive user modelling capabilities, we hope that the trend of improvements in the social behaviours manifested by the children will prove significant.

References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. 4th edn., Text Revision (DSM-IV-TR) (2000)
2. Bosseler, A., Massaro, D.: Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. Journal of Autism and Developmental Disorders 33(6), 653–672 (2003)
3. 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)
4. Milne, M., Luerssen, M., Lewis, T., Leibbrandt, R., Powers, D.: Development of a virtual agent based social tutor for children with autism spectrum disorders. In: Proc. of the International Joint Conference on Neural Networks, pp. 1–9 (2010)
5. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press (1988)
6. Prizant, B., Wetherby, A., Rubin, E., Laurent, A.: The scerts model: A transactional, family-centered approach to enhancing communication and socioemotional ability in children with autism spectrum disorder. Infants and Young Children 16(4), 296–316 (2003)
7. Prizant, B., Wetherby, A., Rubin, E., Laurent, A., Rydell, P.: The SCERTS® Model: A Comprehensive Educational Approach for Children with Autism Spectrum Disorders. Brookes (2006)
8. Putnam, C., Chong, L.: Software and technologies designed for people with autism: what do users want? In: 10th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 3–8 (2008)
9. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Prentice Hall (2003)
10. Smith, C.A., Lazarus, R.S.: Emotion and adaptation. In: Vlahavas, I., Vrakas, D. (eds.) Handbook of Personality: Theory and Research, ch. 23, pp. 609–637. Guilford, New York (1990) L.A. Pervin (Ed.)
11. Tartaro, A., Cassell, J.: Playing with virtual peers: bootstrapping contingent discourse in children with autism. In: Proc. ICLS 2008, pp. 382–389 (2008)
12. Walker, M., Armfield, A.: What is the makaton vocabulary? Special Education: Forward Trends 8(3), 19–20 (1981)
13. Wooldridge, M., Jennings, N.R.: Intelligent agents: Theory and practice. Knowledge Engineering Review 10(2), 115–152 (1995)

The Effect of Virtual Agents' Emotion Displays and Appraisals on People's Decision Making in Negotiation

Celso M. de Melo[1], Peter Carnevale[2], and Jonathan Gratch[1]

[1] Institute for Creative Technologies, USC, 12015 Waterfront Drive, Building #4 Playa Vista,
CA 90094-2536, USA
{demelo,gratch}@ict.usc.edu
[2] USC Marshall School of Business, Los Angeles, CA 90089-0808, USA
peter.carnevale@marshall.usc.edu

Abstract. There is growing evidence that emotion displays can impact people's decision making in negotiation. However, despite increasing interest in AI and HCI on negotiation as a means to resolve differences between humans and agents, emotion has been largely ignored. We explore how emotion displays in virtual agents impact people's decision making in human-agent negotiation. This paper presents an experiment (N=204) that studies the effects of virtual agents' displays of joy, sadness, anger and guilt on people's decision to counteroffer, accept or drop out from the negotiation, as well as on people's expectations about the agents' decisions. The paper also presents evidence for a mechanism underlying such effects based on appraisal theories of emotion whereby people retrieve, from emotion displays, information about how the agent is appraising the ongoing interaction and, from this information, infer about the agent's intentions and reach decisions themselves. We discuss implications for the design of intelligent virtual agents that can negotiate effectively.

Keywords: Emotion Displays, Decision Making, Negotiation, Appraisal Theories, Reverse Appraisal.

1 Introduction

Negotiation, defined by Pruitt and Carnevale [1] as "a discussion among two or more parties aimed at reaching agreement when there is a perceived divergence of interest", is a common mechanism people use for conflict resolution. Negotiation has, thus, drawn considerable attention in the artificial intelligence and human-computer interaction fields as means to resolve conflict in agent-agent and human-agent interaction [2-4]. Automated computer agents can assist less qualified individuals in the negotiation process and, in some situations, even replace human negotiators altogether. Virtual agents, by virtue of having bodies, can add further dimensions to these automated negotiators. One such dimension is the expression of emotions. Effectively, recent decades have seen growing interest on the interpersonal effect of emotion in negotiation (e.g., [5-7]). Complementing research on the impact of emotion in one's own decision making (e.g., [8]), this research explores how one's emotion displays impact

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 53–66, 2012.

another's decision making and emphasizes that emotional expressions are not simple manifestations of internal experience; rather, expressions are other-directed and communicate intentions, desired courses of actions, expectations and behaviors [9-11]. There is now plenty of empirical evidence demonstrating the differentiated effects of emotions on negotiation outcome (e.g., [5, 12-14]). However, despite the interest artificial intelligence and human-computer interaction has shown in automating negotiation, emotion has been notoriously absent in this endeavor.

To study the social effects of emotion displays in virtual agents on people's decision making in negotiation we build on seminal work by Van Kleef, De Dreu and Manstead on the effects of joy, sadness, anger and guilt on people's concession-making in multi-issue bargaining [12-14]. These studies showed that people conceded more to angry or sad counterparts than happy or guilty counterparts. When facing an angry counterpart, people were argued to infer the other to have high limits and, thus, to avoid costly impasse, were forced to concede; when facing a sad counterpart, people tried to relieve the other's pain by making concessions; when facing a happy counterpart, people inferred the other to have low limits and, thus, could afford to be strategically more demanding; finally, when facing a guilty counterpart, people inferred the other was trying to make amends for a previous transgression and, thus, were more demanding with them. In the past, we have shown that the effects of anger and joy can also occur in human-agent negotiation when emotion is displayed by virtual agents [15]. In the present paper, we complement this work and further study the effects of sadness and guilt in human-agent negotiation. Moreover, in real-life negotiation, people usually can, aside from making counteroffers, accept the counterpart's offer or drop out from the negotiation and it is important to understand when such decisions are made [1]. Therefore, in contrast to the previous study, here we also study how emotion displays impact people's decision to accept or drop out from the negotiation. The paper presents a novel experiment where participants imagined engaging in negotiation with virtual agents that reacted emotionally to the participants' initial offer. We measure how agents' displays of joy, sadness, anger or guilt impact people's decisions to counteroffer, accept or drop out from the negotiation. We also measure how these displays impact people's expectations about the agents' behavior.

This paper also studies the *mechanism* underlying the effects of agents' emotion displays on people's decision making in negotiation. Recently, we proposed a mechanism for the social effects of emotion in the prisoner's dilemma based on appraisal theories of emotion [16]. In appraisal theories [17], emotion displays arise from cognitive appraisal of events with respect to one's goals (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced and displayed. Since displays reflect the agent's intentions through the appraisal process, we argued people could infer, from emotion displays, how virtual agents were appraising the ongoing interaction and, from this information, make inferences about the agents' intentions. This view is compatible with Hareli and Hess' [18] findings that people can, from emotion displays, "reverse engineer" the appraisal process and, from information about appraisals, make inferences about someone's character. This proposal was, thus, referred to as reverse appraisal. Indeed, the results showed that perception of how the agent was appraising the interaction mediated the effect of emotion displays on perception of the agent's intention to cooperate in the prisoner's dilemma. Moreover, in a follow-up experiment, we showed that identical effects could be achieved if, instead of expressing emotions in the face, agents expressed how they were appraising the interaction directly through text (e.g., "I really don't like this outcome and I blame you for it."). This result, thus, further supports reverse appraisal. Generalizing these findings, reverse appraisal suggests that what matters for the social effects of emotion in decision making are *not* the displays per se but, the information they communicate about the agent's appraisals. Applying this to negotiation, reverse appraisal suggests that the effects of displays on people's decision making in negotiation should also be mediated by people's perceptions of the agents' appraisals. Therefore, in this paper we present a multiple mediation analysis [19] that tests whether reverse appraisal can explain the social effects of emotion in human-agent negotiation.

2 Experiment

Scenarios. Participants imagined engaging in a standard multi-issue bargaining task [12] with virtual agents that reacted emotionally to the participant's initial offer. In this task, the participant played the role of a seller of a consignment of mobile phones whose goal was to negotiate on three issues: price, warranty period and service contract. Each issue had 9 levels, being the highest level the most valuable for the participant, and the lowest level the least valuable. Level 1 on price ($110) yielded 0 points and level 9 ($150) yielded 400 points (i.e., each level corresponded to a 50 point increment). Level 1 on warranty (9 months) yielded 0 points and level 9 (1 month) yielded 120 points (i.e., each level corresponded to a 15 point increment). Finally, for duration of service contract, level 1 (9 months) yielded 0 points, and level 9 (1 month) yielded 240 points (i.e., each level corresponded to a 30 point increment). It was pointed out to the participant that the best deal was, thus, 9-9-9 for a total outcome of 760 points (400 + 120 + 240). The participant engaged in this task with a virtual agent, referred to as David, which had different payoffs that were not known. The negotiation supposedly proceeded according to the alternating offers protocol, being the participant the first to make an offer. The participant was informed that the negotiation would proceed until one player *accepted* the other's offer, *dropped out* from the negotiation or time expired. In reality, the scenario pertained only to the first offer made by the participant which corresponded to 9-6-7, worth 400 + 75 + 180 = 655 points to the participant. After being shown the offer, the participant watched a video of the agent reacting emotionally. After watching the video, the participant was asked several questions and, then, explained that the scenario was over.

Conditions. The experiment had one between-participants factor: Emotion Display (Neutral vs. Joy vs. Sadness vs. Anger vs. Guilt). The emotion displays were shown in the face of a virtual agent and the respective displays are shown in Figure 1. These facial displays were animated using a muscular model of the face with blushing and wrinkles [20].

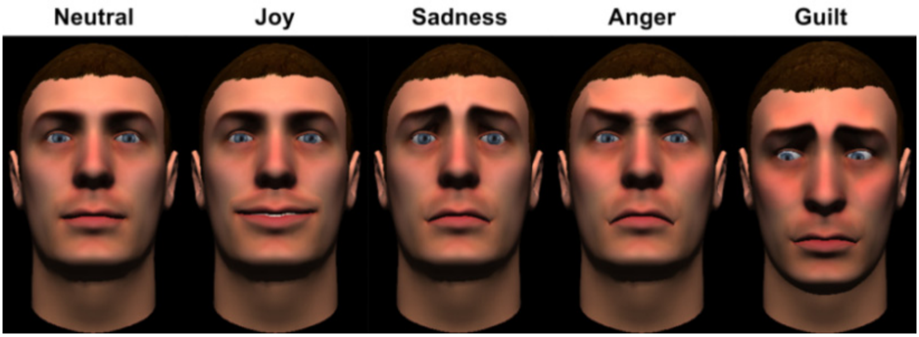


Fig. 1. The facial displays of emotion

Measures. After watching the video of the agent's emotional reaction, we asked participants the following questions: How much did David experience each of the following emotions a) Sadness b) Joy c) Anger d) Guilt? (scale went from 1, *not at all*, to 7, *very much*).

Even though several appraisal theories have been proposed [17, 21-23], there tends to be agreement on which appraisals predict the emotions we considered in this experiment: joy occurs when the event is conducive to one's goals; sadness occurs when the event is not conducive to one's goals; anger occurs when the event is not conducive to one's goals and is caused by another agent; guilt occurs when the event is not conducive to one's goals and is caused by the self. Thus, two appraisal variables are of relevance here: (a) *conduciveness to goals*, which measures whether the event is consistent or inconsistent with the individual's goals; and, (b) *blameworthiness*, which measures whether the self or another agent is responsible for the event. After watching the video of the agent's emotional reaction, participants were asked the following questions about how was the agent appraising the outcome (all questions on a scale from 1, *not at all*, to 7, *very much*):

1. How pleasant for David was it to be in this situation? [21]
2. At the time of experiencing the emotion, do you think David perceived that the consequences of the event did or would bring about positive, desirable consequences for him (e.g., helping him reach a goal, or giving pleasure)? [22]
3. Was the situation obstructive or conducive to David's goals? [23]
4. Was what happened something that David regarded as fair? [23]
5. How much did you think David blamed himself for the event? [21]
6. How much did you think David blamed you for the event? [21]

Following the appraisal perception questions, we asked several questions about the participant's decision and expectations regarding the agent's decision (Questions 7, 8, 11 and 12 were on a scale from 1, *not at all*, to 7, *very much*):

7. How likely is David to accept your offer?
8. How likely is David to drop out from the negotiation?
9. If David were to make a counter-offer, what would that likely be?
10. How likely are you to accept David's counter-offer?

11. How likely are you to drop out from the negotiation?
12. If you were to make a counter-offer what would that be?

Finally, two questions were asked that characterized the interaction with the agent (scale went from 1, *not at all*, to 7, *very much*):

13. How cooperative was David?
14. Would you negotiate again with David in a future occasion?

Participants. We recruited 204 participants online using Amazon Mechanical Turk. This resulted in approximately 41 participants for each emotion. Regarding gender, 56.4% of the participants were male. Age distribution was as follows: *18 to 21 years*, 16.2%; *22 to 34 years*, 56.9%; *35 to 44 years*, 15.7%; *45 to 54 years*, 7.4%; *55 and over*, 3.9%. Most participants were from the United States (40.7%) and India (42.2%). The education level distribution was as follows: *high school*, 18.6%; *college*, 57.8%; *Masters*, 19.6%; *Ph.D. or above*, 3.9%. Education majors and profession were quite diverse. Participants were paid USD $1.02 and average participation time was 15 minutes.

3 Results

Manipulation Check. To validate that participants took the task seriously we looked at outliers and performance measures (e.g., participation time). No participants were excluded under these criteria. To validate that participants were interpreting the virtual agent's facial displays as intended, we subjected the emotion interpretation measures to a MANOVA across Emotion Display. Table 1 shows the means and standard deviations for this analysis. The results for the multivariate test showed a significant effect ($p < .05$, Pillai's Trace). ANOVAs confirmed significant differences for the emotion interpretation measures (see last column of Table 1), and subsequent Bonferroni post-hoc tests revealed that: perception of joy was highest for Joy, perception of sadness was highest for Sadness, perception of anger was highest for Anger and, perception of guilt was highest for Guilt (all tests, $p < .05$). Thus, the agents' emotion displays were being interpreted as intended.

Table 1. Perceived emotion in the facial displays (manipulation check)

	Emotion Display					
Perceived Emotion	Neutral	Joy	Sadness	Anger	Guilt	ANOVA
Joy	2.20	5.71	1.95	2.10	1.98	66.63*
	(1.38)	(1.08)	(1.28)	(1.53)	(1.07)	
Sadness	2.00	1.34	5.27	3.32	4.53	52.25*
	(1.32)	(0.88)	(1.36)	(1.89)	(1.65)	
Anger	1.93	1.34	2.41	4.51	2.48	35.78*
	(1.39)	(0.86)	(1.50)	(1.38)	(1.18)	
Guilt	1.93	1.24	2.59	2.76	3.78	17.95*
	(1.23)	(0.66)	(1.52)	(1.92)	(1.51)	

* $p < .0125$ (Bonferroni correction applied).

Effects on Decision Measures. To study the participant's counteroffer and the agent's expected counteroffer we looked at *demand difference*, a standard measure that is defined as the difference between demand (i.e., the value of the offer in points) in the participant's first offer (655 points) and the counteroffer. The means and standard errors for demand difference are shown in Figure 2. The means and standard errors for the remaining decision measures and agent characterization measures are shown in Figure 3. Table 2 summarizes descriptive statistics for all these measures. To analyze these results we ran one-way ANOVAs, across Emotion Display, for each measure. Tukey HSD post-hoc tests were used for pairwise comparisons.

Regarding participant's demand difference, there was a (main) effect of Emotion Display ($F(4,199) = 2.85$, $p < .05$) and post-hoc tests revealed that demand difference was lower after the agent displayed Joy than Anger ($p < .05$) or Guilt ($p < .10$). Regarding (expected) agent's demand difference, there was an effect of Emotion Display ($F(4,199) = 6.02$, $p < .05$) and post-hoc tests revealed that expected demand difference was lower after the agent displayed Joy than Anger ($p < .05$), Guilt ($p < .05$) or Sadness ($p < .10$).

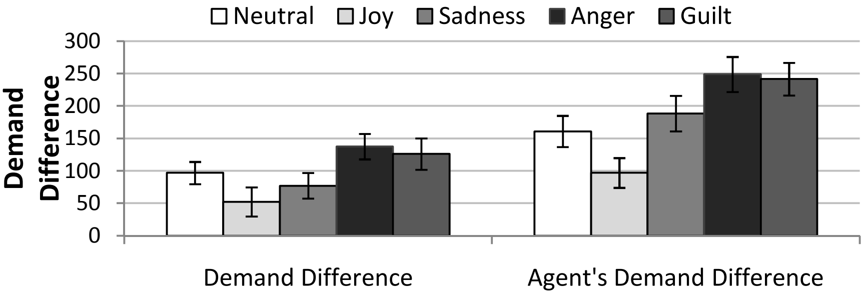


Fig. 2. Means (and standard errors) for demand difference (difference in points between the participant's first offer and the counteroffer)

Regarding the participant's decision to accept, there was a (main) effect of Emotion Display ($F(4,199) = 5.47$, $p < .05$) and post-hoc tests revealed the participant was more likely to accept after Joy than Sadness, Anger or Guilt (all tests, $p < .05$). Regarding the (expected) agent's decision to accept, there was an effect of Emotion Display ($F(4,199) = 48.19$, $p < .05$) and post-hoc tests revealed that the agent was more likely to accept after displaying Joy than the Neutral display and, less likely to accept after displaying Sadness, Anger or Guilt than the Neutral display (all tests, $p < .05$). Regarding the participant's decision to drop out, there was an effect of Emotion Display ($F(4,199) = 2.86$, $p < .05$) and post-hoc tests revealed the participant was more likely to drop out after the a display of Sadness than Joy ($p < .05$). Regarding the (expected) agent's decision to drop out, there was an effect of Emotion Display ($F(4,199) = 19.52$, $p < .05$) and post-hoc tests revealed that the agent was more likely to drop out after displaying Sadness, Anger or Guilt than the Neutral display and, less likely to drop out after displaying Joy than the Neutral display (all tests, $p < .05$).

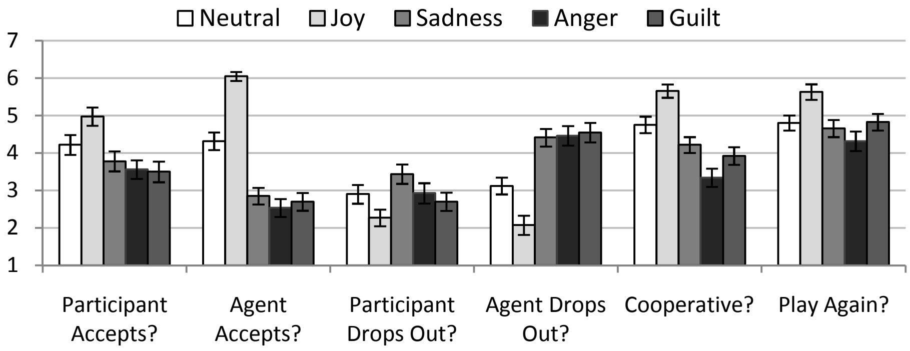


Fig. 3. Means (and standard errors) for accept, drop out and characterization measures

Table 2. Means (and standard deviations) for decision measures

	Neutral	Joy	Sadness	Anger	Guilt
Participant's Demand Difference *	96.95 (109.15)	52.20 (143.68)	77.07 (126.63)	137.56 (125.21)	126.13 (152.96)
Agent's Demand Difference *	160.98 (154.35)	96.95 (147.02)	188.17 (177.96)	248.90 (173.10)	241.63 (158.78)
Participant Accepts *	4.22 (1.71)	4.98 (1.56)	3.78 (1.71)	3.56 (1.60)	3.50 (1.75)
Agent Accepts *	4.32 (1.51)	6.05 (0.77)	2.85 (1.42)	2.54 (1.53)	2.70 (1.51)
Participant Drops Out *	2.90 (1.61)	2.27 (1.43)	3.44 (1.66)	2.93 (1.74)	2.70 (1.54)
Agent Drops Out *	3.12 (1.44)	2.07 (1.65)	4.41 (1.50)	4.46 (1.67)	4.55 (1.65)
Play Again *	4.80 (1.29)	5.63 (1.32)	4.66 (1.46)	4.32 (1.66)	4.83 (1.41)
Cooperative *	4.76 (1.41)	5.66 (1.13)	4.22 (1.35)	3.34 (1.56)	3.93 (1.49)

Note. Demand difference is the difference in points between the participant's first offer (9-6-7) and the counteroffer.
* $p < .05$.

Finally, Regarding the participant's willingness to play again with the agent, there was a (main) effect of Emotion Display ($F(4,199) = 4.67$, $p < .05$) and post-hoc tests revealed that the participant was more likely to play again with the agent that displayed Joy than Neutral ($p < .10$), Sadness ($p < .05$), Anger ($p < .05$) and Guilt ($p < .10$). Regarding perception of how cooperative the agent was, there was an effect of Emotion Display ($F(4,199) = 16.21$, $p < .05$) and post-hoc tests revealed that the agent that displayed Joy was perceived as more cooperative than the agent that showed the

Neutral display ($p < .05$) and the agent that showed Anger ($p < .05$) or Guilt ($p < .10$) was perceived to be less cooperative than the agent that showed the Neutral display.

Effects on Perception of Appraisals Measures. Questions 1 to 4 were highly correlated (α = .903) and, thus, were collapsed (averaged) into a single measure called conduciveness to goals. The means, standard deviations and standard errors for conduciveness to goals, self-blame (Question 5) and participant-blame (Question 6) are displayed in Figure 4 and presented in Table 3. For our main analysis, we subjected the perception of appraisal measures to a one-way ANOVA across Emotion Display. Tukey HSD post-hoc tests were used for pairwise comparisons.

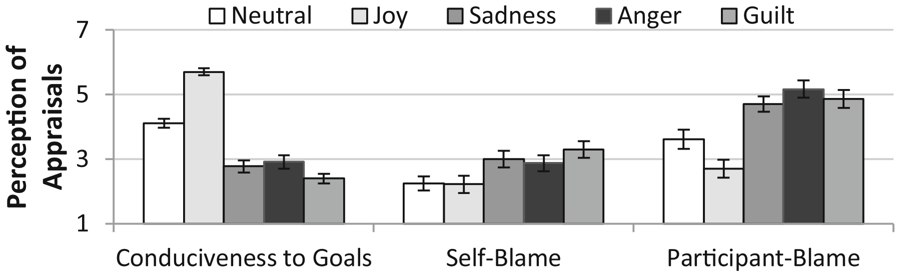


Fig. 4. Means (and standard errors) for perception of appraisals

Table 3. Means (and standard deviations) for perception of appraisals

	Neutral	Joy	Sadness	Anger	Guilt
Conduciveness to Goals *	4.11	5.71	2.77	2.92	2.40
	(0.92)	(0.68)	(1.26)	(1.33)	(0.97)
Self-Blame *	2.24	2.22	3.00	2.88	3.30
	(1.43)	(1.68)	(1.63)	(1.58)	(1.68)
Participant-Blame *	3.61	2.71	4.71	5.17	4.88
	(1.90)	(1.76)	(1.52)	(1.67)	(1.76)

* $p < .05$.

Regarding conduciveness to goals, the results showed a (main) effect of Emotion Display ($F(4, 199) = 66.29$, $p < .05$). Post-hoc tests revealed that when the agent displayed Joy, it was perceived to find the outcome more conducive than when it showed the Neutral display; moreover, when the agent displayed Sadness, Anger or Guilt, the agent was perceived to find the outcome less conducive than when it showed the Neutral display (all tests, $p < .05$). Regarding self-blame, the results showed an effect of Emotion Display ($F(4, 199) = 3.62$, $p < .05$). Post-hoc tests revealed that when the agent displayed Guilt it was perceived to blame itself more than when it showed Joy or the Neutral Display (all tests, $p < .05$). Regarding participant-blame, the results showed an effect of Emotion Display ($F(4, 199) = 14.52$, $p < .05$). Post-hoc tests revealed that when the agent displayed Joy it was perceived to blame the participant the least and, when it showed Sadness, Anger or Guilt it was perceived to blame the participant more than when it displayed the Neutral display (all tests, $p < .05$).

Mediation Analysis. Here we present a causal steps approach multiple mediation analysis [19] of perceptions of appraisal on the effect of emotion displays on participant's demand difference and (expected) agent's demand difference. This method is an extension to multiple mediators of the single-mediation analysis proposed by Baron and Kenny [24]. Figure 5 summarizes the mediation model. The independent variables (IVs) were the classification questions for perception of joy, sadness, anger and guilt. The dependent variables (DV) were the demand difference measures. The proposed mediators were the perception of appraisal variables: conduciveness to goals, self-blame and participant-blame. According to this approach, there is mediation by a specific mediator M_x if: (1) the path, a_x, from the IV to the mediator is significant; (2) the path, b_x, from the mediator to the DV, when controlling for the IV, is significant; (3) the indirect effect, a_xb_x, from the IV to the DV, when controlling for the mediator, is significantly different than zero and greater than zero by a non-trivial amount. Moreover, there is mediation of the *set* of mediators when the sum of the indirect effects of all mediators is significantly different than zero. There is full mediation when the total effect, *c*, from the IV to the DV (not considering any mediators), becomes non-significant, i.e., the direct effect, *c'*, from the IV to the DV (when accounting for all mediators), is non-significant. Finally, in the original paper, Baron and Kenny also require the total effect, *c*, to be significant. However, many authors advocate this path need not be significant, in the multiple mediation case, for mediation to occur [19].

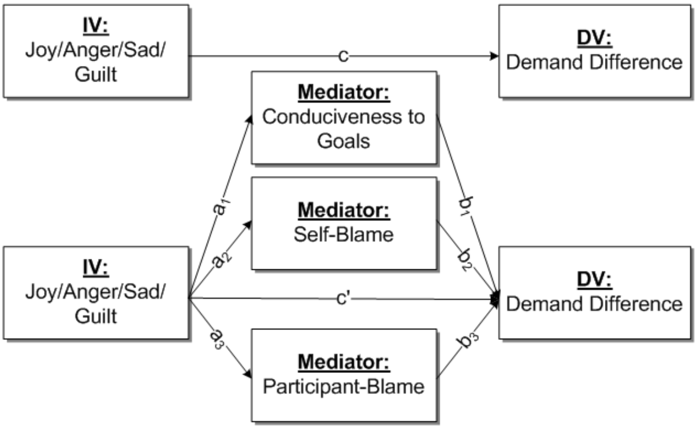


Fig. 5. The multiple mediation model

Table 4 shows the analysis: the shaded cells on the *a*, *b* and *ab* path columns represent that the causal-step requirement on the respective path has been passed. Regarding the participant's demand difference, the results showed full mediation of conduciveness to goals on the effect of Joy, Sadness and Anger; there was also evidence of mediation of conduciveness to goals and self-blame on the effect of Guilt. Regarding (expected) agent's demand difference, the results showed full mediation of conduciveness to goals and participant-blame on the effect of Joy, Sadness and Anger; there was also evidence of mediation of conduciveness to goals on the effect of Guilt.

Table 4. Mediation analysis of perceptions of appraisals on the effect of emotion displays

		IV→ Mediators			Mediators → DV			Total	Direct	Indirect Effect			
		(a paths)			*(b paths)*			Effect	Effect	*(ab paths)*			
		Cn	SB	PB	Cn	SB	PB	*(c path)*	*(c' path)*	Tot	Cn	SB	PB
Participant's Demand Difference	Joy	.62**	-.01	-.36**	-21.15**	-5.03	-1.70	-9.42**	3.20	-12.62**	-13.36**	.21	.53
		(.000)	(.829)	(.000)	(.004)	(.227)	(.670)	(.008)	(.567)	(.004)	(.003)	(.440)	(.669)
	Sad	-.39**	.22**	.29**	-17.11**	-5.65	-1.81	6.08*	1.69	4.39**	6.11**	-1.32	-.41
		(.000)	(.000)	(.000)	(.001)	(.198)	(.651)	(.066)	(.665)	(.048)	(.002)	(.204)	(.650)
	Anger	-.29**	.12*	.42**	-17.01**	-5.67	-2.62	7.58*	4.28	3.29*	4.95**	-.74	-.92
		(.000)	(.077)	(.000)	(.001)	(.178)	(.521)	(.052)	(.310)	(.077)	(.002)	(.220)	(.521)
	Guilt	-.19**	.55**	.14*	-17.21**	-7.7*	-1.82	4.56	5.31	-.75	3.03**	-3.60*	-.18
		(.005)	(.000)	(.097)	(.000)	(.099)	(.648)	(.264)	(.256)	(.766)	(.008)	(.098)	(.656)
Agent's Demand Difference	Joy	.62**	-.01	-.36**	-32.95**	-11.72*	12.21*	-31.93**	-7.16	-24.77**	-20.55**	.15	-4.37*
		(.000)	(.829)	(.000)	(.004)	(.074)	(.055)	(.000)	(.400)	(.000)	(.004)	(.830)	(.066)
	Sad	-.39**	.22**	.29**	-44.53**	-9.60	12.71**	10.93*	-7.98	18.91**	17.27**	-2.06	3.69*
		(.000)	(.000)	(.000)	(.000)	(.157)	(.046)	(.059)	(.201)	(.000)	(.000)	(.177)	(.063)
	Anger	-.29**	.12*	.42**	-39.74**	-12.09*	11.85*	15.83**	.93	14.90**	11.37**	-1.49	5.01*
		(.000)	(.077)	(.000)	(.000)	(.067)	(.072)	(.028)	(.895)	(.000)	(.001)	(.198)	(.083)
	Guilt	-.19**	.55**	.14*	-41.14**	-7.79	12.13*	-2.60	-7.77	5.17	7.81**	-4.31	1.67
		(.005)	(.000)	(.097)	(.000)	(.319)	(.056)	(.723)	(.330)	(.363)	(.013)	(.319)	(.206)

Note. Cn = Conduciveness to goals; SB = Self-Blame; PB = Participant-Blame; CP = Coping Potential.
Demand difference is the difference in points between the participant's first offer (9-6-7) and the counteroffer.
Values correspond to unstandardized regression coefficients (*p* values in parentheses).
** $p < .05$. * $p < .10$.

4 Discussion

The results showed that emotion displays in virtual agents can impact people's decision making in negotiation. In line with previous findings [12, 15], our results indicated that people conceded more to an angry agent than to a happy agent. The rationale here is that people infer the angry agent to have high aspirations and, thus, to avoid costly impasse are forced to concede; in contrast, people infer the happy agent to have low aspirations and, therefore, can afford to be strategically more demanding. However, our results suggested displays of guilt led people to concede more. This finding seemingly contrasts with Van Kleef et al.'s findings [13] that suggest people concede less to a guilty counterpart because guilt is signaling an apology and a willingness to make amends for a previous transgression. However, our scenarios pertained to an emotional reaction to the participant's first offer and, therefore, there is no obvious previous transgression. In this case, participants are more likely interpreting the display of guilt as serving a supplication function (i.e., a cry for help) and, therefore, comply by helping the agent by making more generous offers. These different interpretations of guilt reinforce the importance of context for the interpretation of emotion and the social effects of emotion displays [6, 14, 25].

Regarding the decisions to accept or drop out, the results showed that people were more likely to accept and expect the agent to accept after a display of joy than a display of sadness, anger or guilt; complementary, people were more likely to drop out from the negotiation or expect the agent to drop out after a display of sadness, anger or guilt than a display of joy. Regarding characterization, the results indicated people, if given a choice, would likely negotiate again with the smiling agent, which they found to be the most cooperative, and would likely not negotiate again with the angry agent, which they found to be the least cooperative.

To understand the mechanism behind such effects, we looked at the reverse appraisal proposal [16, 18] whereby people retrieve, from emotion displays, information about how agents are appraising the ongoing interaction and, from this information, infer the agents' intentions. Effectively, our results confirmed that people were able to retrieve, from emotion displays, information about agent's appraisals: the offer was perceived to be more conducive to the agent's goals after a display of joy than a display of sadness, anger or guilt; the agent was perceived to blame more itself after a display of guilt; and, the agent was perceived to blame the participant more after displaying anger, guilt or sadness than joy. These results are, in general, compatible with expectations from appraisal theories [17] and with previous findings that show people can retrieve information about appraisals from facial displays of emotion [26]. Moreover, our results show that perceptions of the agents' appraisals mediated the effects of emotion displays on people's decisions and expectations about the agents' decisions. For instance, the effect of joy or anger on people's demand was fully mediated by perceptions about how conducive the offer was to the agent. This is compatible with Van Kleef et al.'s argument that people use displays of joy or anger to infer the counterpart's limits in negotiation [12]. These results, thus, suggest that information about the agents' appraisals are a critical component people retrieve from emotion displays to inform their decision making in negotiation.

Overall, the results emphasize the importance of valence for the social effects of emotion displays in human-agent negotiation. Whereas it is easy to distinguish between the effects of joy and the negative displays, the differences between the effects of sadness, anger and guilt are more subtle. The mediation analysis also lends some support to this dichotomy in that conduciveness to goals, which can be viewed as valence, is the most significant mediator for the social effects of emotion. However, it is important not to over-generalize these findings and conclude that valence is all that matters. Effectively, there is consistent evidence in the literature for the differentiated effects of negative displays of emotion on people's decision making (for a review, see [14]). Our own research has shown that sadness, anger, and guilt can have distinct effects on people's decision to cooperate in the prisoner's dilemma and, that these effects are mediated by, not only conduciveness to goals, but other appraisals as well [16]. Finally, though currently lacking, research on displays of positive emotions (joy, pride, admiration, gratitude, etc.) is likely to reveal, as well, differentiated social effects on people's decision making.

Regarding limitations and future work, a methodology based in scenarios can always be subjected to the criticism that results only tap into people's naïve theories of the social effects of emotion and not actual behavior [27]. In this regard, we point out that the results presented here are compatible with other findings where people actually engaged in negotiation with emotional human [12, 13] and virtual [15] counterparts. Still, we plan to replicate and extend this work with experiments, with proper financial incentives [28], where people engage in negotiation with emotional virtual agents. Furthermore, these experiments should pay special attention to the contextual effects of emotion. As discussed above, the interpretation of emotion displays is highly contextual [6, 14, 25]; for instance, a display of guilt after the human's first offer can have a very different meaning than the same display of guilt after the agent has made an unreasonable counter-offer.

Finally, the results presented in this paper have important implications for the design of agents that can negotiate. First, the results report which emotions agents should express to systematically influence humans' concession-making, likelihood of accepting or dropping out and, perceptions of the agent's cooperativeness. Second, the results also emphasize tradeoffs designers need to be aware of when endowing agents with emotions: Whereas an agent that displays anger, guilt or sadness might be effective in eliciting concessions from people, people are less willing to interact with such agents in the future; on the other hand, whereas people might prefer to interact with an agent that displays joy, the agent might not be very successful in reaching profitable agreements for itself, at least in the short run. Finally, the mediation results lend support to reverse appraisal which suggests that what is essential for the social effects of emotion in negotiation are not the emotion displays per se but, the information about appraisals that is communicated. This emphasizes that designers need not necessarily simulate facial displays of emotion to achieve the effects of emotion on negotiation we see in people; all that is required is to effectively convey the underlying appraisals.

Acknowledgments. This work was sponsored by the Fundação para a Ciência e a Tecnologia grant #SFRH-BD-39590-2007; the Air Force Office of Scientific Research under grant FA9550-09-1-0507; and, the NSF under grant IIS-0916858. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Pruitt, D., Carnevale, P.: Negotiation in social conflict. Brooks/Cole Publishing Co., Pacific Grove (1993)
2. Jennings, N., Faratin, P., Lomuscio, A., Parsons, S., Wooldridge, M., Sierra, C.: Automated negotiation: Prospects, methods and challenges. Group Decis. Negot. 10(2), 199–215 (2001)
3. Faratin, P., Sierra, C., Jennings, N.: Using similarity criteria to make issue trade-offs in automated negotiations. Artificial Intelligence 142(2), 205–237 (2002)
4. Lin, R., Kraus, S.: Can automated agents proficiently negotiate with humans? Communications of the ACM 53(1), 78–88 (2010)
5. Van Kleef, G., De Dreu, C., Manstead, A.: An interpersonal approach to emotion in social decision making: The emotions as social information model. Adv. Exp. Soc. Psychol. 42, 45–96 (2010)
6. Hareli, S., Hess, U.: The social signal value of emotions. Cognition & Emotion 26(3), 285–289 (2012)
7. Morris, M., Keltner, D.: How emotions work: An analysis of the social functions of emotional expression in negotiations. Res. Organ. Behav. 22, 1–50 (2000)
8. Blanchette, I., Richards, A.: The influence of affect on higher level cognition: A review of research on interpretation, judgment, decision making and reasoning. Cognition & Emotion 15, 1–35 (2010)
9. Frijda, N., Mesquita, B.: The social roles and functions of emotions. In: Kitayama, S., Markus, H. (eds.) Emotion and Culture: Empirical Studies of Mutual Influence, pp. 51–87. American Psychological Association, Washington, DC (1994)
10. Keltner, D., Haidt, J.: Social functions of emotions at four levels of analysis. Cognition & Emotion 13(5), 505–521 (1999)
11. Keltner, D., Kring, A.M.: Emotion, social function, and psychopathology. Rev. Gen. Psychol. 2, 320–342 (1998)
12. Van Kleef, G., De Dreu, C., Manstead, A.: The interpersonal effects of anger and happiness in negotiations. J. Pers. Soc. Psychol. 86, 57–76 (2004)
13. Van Kleef, G., De Dreu, C., Manstead, A.: Supplication and appeasement in negotiation: The interpersonal effects of disappointment, worry, guilt, and regret. J. Pers. Soc. Psychol. 91, 124–142 (2006)
14. Van Kleef, G., De Dreu, C., Manstead, A.: An interpersonal approach to emotion in social decision making: The emotions as social information model. Advances in Experimental Social Psychology 42(10), 45–96 (2010)
15. de Melo, C., Carnevale, P., Gratch, J.: The effect of expression of anger and happiness in computer agents on negotiations with humans. In: Proceedings of Autonomous Agents and Multiagent Systems, AAMAS (2011)
16. de Melo, C., Carnevale, P., Read, S., Gratch, J.: Reverse appraisal: The importance of appraisals for the effect of emotion displays on people's decision-making in a social dilemma. To be Presented at The 34th Annual Meeting of the Cognitive Science Society, CogSci 2012 (2012)

17. Ellsworth, P., Scherer, K.: Appraisal processes in emotion. In: Davidson, R., Scherer, K., Goldsmith, H. (eds.) Handbook of Affective Sciences, pp. 572–595. Oxford University Press, New York (2003)
18. Hareli, S., Hess, U.: What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. Cognition & Emotion 24, 128–140 (2010)
19. Preacher, K., Hayes, A.: Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav. Res. Meth. 40, 879–891 (2008)
20. de Melo, C., Gratch, J.: Expression of Emotions Using Wrinkles, Blushing, Sweating and Tears. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 188–200. Springer, Heidelberg (2009b)
21. Smith, C., Ellsworth, P.: Patterns of cognitive appraisal in emotion. J. Pers. Soc. Psychol. 48, 813–838 (1985)
22. Scherer, K.: Appraisal considered as a process of multi-level sequential checking. In: Scherer, K., Schorr, A., Johnstone, T. (eds.) Appraisal Processes in Emotion: Theory, Methods, Research, pp. 92–120. Oxford University Press, New York (2001)
23. Frijda, N.: Relations among emotion, appraisal, and emotional action readiness. J. Pers. Soc. Psychol. 57, 212–228 (1989)
24. Baron, R., Kenny, D.: The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. J. Pers. Soc. Psychol. 51, 1173–1182 (1986)
25. Aviezer, H., Hassin, R., Ryan, J., Grady, C., Susskind, J., Anderson, A., et al.: Angry, disgusted, or afraid? Studies on the malleability of emotion perception. Psychological Science 19(7), 724–732 (2008)
26. Scherer, K., Grandjean, D.: Facial expressions allow inference of both emotions and their components. Cognition & Emotion 22(5), 789–801 (2008)
27. Parkinson, B., Manstead, A.S.R.: Making sense of emotion in stories and social life. Cognition & Emotion 7(3/4), 295–323 (1993)
28. Hertwig, R., Ortmann, A.: Experimental practices in economics: A methodological challenge for psychologists? Behav. Brain Sci. 24(3), 383–451 (2001)

First Impressions: Users' Judgments of Virtual Agents' Personality and Interpersonal Attitude in First Encounters

Angelo Cafaro[1], Hannes Högni Vilhjálmsson[1], Timothy Bickmore[4], Dirk Heylen[5], Kamilla Rún Jóhannsdóttir[2], and Gunnar Steinn Valgarðsson[1,3]

[1] Center for Analysis and Design of Intelligent Agents, School of Computer Science
[2] School of Business
Reykjavik University, Iceland
[3] Icelandic Institute for Intelligent Machines, Reykjavik, Iceland
{angelo08,hannes,kamilla,gunnarsv08}@ru.is
[4] College of Computer and Information Science, Northeastern University, USA
bickmore@ccs.neu.edu
[5] Human Media Interaction, University of Twente, The Netherlands
d.k.j.heylen@utwente.nl

Abstract. In first encounters people quickly form impressions of each other's personality and interpersonal attitude. We conducted a study to investigate how this transfers to first encounters between humans and virtual agents. In the study, subjects' avatars approached greeting agents in a virtual museum rendered in both first and third person perspective. Each agent exclusively exhibited nonverbal immediacy cues (smile, gaze and proximity) during the approach. Afterwards subjects judged its personality (extraversion) and interpersonal attitude (hostility/friendliness). We found that within only 12.5 seconds of interaction subjects formed impressions of the agents based on observed behavior. In particular, proximity had impact on judgments of extraversion whereas smile and gaze on friendliness. These results held for the different camera perspectives. Insights on how the interpretations might change according to the user's own personality are also provided.

Keywords: first impressions, personality traits, interpersonal attitude, empirical evaluation, nonverbal behavior, camera perspectives.

1 Introduction

In first encounters our initial impressions of another person may determine whether there are subsequent encounters and, importantly, what expectations we bring to future encounters [1]. Therefore, it is not surprising that individuals attempt to manage the impressions that others form of them [2]. First impressions can be shaped by both *static* individual characteristics and stereotypes, such as height, clothing or, generally, visual appearance [3,4,5] and by *dynamic* characteristics such as verbal [6] and nonverbal behavior [1,4,7]. These characteristics differ in the way they can be controlled. Individuals, for example, can carefully plan how to present themselves visually in a first

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 67–80, 2012.

encounter, but then it may be difficult to have full control over all nonverbal cues [2] during the interaction. In fact, one of the most interesting properties of nonverbal cues in social interaction is that they are irrepressibly impactful. Try as they might, people cannot refrain from behaving nonverbally. If, for example, they try to be as passive as possible, they are likely to be perceived as unexpressive, inhibited, withdrawn, and uptight [2]. Therefore, nonverbal behavior plays a fundamental but, at the same time, subtle role in the dynamics of impression management. After just a few seconds of observing someone's nonverbal behavior, we can pick up with a remarkable accuracy a variety of information including, for instance, a person's skills [8], sexual orientation [9], political view [10] and, personality or attitudes towards other persons [4,11,12,13]. Intelligent agents are not immune to these judgments. During initial encounters with agents, such as the receptionist at a virtual museum in Fig. 1, users form impressions of them based on exhibited nonverbal behavior.

During even the most fleeting interactions, perceivers rapidly form impressions of another's personality traits [14], which can be defined as broad themes in behavior, thoughts, and emotions that distinguish one person from another and endure over time [5]. The most used theory of personality, the so called "Big 5" model [15], is based on the results of factor analyses that demonstrate that five factors are sufficient for providing the best compromise between explanatory power and parsimony. These 5 factors are: extraversion, neuroticism, agreeableness, conscientiousness and openness to experience. Accurate first impressions of personality traits can be formed [11] and extraversion (the extent to which people are outgoing, gregarious, talkative, and sociable) seems to be one of the easiest trait to pick up [14] through rapid interpersonal judgments of nonverbal behavior [13], including interpersonal distance, smile, gaze and posture [4,7,12].

Compared to personality, attitudes are subject to a greater degree of variation over time. Interpersonal attitudes are essentially an individual's conscious or unconscious judgment of how they feel about and relate to another person while interacting with them [4]. Argyle identifies two fundamental dimensions of interpersonal attitudes that can account for a great variety of non-verbal behavior: affiliation (ranging from friendly to hostile) and status (from dominant to submissive) [4]. Affiliation, in particular, can be broadly characterized as liking or wanting a close relationship. Most categories of nonverbal behavior that can be used to regulate this aspect fall under the category of *"immediacy behavior"*. These include proximity, gaze, and certain facial expressions such as smiles. Immediacy is similarly defined as the degree of perceived physical or psychological closeness between two people [12]. Greater affiliation or immediacy, for example, is conveyed by standing close instead of far, having eye contact and smiling in interpersonal encounters [4,12].

These theoretical underpinnings suggest that the specific set of nonverbal cues composed by smile, gaze and proximity can be used to manage impressions of both "long-term" (personality traits) and "more immediate" (interpersonal attitudes) individual characteristics. We will exploit this duality in the context of initial greeting encounters between humans and agents. The main research questions behind our work are the following. (1) What is the role of smile, gaze and proximity when managing impressions of extraversion and affiliation? (2) How do those cues combine in user interpretations? (3) Does the interpretation of nonverbal behavior change according to users' own personality?

2 Related Work

Expression of Personality and Interpersonal Attitudes. There has been considerable previous work developing expressive virtual characters capable of reflecting a personality consistent with the verbal and nonverbal cues exhibited. Neff et al. exploited the extraversion [16] and neuroticism [17] traits of the Big Five model in multimodal characters evaluating the effects of verbal and nonverbal behavior in personality perception studies. Similarly, Paiva et al. [18] presented a model of personality, based on the Big 5, aimed at creating distinct traits that in turn can influence an agent's cognitive and behavioral processes. Pelachaud et al. proposed a real-time backchannel selection algorithm for choosing the type and frequency of backchannels to be displayed according to the personality of the virtual character used [19]. Regarding interpersonal attitudes, Gillies and Ballin [20] concentrated on a general framework based on Argyle's status and affiliation model for animating nonverbal behavior of virtual characters in improvisational visual media production and expressing interpersonal attitudes toward to one another. Finally, Lee and Marsella [21] proposed an analysis framework of nonverbal behavior for modelling side participants and bystanders. They based their analysis on the Argyle's status and affiliation model and considered agents' interpersonal relationships, communicative acts and conversational roles.

These works dealt with either incorporating personality traits [16,17,18,19] or interpersonal attitudes [20,21] separately. The virtual agents were mainly designed for face-to-face interactions or interactive drama. Our work focuses on interpretations of nonverbal behavior when both personality (extraversion) and attitude (affiliation) are expressed at the same time. Furthermore, our agents are exclusively exhibiting nonverbal behavior in the formative moments of the first virtual encounter between the user and agent.

Impression Management and Nonverbal Behavior. Heylen et al. [22,23] showed how a realization of a simple communicative function (managing the interaction) could influence users' impressions of an agent. They focused on impressions of personality (agreeableness), emotion and social attitudes through different turn-taking strategies in human face-to-face conversations applied to their virtual agents in order to create different impressions of them. In [24], Fukayama et al. proposed and evaluated a gaze movement model that enabled an embodied interface agent to convey different impression to users. They used an "eyes-only" agent on a black background and the impressions they focused were affiliation (friendliness, warmth) and status (dominance, assurance). Similarly, Takashima et al. [25] evaluated the effects of different eye blinking rates of virtual agents on the viewers subjective impressions of friendliness, nervousness and intelligence.

The work of Heylen et al. emphasizes the "side-effect" of different nonverbal choices in the realization of a communicative function (i.e. turn taking), whereas our purpose is to intentionally manipulate specific agents' immediacy cues (smile, gaze and proximity) and see how users interpret them. The interest is on the impressions they form of personality/affiliation but also keeping an eye on extra types of judgments that could arise. As opposed to Fukuyama et al., we are using full body virtual agents to exhibit our nonverbal behavior (in particular to be able to exhibit proximity cues), which is not narrowed down to specific behaviors such as eyes-only gaze [24] or eyes blinking [25].

Impact of User's Personality on Agent Evaluation. Bickmore et al. [26] showed that an agent's use of small talk increased trust in it for extraverted users, but for introverted users it had no effect. According to Von Der Pütten et al.[27], users' personality influences their subjective feeling after the interaction with a virtual agent, as well as their evaluation and actual behavior. The effects of an agent's behavior also depends on the personality of the user, in particular people with high values in agreeableness and extraversion (among other findings) judged agents more positively compared to people with high values in shyness. Kang and colleagues suggested that users' personality traits crucially affect their perceptions of virtual agents. They explored how users' shyness [28] and Big 5 personality traits [29] are associated with their feelings of rapport when they interacted with different versions of virtual agents capable of exhibiting nonverbal feedback. In [28] they found that more anxious people (high in social anxiety, i.e. shyness) felt less rapport, while feeling more embarrassment, when they interacted with a *non-contingent* agent. On the other hand, in [29] more agreeable people felt strong rapport when interacting with a rapport agent embodying agreeable features (i.e. nonverbal *contingent* feedback while listening).

As opposed to the typology of studies investigating the benefits of matching-up user and agent personality (e.g. [30]), we aim to understand the role of a user's personality when interacting with a virtual agent, similar to [28,29]. However, in our context we are interested in the possible blending effect that user personality may have on snap judgments of personality/affiliation after observations of solely body language in the very first moments of interaction.

3 Experimental Design

In order to evaluate users' impressions of a greeting agent's extraversion and affiliation in a first encounter we conducted an empirical study in which subjects approached a series of agents with their own avatar. The agents exclusively exhibited a set of nonverbal immediacy cues that were systematically manipulated during approaches of 12.5 seconds each (the length has been chosen after a prior validation study described later in this section). The study was split in two trials differing only in the camera perspective used (1st or 3rd). Our hypotheses, for both trials, were the following:

- **H1:** The amount of extraversion that subjects attribute to a greeting agent **(a)** depends on the unique combination of smile, gaze and proximity it exhibits towards the subject during the first 12.5 seconds of the interaction and **(b)** is further moderated by the subject's own personality.
- **H2:** The amount of friendliness that subjects attribute to a greeting agent **(a)** depends on the unique combination of smile, gaze and proximity it exhibits towards the subject during the first 12.5 seconds of the interaction and **(b)** is further moderated by the subject's own personality.

3.1 Apparatus and Stimuli

The context was a virtual main entrance of a museum. The scene always started with the subject's avatar (AVATAR) outside, in front of automatic sliding doors, and the greeting agent (AGENT) standing inside, close to a reception desk watching a computer screen. Figure 1 (left) shows this setting in first person perspective when the approach has already started. To conduct the study in a fully controlled fashion and have subjects focusing exclusively on the AGENT, their level of interaction was limited to deciding when to start the approach by pressing a specific button. This triggered a locomotion behavior of their AVATAR towards the AGENT that automatically ended when the AVATAR reached the encounter space. We limited the control of the AVATAR to this simple choice to ensure that all approaches were performed in the same way across all conditions and subjects. To control for possible bias of the agent's visual appearance on the impressions formed, the agents were always graphically identical and not wearing any clothes. We used a male gendered model having human resemblance. Body movements were generated with procedural animation techniques and included a default eye blinking behavior and a slight body oscillation movement. All AGENTS were *always* holding the arms at the back with hands unclenched (as shown in Fig. 1 (left)). To give the idea of interaction with different entities we assigned them the name "Agent" followed by a progressing number shown at the beginning of each approach and in the top-left corner of the screen.

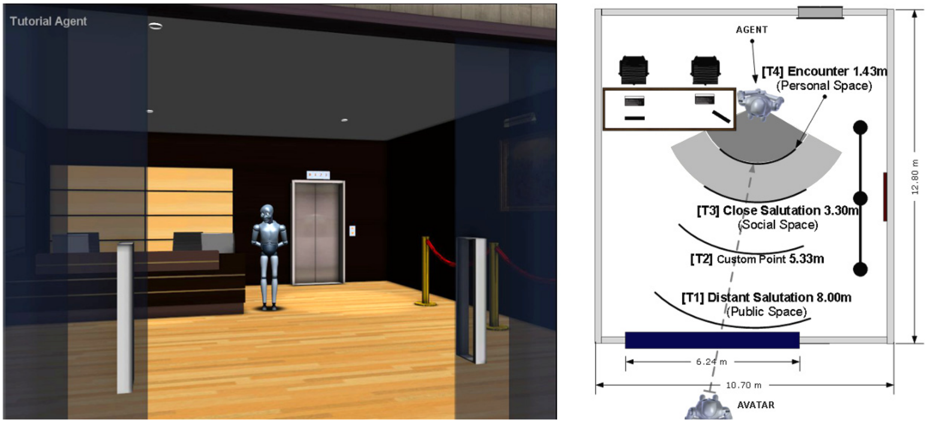


Fig. 1. The setting of our study with the user's avatar entering the virtual museum entrance in first person mode and the greeting agent waiting inside. The schematic shows points where specific behaviors were exhibited by the agent during the avatar's approach.

Our independent variables were **smile** (no vs. yes), **gaze** (low % vs. high %) and **proximity** (no approach vs. approach). We conducted an informal manipulation check (N = 10, 2 females and 8 males, every subject tested both 1P and 3P perspectives) where we deployed a simplified version of the 3D environment and the agent exhibiting each behavior separately to verify that differences between the levels were correctly perceived by subjects within a certain time limit. The exact timing and location for triggering each behavior was based on Kendon's observations of human greetings (distant and close salutation model) [31] and Hall's proxemics theory [32]. Figure 1 (right) shows a schematic top view of the scene with the AVATAR and the AGENT in their initial positions. The grayed dotted line shows the path followed by the AVATAR, black arcs are points were specific behaviors were exhibited. The description on top of them includes: a short reference name (in square brackets), the corresponding stage in Kendon's model (except for the custom point T2), the distance (in meters) from the AGENT and the name of the corresponding space in Hall's model (when overlapping). The arc without description was added to manipulate gaze (as described later) and the gray circular sections represents the AGENT's social and personal space according to Hall's proxemics theory. The duration of 12.5 seconds for each approach came naturally from the two models chosen: It was the time needed by the AVATAR to walk from its initial position (slightly off T1) to the *encounter* point (T4), that coincided with Hall's *personal space* boundary (humans usually do not allow others to cross this space, in particular during a first encounter). The duration was also determined by the AVATAR's speed, that was fine-tuned in the manipulation check to make sure that subjects were able to observe all the nonverbal cues exhibited by the AGENT, while keeping a walking speed for the AVATAR as much natural as possible.

We created a *baseline behavior* for the AGENT that was exhibited across all conditions of the study when the AVATAR approached it. This consisted of watching the computer screen at the beginning with both head and eyes towards it, gazing at the AVATAR for 2 seconds when it was at T1 (8m), looking back at the screen moving only the eyes and, finally, gazing at the AVATAR at T3 (3.30m). The AVATAR always stopped at T4 (1.43m). In a smiling condition the AGENT started smiling at T1. The *"high %"* gaze was obtained with a 2 seconds eye glance at T2. It follows that the difference between "low %" and "high %" gaze conditions was simply related to their duration, in the former the AGENT looked at the subject's AVATAR for a shorter time compared to the latter (in the manipulation check we validated whether subjects were able to distinguish between the two). The *"approach"* condition was simply a step towards the AVATAR when it was at T2 keeping the arms at the back. Since we had eight different conditions, we adopted a latin square design to partially counter balance the treatment order and avoid first order carryover effects [33].

3.2 Measures

A summary of our measures is provided in Tab.1. Agent Extraversion was assessed using 4 items from the Saucier's Mini-Markers [34] set of adjectives for measuring the Big 5. Two with positive (bold and extroverted) and two with negative (shy and withdrawn) valence. For the analysis the negative valence items scores were flipped and averaged with the positive ones to provide a final score. As exploratory variables, we included the ***Extra Impressions*** formed by subjects right after every approach and a measure of ***Agent Likeability***.

Table 1. Summary of measures. Points refer to number of points on Likert scales.

MEASURE	QUESTION	POINTS	LEFT ANCHOR	RIGHT ANCHOR
Agent Extraversion	I think the agent is [bold, extraverted, shy, withdrawn]	9	Extremely inaccurate	Extremely accurate
Agent Friendliness	How hostile/friendly has the agent been towards you?	9	Extremely hostile	Extremely friendly
Agent Likeability	Would you want to continue the interaction with this agent later?	5	No, definitely not	Yes, definitely
Subject Personality	Extraversion, agreeableness, neuroticism (using Saucier's items)			
Extra Impressions	Subjects asked to write adjectives that came to their minds			

3.3 Participants and Procedure

We had 32 participants for each trial recruited via public announcements in our university campus and the surrounding city. In the 1P trial we had 20 males and 12 females representing 11 nationalities[1]. In the 3P trial we had 19 males and 13 females representing 9 nationalities. In both trials, subjects were aged 21-60 with 63% in the 21-30 range. All subjects were well educated and most were at least familiar with computer science and psychology. They were led to a dedicated room at our university facility, seated in front of a 19" LCD monitor, instructed about the procedure and shown a tutorial for familiarization. After this introduction, the investigator monitored the session from an adjacent room. The session consisted of (1) observing each approach and then filling a form that included all measurements except the subject personality, (2) completing the personality inventory and (3) inserting demographic data in separate web forms. Finally, the investigator debriefed them.

3.4 Quantitative Results

We conducted separate statistical analyses for the two trials, further comparison between the two is provided in Sec. 4. For each trial, we conducted a mixed-design ANOVA for each measure (Agent Extraversion and Friendliness) with smile, gaze, and proximity as a within-subjects factors and subject extraversion, agreeableness and neuroticisms as between-subjects factors. We used a full factorial model except that we omitted interactions among the between-subject factors. In order to use the three subject personality traits as between factors, for each measured trait we split our population in tertiles, thus resulting in 3 levels "*low*, *medium* and *high*" for each trait. For quantitative variables this has been shown to be a better practice [35] compared to the median split [36] (*"high"* and *"low"*). Main effects of interactions between factors are tested using Bonferroni adjustments for multiple comparisons. Effect sizes for all comparisons ranged from **.02** to **.73**. Table 2 provides a summary of our quantitative findings for both trials.

[1] As part of the demographic information, we asked participants to select the nation that most represented their cultural identity from a list of all countries in the world.

Table 2. A summary of our results. The first column indicates the camera perspective of the trial, second and third refer to our two measures: agent extraversion and friendliness. For each measure relevant main effects and factor interactions, including significance level (p-values in parenthesis), are reported. All main effects positively affected extraversion and friendliness. The factor interactions had different influence depending on the subject personality. The abbreviation S. stands for *"subject"*.

TRIAL	AGENT EXTRAVERSION	AGENT FRIENDLINESS
1P	Proximity (.000)	Smile (.000)
	Gaze (.082)	Gaze (.049)
	Gaze * S. Extraversion (.052)	Gaze * S. Agreeableness (.026)
	Smile * S. Agreeableness (.084)	Smile * Proximity * S. Agreeable. (.031)
3P	Proximity (.000)	Smile (.000)
	Smile * S. Extraversion (.025)	Gaze (.002)
	Gaze * Proximity * S. Extra. (.057)	Smile * S. Extraversion (.002)
	Smile * Proximity * S. Neuro. (.070)	Smile * S. Agreeableness (.064)

First Person Perspective (1P)

Agent Extraversion. The analysis revealed a significant main effect of proximity on agent level of extraversion, $F(1, 25) = 34.75, p < .001$; *approaching* agents were rated higher than *non-approaching* agents (**H1-a supported**). The main effect of gaze was near significant, $F(1, 25) = 3.28, p = .082$. The main effect of smile was not significant, and there were no significant factor interaction effects. However, the factor interaction between gaze and subject extraversion was near significant, $F(2, 25) = 3.35, p = .052$, as was the factor interaction between smile and subject agreeableness, $F(2, 25) = 2.74, p = .084$, therefore **H1-b** is **rejected**.

Agent Friendliness. There was a significant main effect of smile on agent level of friendliness, $F(1, 25) = 34.75, p < .001$; *smiling* agents were rated higher than *not smiling* ones (**H2-a supported**). There was a significant main effect of gaze, $F(1, 25) = 4.27, p < .05$, and a significant factor interaction between gaze and subject agreeableness, $F(2, 25) = 4.2, p < .05$. This would suggest that the effect of gaze depended on the subject personality. A main effects follow-up analysis revealed that gaze affected the ratings of agent friendliness for *low* agreeable subjects, but not *medium* and *high* ones (**H2-b is partially supported**). The main effects of gaze were further analyzed by pairwise comparisons: for subjects with *low* level of agreeableness, the ratings of agent friendliness in the *low gaze* condition were significantly lower than the *high gaze* condition ones. There was also a significant factor interaction between smile, proximity and subject agreeableness, $F(2, 25) = 4.02, p < .05$. The follow-up analysis of proximity main effects was not significant. On the other hand, smile affected the ratings of agent friendliness at all levels of proximity and for all the three subject personality levels, except for *low* agreeable subjects when the agents were *not approaching* them.

Agent Likeability. We ran the same mixed-design ANOVA for the ratings of agent likeability. There was a significant main effect of smile on agent likeability $F(1, 25) = 20.03, p < .001$; subjects preferred to continue the interaction with *smiling* agents.

Third Person Perspective (3P)

Agent Extraversion. Results of the analysis revealed a significant main effect of proximity on agent level of extraversion, $F(1, 25) = 67.20, p < .001$, and this was rated higher in the *approach* condition (**H1-a supported**). The main effects of smile and gaze were not significant. There was a significant factor interaction between smile and subject extraversion, $F(2, 25) = 4.27, p < .05$. This would suggest that the effect of smile depended on the subject personality. However, a main effects follow-up analysis revealed that smile affected the ratings of agent extraversion for *low* extraverted subjects, but not *medium* and *high* ones (**H1-b is partially supported**. A main effects analysis indicated that for subjects with *low* level of extraversion the ratings of agent extraversion when *not smiling* were significantly different from the condition with *smiling*. The factor interaction between gaze, proximity and subject extraversion was near significant, $F(2, 25) = 3.22, p = .057$, as was the factor interaction between smile, proximity and subject neuroticism, $F(2, 25) = 2.97, p = .070$.

Agent Friendliness. There were significant main effects of smile and gaze on agent level of friendliness (Smile. $F(1, 25) = 49.07, p < .001$; Gaze. $F(1, 25) = 12.33$, $p < .005$); friendliness was rated higher either when the agent was *smiling* or when the *amount* of *gaze* was *high* (**H1-a supported**). The main effect of proximity was not significant. There was a significant factor interaction between smile and subject extraversion, $F(2, 25) = 8.00, p < .005$. This would suggest that the effect of smile depended on the subject personality. However, a main effects follow-up analysis revealed that smile affected the ratings of agent friendliness for *medium* and *high* extraverted subjects, but not *low* ones (**H2-b is partially supported**). The main effects of smile were further analyzed: for subjects with *medium* level of extraversion the ratings of agent friendliness when *not smiling* were significantly lower than conditions with *smiling* agents. For subjects with *high* level of extraversion the ratings of agent friendliness when *not smiling* were significantly different from the conditions with *smiling*. The factor interaction between smile and subject agreeableness was near significant, $F(2, 25) = 3.08, p = .064$, as was the factor interaction between gaze, proximity and subject agreeableness, $F(2, 25) = 2.85, p = .077$.

Agent Likeability. There were significant main effects of smile and gaze on agent likeability (Smile. $F(1, 25) = 41.35, p < .001$; Gaze. $F(1, 25) = 9.91, p < .005$); subjects preferred to continue the interaction with agents *smiling* and *gazing* at them *more*. The factor interaction between smile and subject extraversion was near significant, $F(2, 25) = 2.68, p = .088$, as was the factor interaction between proximity and subject extraversion, $F(2, 25) = 2.73, p = .084$.

3.5 Qualitative Results

For the analysis of "Extra Impressions" we grouped synonymous adjectives into different categories. For each of these, we counted the number of different subjects that used adjectives belonging to that category. In both trials subjects' extra impressions revealed that the agent was judged as "bored, annoyed" (Tot. 1P = 24, 3P = 15) mainly when *not smiling* and *not approaching* or exhibiting a *short* amount of *gaze*, "careless, dismissive, uninterested" (Tot. 1P = 12, 3P = 23) when *smiling* but *gazing* for a *short* amount of time and vice versa. Impressions of "aggressive, stern, challenging and unfriendly" were formed (TOT. 1P = 15, 3P = 18) when the agents were *approaching*. In general, subjects judged the agents as "kind, polite, gentle" (Tot. 1P = 20, 3P = 6) and used common human characteristics to define their extra impressions, thus perceiving the agents as believable even though all our behaviors were pre-scripted. Only a few subjects used adjectives such as "fake, deliberated, agent, scripted" (Tot. 1P = 2, 3P = 6) in the specific condition when he was *approaching*, *not smiling* and *gazing* briefly. Furthermore, adjectives such as "authority, powerful, leader, achiever, ambitioned" were used (Tot. 1P = 17, 3P = 10) mainly when *approaching* and "professional, business-like, precise" (Tot. 1P = 12, 3P = 10) when *not smiling* regardless of proximity and gaze levels.

4 Discussion and Future Work

For the first person perspective (1P), H1-a and H2-a were supported. We found that the amount of extraversion and friendliness that subjects attributed to our agents depended on unique combinations of smile, gaze and proximity that they exhibited. In particular, agents *approaching* the subject's avatar were judged as more extraverted than agents *not approaching*, regardless of gaze amounts or whether they were *smiling* or *not*. Smile had a main effect on judgments of friendliness. These results seem quite intuitive but it is important to note that proximity had absolutely no effects on judgments of friendliness even though qualitative impressions of "aggressive, stern, challenging and unfriendly" were formed when subjects judged *approaching* agents. Therefore, we had a sharp distinction between interpretations of proximity and smile. When it came to judging extraversion proximity had the highest weight, whereas smile dominated the impression formation of friendliness. This is an important result if we consider that smile and gaze can also be used to express personality traits (extraversion) as suggested by previous social psychology literature in human-human nonverbal communication [4,13].

The relation between subject personality and behavior interpretation is harder to explain since H1-b was rejected and H2-b only partially supported. The effect of gaze on agent friendliness partially depended on subject agreeableness. *Low* agreeable subjects interpreted more gazing friendlier compared to less gazing. We didn't get significant results for *medium* and *high* agreeable subjects. According to the personality inventory we used, those who scored *low* in agreeableness are likely to be cold, unsympathetic, rude and harsh as opposed to the warm, kind and cooperative highly agreeable people. We think that this might reflect results of a previous study arguing that low sociable people tend to be more accurate in judging others in zero-acquaintance situations [37]. The factor interaction between gaze and subject extraversion was near significant for the agent level of extraversion, and again only for *low* extraverted subjects (shy, quiet, withdrawn).

Gaze is also involved in a possible explanation for the factor interaction between smile, proximity and subject agreeableness when judging the agent friendliness. In fact, smile had effect on all the subjects except the *low* agreeable group in the particular conditions when the agents were *not approaching*. This would suggest that this group gave more importance to gaze in that case. Although non-significant, a similar trend was observed also in the judgments of agent extraversion.

H1-a and H2-a were also supported when moving to third person perspective and with quite similar results. Again agents *approaching* the subject's avatar were judged as more extraverted than agents *not approaching* them, regardless of smile and the amount of gaze they gave. The effects of gaze on agent friendliness were clearer and didn't depend on subjects' personality. They interpreted agents gazing more at them as friendlier. Smile also led to higher ratings of friendliness, except for *low* extraverted subjects that formed impressions of extraversion rather than friendliness when judging a smiling agent. A possible explanation could be still related to the higher accuracy of judgments that low sociable people express, therefore interpreting smile as a cue of higher extraversion in that case. Another reason could be the great variability we had in the subjects level of extraversion (2.25 to 8.13) whereas the level of agreeableness was more compact (5.00 to 8.25). In general, the role of smile and proximity was clearly separated also for this trial.

Our findings indicate that results in social psychology research on the assessment of personality traits and attitudes on the basis of nonverbal behavior [7,1,4] do translate to the context of user-agent interaction. In particular, outcomes of using nonverbal immediacy [12] are preserved in virtual encounters.

Despite a stronger effect of gaze in 3P, results in both trials are similar, thus suggesting that camera perspective does not alter the way our set of nonverbal cues was interpreted. This result reflects our expectations, even though we couldn't formulate a precise hypothesis a priori due to the lack of previous work investigating this particular aspect. Similar research dealt more with immersive virtual environments [38] explored with head mounted displays [39,40] but not with 3D virtual environments experienced in the same way as in our study or in many of the works mentioned in Sec. 2. We think that, in addition to impact the virtual agents community, this result has also implications in the study of human social psychology. It is interesting to see how users in the 3P trial were still able to form impressions of a virtual character (the agent) when this was exhibiting nonverbal cues towards another virtual character shown on the screen (their avatar) and not directly towards them as in 1P, thus putting them-selves completely in the role of a virtual entity external to their body.

Furthermore, in both trials results of agent likeability mirrored those of friendliness, thus agents smiling and gazing more also resulted in more approachable and likeable agents. This is not surprising considering that one of the advantages of immediacy cues is obtaining a more favourable impression [12], but it also foresees that friendliness was considered more important than extraversion by subjects when they had to decide whether to continue the interaction or not.

Some limitations should be considered. When we looked at the relationship between subjects' personality and their interpretations we found interesting trends supporting that personality acted as moderator. However, these speculations are limited by the statistical significance of the results and the specific population obtained. The ideal body of subjects would have consisted of a balanced population with personality equally distributed in the three groups for each trait. Furthermore, we are aware that cultural identity has influence on behavior interpretation and, in particular, in the 1P we had a high variety in the population. Finally, we may also want to look further into possible gender differences.

Future work will continue in two directions. First, we will build on these results exploiting the impact of impression management on users. The goal is to understand whether initial impressions of an agent impact users' desire to interact with it again. Secondly, we will consider the user personality and investigate possible matches with the agent's personality, interpersonal attitude and the combination of the two.

Acknowledgements. We are grateful to the Reykjavik University IT services for their technical support during the preparation of the study and to Daniel Schulman for his assistance in conducting the data analysis. This work is supported by the Center for Analysis and Design of Intelligent Agents and the School of Computer Science at Reykjavik University.

References

1. Riggio, R.E., Friedman, H.S.: Impression formation: The role of expressive behavior. Journal of Personality and Social Psychology 50(2), 421–427 (1986)
2. DePaulo, B.M.: Nonverbal behavior and self-presentation. Psychological Bulletin 111(2), 203–243 (1992)
3. Naumann, L.P., Vazire, S., Rentfrow, P.J., Gosling, S.D.: Personality judgments based on physical appearance. Personality and Social Psychology Bulletin 35(12), 1661–1671 (2009)
4. Argyle, M.: Bodily communication, 2nd edn. Methuen, New York (1988)
5. Miller, R., Perlman, D., Brehm, S.: Intimate Relationships. McGraw-Hill, Boston (2007)
6. Leary, M.R., Kowalski, R.M.: Impression management: A literature review and two-component model. Psychological Bulletin 107(1), 34–47 (1990)
7. 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)
8. Ambady, N., Rosenthal, R.: Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology 64(3), 431–441 (1993)
9. Rule, N.O., Ambady, N.: Brief exposures: Male sexual orientation is accurately perceived at 50 ms. Journal of Experimental Social Psychology 44(4), 1100–1105 (2008)
10. Rule, N.O., Ambady, N.: Democrats and republicans can be differentiated from their faces. PLoS ONE 5(1), e8733 (2010)
11. Levesque, M.J., Kenny, D.A.: Accuracy of behavioral predictions at zero acquaintance: A social relations analysis. J. of Personality and Social Psychology 65(6), 1178–1187 (1993)
12. Richmond, V., McCroskey, J., Hickson, M.: Nonverbal communication in interpersonal relations, 6th edn. Allyn and Bacon (2008)
13. Campbell, A., Rushton, J.P.: Bodily communication and personality. British Journal of Social & Clinical Psychology 17(1), 31–36 (1978)
14. Kammrath, L.K., Ames, D.R., Scholer, A.A.: Keeping up impressions: Inferential rules for impression change across the big five. J. of Experimental Social Psychology 43(3), 450–457 (2007)
15. McCrae, R.R., Costa Jr., P.T.: Personality trait structure as a human universal. American Psychologist 52(5), 509–516 (1997)
16. 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)

17. 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)
18. Doce, T., Dias, J., Prada, R., Paiva, A.: Creating Individual Agents through Personality Traits. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 257–264. Springer, Heidelberg (2010)
19. 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)
20. Ballin, D., Gillies, M., Crabtree, B.: A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In: First European Conference on Visual Media Production IEEE (2004)
21. 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)
22. ter Maat, M., Heylen, D.: Turn Management or Impression Management? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 467–473. Springer, Heidelberg (2009)
23. ter Maat, M., Truong, K.P., Heylen, D.: How turn-taking strategies influence users' impressions of an agent. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 441–453. Springer, Heidelberg (2010)
24. Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., Hagita, N.: Messages embedded in gaze of interface agents — impression management with agent's gaze. In: Proceedings of SIGCHI, pp. 41–48. ACM, New York (2002)
25. Takashima, K., Omori, Y., Yoshimoto, Y., Itoh, Y., Kitamura, Y., Kishino, F.: Effects of avatar's blinking animation on person impressions. In: Proceedings of Graphics Interface 2008, pp. 169–176. Canadian Information Processing Society, Toronto (2008)
26. Bickmore, T., Cassell, J.: Relational agents: a model and implementation of building user trust. In: Proceedings of SIGCHI, pp. 396–403. ACM, New York (2001)
27. 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)
28. Kang, S.-H., Gratch, J., Wang, N., Watt, J.H.: Does the contingency of agents' nonverbal feedback affect users' social anxiety? In: Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems, pp. 120–127. International Foundation for Autonomous Agents and Multiagent Systems (2008)
29. Kang, S.-H., Gratch, J., Wang, N., Watt, J.H.: Agreeable People Like Agreeable Virtual Humans. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 253–261. Springer, Heidelberg (2008)
30. Isbister, K., Nass, C.: Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. International Journal of Human-Computer Studies 53(2), 251–267 (2000)
31. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters (Studies in Interactional Sociolinguistics). Cambridge University Press (1990)
32. Hall, E.T.: The Hidden Dimension. Doubleday (1966)
33. Bradley, J.V.: Complete counterbalancing of immediate sequential effects in a latin square design. Journal of the American Statistical Association 53(282), 525–528 (1958)
34. Saucier, G.: Mini-markers: A brief version of goldberg's unipolar big-five markers. Journal of Personality Assessment 63(3), 506–516 (1994)

35. Gelman, A., Park, D.: Splitting a predictor at the upper quarter or third and the lower quarter or third. The American Statistician 63(1), 1–8 (2009)
36. MacCallum, R., Zhang, S., Preacher, K., Rucker, D.: On the practice of dichotomization of quantitative variables. Psychological Methods 7(1), 19–40 (2002)
37. Ambady, N., Hallahan, M., Rosenthal, R.: On judging and being judged accurately in zero-acquaintance situations. Journal of Personality and Social Psychology 69(3), 518–529 (1995)
38. Mohler, B.J., Bülthoff, H.H., Thompson, W.B., Creem-Regehr, S.H.: A full-body avatar improves egocentric distance judgments in an immersive virtual environment. In: Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization, p. 194. ACM, New York (2008)
39. Salamin, P., Thalmann, D., Vexo, F.: The benefits of third-person perspective in virtual and augmented reality? In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pp. 27–30. ACM, New York (2006)
40. Mohler, B.J., Bülthoff, H.H., Dodds, T.J.: A communication task in hmd virtual environments: Speaker and listener movement improves communication. In: Proceedings of the 23rd Annual Conference on Computer Animation and Social Agents, pp. 1–4 (June 2010)

A Study of Emotional Contagion with Virtual Characters

Jason Tsai[1], Emma Bowring[2], Stacy Marsella[3], Wendy Wood[1], and Milind Tambe[1]

[1] University of Southern California, Los Angeles, CA 90089
{jasontts,tambe}@usc.edu
[2] University of the Pacific, Stockton, CA 95211
ebowring@pacific.edu
[3] USC Institute for Creative Technologies, Playa Vista, CA 90094
marsella@ict.usc.edu

Abstract. In social psychology, emotional contagion describes the widely observed phenomenon of one person's emotions mimicking surrounding people's emotions [10]. In this paper, we perform a battery of experiments to explore the existence of agent-human emotional contagion. The first study is a between-subjects design, wherein subjects were shown an image of a character's face with either a neutral or happy expression. Findings indicate that even a still image induces a very strong increase in self-reported happiness between Neutral and Happy conditions with all characters tested.

In a second study, we examine the effect of a virtual character's presence in a strategic decision by presenting subjects with a modernized Stag Hunt game. Our experiments show that the contagion effect is substantially dampened and does not cause a consistent impact on behavior. A third study explores the impact of the strategic situation within the Stag Hunt and conducts the same experiment using a description of the same strategic situation with the decision already made. We find that the emotional impact returns, implying that the contagion effect is substantially lessened in the presence of a strategic decision.

1 Introduction

Emotional contagion is defined as the tendency to catch the emotions of other people [10]. While initial work focused on documenting its existence, recent research has moved to understanding its impacts on everyday life. Research in administrative sciences has shown emotional contagion to improve cooperation, decrease conflict, and increase perceived task performance in groups and organizations [2]. Small et al. have shown substantial impacts on charitable donation amounts with only a still image [17]. Though its effects are often felt, in-depth understanding of emotional contagion remains an open area of research.

The vast majority of emotional contagion research, however, has come from the social sciences and examines the spread of emotions from humans to other humans. Emotional contagion's impact in virtual agents' interactions with humans, however, is a largely untouched area of research. The effects are assumed to either be nonexistent and therefore overlooked entirely or to mimic human-human emotional influences. However, these assumptions are not supported by our experiments. As virtual

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 81–88, 2012.

agents enter high-risk and emotionally delicate applications such as virtual psychotherapy [14,15,16], for example, researchers must be cognizant of all potential emotional influences characters can have on users.

This work serves as a first study to find experimental support for the aforementioned results in agent-human emotional contagion. Pursuant of this goal, three sets of studies are conducted. The first study examines the pure contagion case by simply showing subjects a still image of a virtual character with either a happy expression or a neutral expression and then assessing the subject's self-reported happiness thereafter. The use of a still image as a manipulation follows from previous studies in emotional contagion [17,20]. The second study adds the presentation of a game-theoretic situation known as a Stag Hunt along with the character image to assess both the contagion and the behavioral impact of the virtual character in a strategic setting. While studies have shown that emotional contagion can impact one's propensity to trust and enhance perceived cooperation among other findings [2,7], there has been far less work showing behavioral impacts in strategic decisions. Thus, we also attempt to examine whether behavioral impacts arise in strategic decisions to better understand its potential impacts in real-world agent applications. Finally, the third study examines the post-hoc hypothesis that the presentation of a decision to the user dampens the emotional contagion effect. Specifically, we present the same strategic situation as in the second study, but with the decision already made for the subject.

In this work, we provide the first experimental results supporting the existence of emotional contagion between virtual agents and humans. Results show a very large increase in self-reported happiness from only adding a smile to an otherwise identical still image of a virtual character. In the second study, when the character is placed in the context of a strategic decision, both subject behavior and subject self-reports of happiness are only impacted significantly by one character. The last study, which removes the user's decision from the previous experiment, finds that the character's expression's affect on emotion returns significantly, implying that a strategic decision posed to users will dampen the emotional contagion effect beyond only reading about a situation. These results serve as a preliminary study to alert agent researchers to the impacts that virtual character emotions may have on human users.

2 Related Work

Models of emotional contagion have been explored in a computational context that focus on crowd or society simulation. For example, [4,8,13] each present alternative models of emotional contagion in agent crowds, while [18] proposes a comparison technique to evaluate such models. This body of work is an attempt to mimic human-human contagion and not an exploration of agent-human contagion which we seek to understand here. There also exists a large body of work on the interaction between virtual agents and humans [5,9,19]. The entire area of virtual rapport [9,19], for example, focuses on user opinions of the virtual agents and their interaction. The primary goal is to create agents that users enjoy, appreciate, and relate to. Recent work has looked at the impact of agent expressions in a strategic negotiation setting [5] as well. However, their work focuses on the behavioral impact of varying the intent of agent expressions on user behavior without examining the emotional impact on users.

In the social sciences, the literature on emotional contagion is more expansive. Hatfield et al. [10] popularized the area by compiling a plethora of situations in which the phenomenon had been observed in their work as well as the work of other researchers. Follow-up research by the co-authors as well as researchers in related fields such as managerial and occupational sciences [2,17] continued to detail the effects of the phenomenon in new domains. Recently, there have been works beginning to quantify emotional contagion and explore cross-cultural variations in attributes that affect emotional contagion [6,12]. A large body of social psychological studies of emotional contagion features an image or video of only a person's face as the origin of the contagion [11,17,20]. We also expect to see a contagion of emotions from an image of a virtual agent's face to humans. Thus, the primary hypothesis of this work is: *The facial display of an emotion by a virtual character will result in emotional contagion with a human.*

3 Pure Contagion Study

In this study, we test the existence of and factors contributing to emotional contagion between an image of a virtual character's facial expression and a human subject. The experiment setup involved a still image of a character, a self-report of emotion, and a character assessment. Participants were randomly assigned to see one of the images shown in Figure 1, and participants were informed that they would be questioned about the character later. Thus, the study was a 4 (characters) × 2 (expressions) between-subjects design. Ellie is part of the SimCoach[1] project, while Utah is part of the Gunslinger[2] project. Dia was taken from screenshots from Final Fantasy XIII.[3] Finally, Roy was taken from screenshots of the game L.A. Noire.[4] In the self-report of emotion, we asked subjects how strongly they felt each of 8 emotions on a 0-8 Likert scale: angry, joyful, upset, sad, happy, gloomy, irritated, and calm. Only the measure of Happy was used as the other emotions were only included for compliance checking. Specifically, participants that rated both Angry and Joyful higher than 5 and participants that rated Happy and Joyful more than 3 points apart were considered not in compliance.

Finally, a 15-question survey was administered to gauge subjects' perception of the characters shown. Attributes were drawn primarily from the BSRI [3] and included: Aggressive, Affectionate, Friendly, Attractive, Self-Reliant, Warm, Helpful, Understanding, Athletic, Gentle, and Likable. Every question was asked on a 0-8 Likert scale. Compliance tests included duplicating the Attractiveness question and ensuring both occurrences were within 2 points of each other, an Unattractiveness question which could not exceed 5 if Attractiveness exceeded 5, and finally a question that simply asked participants to 'Pick number eight'. Participants were also asked to rate how happy the character seemed.

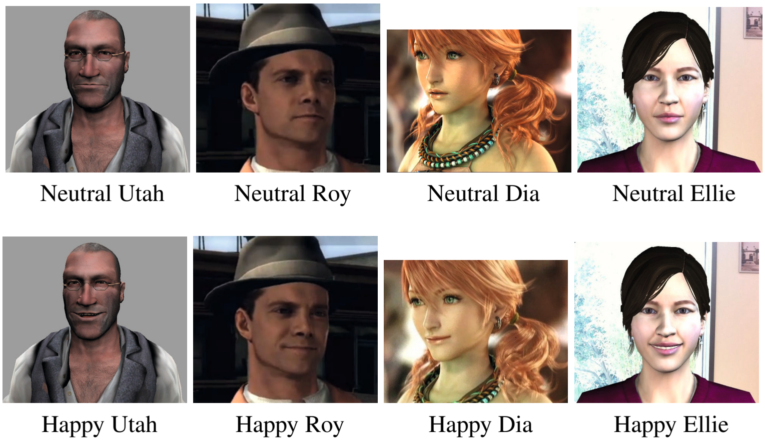
A total of 415 participants that responded to the experiment, conducted on Amazon Mechanical Turk, passed the compliance tests. Participants were required to be over 18

[1] http://ict.usc.edu/projects/simcoach

[2] http://ict.usc.edu/projects/gunslinger/

[3] www.finalfantasyxiii.com

[4] www.rockstargames.com/lanoire/



Neutral Utah Neutral Roy Neutral Dia Neutral Ellie

Happy Utah Happy Roy Happy Dia Happy Ellie

Fig. 1. Characters used, neutral and happy expressions (color)

years of age and were compensated $0.25. The gender distribution was approximately one-third female and two-thirds male, and approximately two-thirds of respondents indicated their ethnicity as Indian.

3.1 Results

We examined whether the facial emotion expressed affected subjects' self-report of emotion. For each of the characters used, participants rated the image used in the Happy condition as significantly happier than the image used in the Neutral condition ($p < 0.001$ for all characters). Based on previous findings in human-human contagion [20], participants should report greater happiness in the Happy condition compared to the Neutral condition. Table 1 shows the means, standard deviations, sample size, and p-values for each experiment. As can be seen, greater happiness was reported in the Happy condition for every character and one-way ANOVA tests revealed significance in every case. This supports our primary hypothesis that an image of a virtual character will cause emotional contagion with a human viewer, since the display of happiness

Table 1. Happiness statistics for Pure Contagion Study

	Condition	Mean	SD	n	p
Utah	Neutral	3.96	2.54	57	< 0.001
	Happy	5.60	2.12	52	
Roy	Neutral	4.00	2.45	45	< 0.001
	Happy	5.75	1.86	55	
Dia	Neutral	4.04	2.26	46	< 0.001
	Happy	5.96	2.19	47	
Ellie	Neutral	4.49	2.37	66	< 0.001
	Happy	5.27	2.10	47	

resulted in reports of higher happiness in subjects as compared to the neutral display. Analysis was also conducted to examine a number of additional hypotheses that have been observed in human-human contagion, but none yielded consistent, statistically significant results. These included differences in contagion strength depending on subject gender, ethnicity, perceived character happiness, and perceived character attractiveness.

4 Strategic Decision Study

Having found preliminary experimental support for the existence of agent-human emotional contagion, we extend the research to include a strategic interaction. Studies into the effects of emotional contagion have primarily been in mimicry, self-reports of emotion, and other non-decision-based effects such as changes in trust inventory responses and judge ratings of 'cooperativeness' [2,7]. While there has been some work in behavioral changes due to emotional contagion, such as its impact on donation amounts [17], our work is the first to consider impacts in a strategic context. The experimental setup involved a still image of a character along with the presentation of a strategic decision, followed finally by a self-report of emotion.

We used a cooperation situation based on the standard game-theoretic Stag Hunt situation. The actual story used in this experiment casts the Stag Hunt scenario in a less outlandish context in which the subject and a coworker he/she has never met are tasked with decorating specific rooms in the office and can either choose to work separately (taking more time) or work together through both of their assigned rooms (taking less time). The amount of time it would take to perform the decoration task was not explicitly stated. The coworker in question was the character whose image is presented with the situation. Subjects were asked how likely they were to help the character with the task on a 0-8 Likert scale. A total of 572 participants responded to the experiment, which was again conducted via Amazon Mechanical Turk, passed the compliance tests.

4.1 Results

In light of the very strong effect found in the Pure Contagion Study and research indicating that the emotional contagion of happiness leads to more trust [7], we expect to see increased happiness in Happy conditions lead to increased likelihood of cooperation. Indeed, we do find a tight link between likelihood of cooperation and participant happiness as shown in Figure 2. The x-axis plots the happiness rating, and the y-axis indicates the average likelihood of cooperation for all respondents with the given happiness rating across all conditions. As the regression's very high R-squared of 0.852 indicates, the two measures are very tightly linked.

However, only the experiment with Dia yielded a statistically significant change in responses. This suggests that the change results from a character-specific attribute and not simply an expression-based mechanism. The lack of effect for the other characters is due partially to the regression's low coefficient of 0.147, which implies that huge changes in happiness are required to induce changes in the likelihood of cooperation. However, the Pure Contagion Study *did* find very large changes in happiness that should have been sufficient. A closer look at the emotional influence of our manipulation reveals the second half of the story.

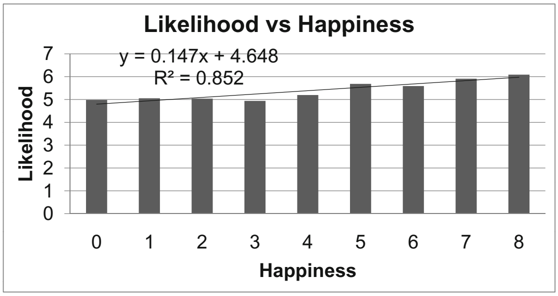


Fig. 2. Likelihood of cooperation versus happiness

While the Pure Contagion Study reported astoundingly large effects of a smile in a still image of a virtual character, the addition of a strategic decision may have altered the contagion effect. Thus, we examine them in this experiment again. We summarize the overall results for each character in Table 2a. As before, we expect subjects in the Happy condition to report higher happiness than subjects in the Neutral condition across all characters. This was indeed the case, however, the effect sizes are much smaller than in the Pure Contagion Study and, in fact, statistical significance was found only in the experiment using Dia, indicating that something character-specific is allowing her to retain more of her emotional impact while all other characters experienced a much greater dampening of emotional impact. In exploring the attributes surveyed in this work, no candidate for a consistent explanatory variable was found.

Table 2. Self-reported happiness of participants

(a) Strategic Decision Study

	Condition	Mean	SD	*n*	*p*
Utah	Neutral	4.92	2.56	105	0.7638
	Happy	5.02	2.48	125	
Roy	Neutral	4.53	2.38	36	0.2098
	Happy	4.86	2.76	49	
Dia	Neutral	4.37	2.57	41	0.019
	Happy	5.68	2.30	38	
Ellie	Neutral	5.24	2.59	93	0.2231
	Happy	5.69	2.39	85	

(b) Strategic Situation Study

	Condition	Mean	SD	*n*	*p*
Utah	Neutral	4.04	2.67	27	0.1329
	Happy	5.09	2.63	32	
Roy	Neutral	4.83	2.33	24	0.2247
	Happy	5.66	2.53	29	
Dia	Neutral	5.88	2.11	48	0.3485
	Happy	6.28	2.08	46	
Ellie	Neutral	4.76	2.33	46	0.008
	Happy	5.95	1.77	41	

These results suggest that the presentation of a trust-based strategic decision dampens the emotional contagion effect. This is in line with findings by researchers in social psychology [17,21] that found that deliberative thinking can dampen emotional influences. However, in light of the tight correlation between the decision and reported happiness, we hypothesize that the decision itself contributed to the dampening effect beyond the impact of simply reading about the situation.

5 Strategic Situation Study

This study was pursued to disentangle the novel effect of making a strategic decision from the previously found effect of reading a situation description [17,21]. It presents subjects with the same situation as in the Strategic Decision Study but removes the decision element from it and simply states that the subject will be cooperating with the character shown to complete the office decoration task. In Table 2b, the overall results of the experiment are shown. As would be expected following findings in social psychology that even reading additional material can dampen emotional influence [17,21], the effect observed in the Pure Contagion Study has not returned in full force. However, the average happiness reported by participants shows a much larger differential than in the Strategic Decision Study, supporting the hypothesis that the decision itself contributed substantially to the dampening of emotional contagion.

6 Conclusion

In this work, we provide a preliminary examination of agent-human emotional contagion across a wide variety of character types. We find support for its existence with a pure contagion study with strong results. In a second study, a strategic decision is added that greatly dampens the contagion effect and, with one exception, did not impact behavior. The final study, which removes the user's decision from the previous experiment, finds that the emotional contagion effect returns. This supports the hypothesis that a strategic decision posed to users will dampen the emotional contagion effect beyond the dampening effect of reading the situation itself.

Our findings suggest a number of insights for virtual agent researchers. First, emotional contagion with virtual agents appears to be substantial and applications should accurately account for it. We have shown that in some domains even a still image can have an emotional effect. Second, researchers should be wary about assuming that human-human social psychology will directly translate into agent-human interactions. Finally, our work has looked at smiles that are perceived as happy, but there are different types of smiles [1]. Further investigations should be carried out to understand the different effects of character expressions. As virtual agent applications extend beyond entertainment into emotionally-charged domains with very serious repercussions such as psychotherapy and military training, researchers must be ever-vigilant of the emotional impacts their characters might have on users.

References

1. Ambadar, Z., Cohn, J.F., Reed, L.I.: All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of Nonverbal Behavior 33(1), 17–34 (2009)
2. Barsade, S.G.: The Ripple Effect: Emotional Contagion and Its Influence on Group Behavior. Administrative Science Quarterly 47, 644–675 (2002)
3. Bem, S.L.: The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology 42, 155–162 (1974)

4. Bosse, T., Duell, R., Memon, Z.A., Treur, J., van der Wal, C.N.: A Multi-agent Model for Emotion Contagion Spirals Integrated within a Supporting Ambient Agent Model. In: Yang, J.-J., Yokoo, M., Ito, T., Jin, Z., Scerri, P. (eds.) PRIMA 2009. LNCS, vol. 5925, pp. 48–67. Springer, Heidelberg (2009)
5. de Melo, C., Carnevale, P., Gratch, J.: The Effect of Expression of Anger and Happiness in Computer Agents on Negotiations with Humans. In: AAMAS 2011 (2011)
6. Doherty, W.: The Emotional Contagion Scale: A Measure of Individual Differences. Journal of Nonverbal Behavior 21(2) (1997)
7. Dunn, J.R., Schweitzer, M.E.: Feeling and Believing: The Influence of Emotion on Trust. Psychology of Personality and Social Psychology 88(5), 736–748 (2005)
8. Durupinar, F.: From Audiences to Mobs: Crowd Simulation with Psychological Factors. PhD thesis, Bilkent University (2010)
9. 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)
10. Hatfield, E., Cacioppo, J.T., Rapson, R.L.: Emotional Contagion. Cambridge University Press (1994)
11. Hess, U., Philippot, P., Blairy, S.: Facial Reactions to Emotional Facial Expressions: Affect or Cognition? Cognition and Emotion 12(4), 509–531 (1998)
12. Lundqvist, L.-O.: Factor Structure of the Greek Version of the Emotional Contagion Scale and its Measurement Invariance Across Gender and Cultural Groups. Journal of Individual Differences 29(3), 121–129 (2008)
13. Pereira, G., Dimas, J., Prada, R., Santos, P.A., Paiva, A.: A Generic Emotional Contagion Computational Model. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 256–266. Springer, Heidelberg (2011)
14. Riva, G.: Virtual Reality in Psychotherapy: Review. CyberPsychology & Behavior 8(3), 220–230 (2005)
15. Rizzo, A., Pair, J., McNerney, P.J., Eastlund, E., Manson, B., Gratch, J., Hill, R.W., Swartout, W.: Development of a VR Therapy Application for Iraq War Military Personnel with PTSD. In: 13th Annual Medicine Meets Virtual Reality Conference, Long Beach, CA. Medicine Meets Virtual Reality, vol. 111, pp. 407–413. IOS Press (2005)
16. Rothbaum, B.O., Hodges, L.F., Ready, D., Graap, K., Alarcon, R.D.: Virtual reality exposure therapy for vietnam veterans with posttraumatic stress disorder. Journal of Clinical Psychiatry 63(8), 617–622 (2001)
17. Small, D.A., Verrochi, N.M.: The Face of Need: Facial Emotion Expression on Charity Advertisements. Journal of Marketing Research 46(6), 777–787 (2009)
18. Tsai, J., Bowring, E., Marsella, S., Tambe, M.: Empirical Evaluation of Computational Emotional Contagion Models. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 384–397. Springer, Heidelberg (2011)
19. Wang, N., Gratch, J.: Rapport and Facial Expression. In: ACII 2009 (2009)
20. Wild, B., Erb, M., Bartels, M.: Are emotions contagious? evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. Psychiatry Research 102(2), 109–124 (2001)
21. Wilson, T.D., Lindsey, S., Schooler, T.Y.: A Model of Dual Attitudes. Psychological Review 107(1), 101–126 (2000)

Longitudinal Affective Computing
Virtual Agents That Respond to User Mood

Lazlo Ring, Timothy Bickmore, and Daniel Schulman

College of Computer and Information Science, Northeastern University
360 Huntington Ave - WVH 202, Boston MA 02115
{lring,bickmore,schulman}@ccs.neu.edu

Abstract. We present two empirical studies which examine user mood in long-term interaction with virtual conversational agents. The first study finds evidence for mood as a longitudinal construct independent of momentary affect and demonstrates that mood can be reliably identified by human judges observing user-agent interactions. The second study demonstrates that mood is an important consideration for virtual agents designed to persuade users, by showing that favors are more persuasive than direct requests when users are in negative moods, while direct requests are more persuasive for users in positive moods.

Keywords: Affect, Agents, Mood, Longitudinal Study, Persuasive Technology.

1 Introduction

As virtual agents spend more time with users—as educators, counselors, and companions—they will need to adapt to changes in users' beliefs, attitudes, and feelings over relatively long periods of time. This is particulary true for agents that enlist the user in behavior change through persuasion. To date, most virtual agent systems treat psychological construct as static, immobile trait confined to a single conversation. Virtual agents that have been developed to sense and respond to user affect also suffer a limitation, they only track and respond to user affective states that change within seconds or minutes [1][2]. In systems that respond to changes in affective state, most agents acknowledge the change and (if warranted) empathize with the user [3]. Expanding upon the idea of responding to changes in affective state, we look to explore other useful adaptations for agent behavior that could be made in response to user mood.

We are interested in the interaction between agent persuasion and the longitudinal affective construct, mood. For the purpose of this study, we define mood (following Larsen [4]) as an affective state that differs from emotion—the external expression of affect[5]—in two distinctive ways: duration and intensity. Whereas emotions last only a matter of seconds (from initial perception and reaction to decay), moods can last hours or days. Moods are also usually perceived to be less intense than emotions and are generally less-specific [6]. This lack of specificity, however, has led to difficulty in modeling mood, unlike emotion which

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 89–96, 2012.

has a variety of well-established models (e.g., basic emotions [7], categories of cognitive elicitors [8], etc.). Because of this, most researchers categorize mood using Russell's & Posner's circumplex model of affect [9], which models affective state (including mood) in terms of valence and arousal.

In this paper we describe a series of empirical studies exploring the detection and use of mood in long-term interactions with virtual agents. In section 3, we investigate whether the assessment of mood in user-agent interactions through verbal and nonverbal behavior is feasible. To answer this, we see if expert human judges can reliably assess the mood of users interacting with a virtual agent. We then investigate whether there is evidence of a mood construct in user-agent interactions, independent of momentary affect that can be assessed using the circumplex model. Then in section 4, we investigate the relationship between user mood and persuasion by an agent. In this study, we explore whether persuasive messages used by an agent are more effective if they are tailored in response to user mood.

2 Related Work

There have been many attempts to detect emotion and affect during agent based interactions. D'Mello investigated the automated detection of affect through various sensors [1]. Three studies were conducted to collect data on how affect expresses itself through body postures and eye movements, and how an automated affect detector could be created. Participants interacted with the AutoTutor pedagogical agent, and then were judged by one of three techniques: experts trained in Ekman's facial action coding system [7], self-report of their affective state, or judgement by their peers. Their results found that posture features and the tracking of eye movement could predict a participant's affective state with 70% accuracy.

The relationship between mood and persuasion has also been explored by multiple researchers. Aderman [10] found that the form of a request significantly impacted a participant's willingness to comply based upon their mood. Following a positive or negative mood induction, participants were asked to sort cards, with the request being phrased as either a study requirement or a favor to the experimenter. Participants in the negative mood condition were found to sort significantly more cards when the task was phrased as a requirement, where as those in the positive condition were found to sort significantly more cards in the favor condition.

3 User Mood Classification by Human Judges

To begin our exploration of user mood in long-term interactions with virtual agents we first wanted to determine whether human observers could reliably identify user moods in these interactions, based on their verbal and nonverbal conversational behavior.

Reliability Analysis: We used videotaped recordings of longitudinal user-agent conversations collected as part of a study of an eldercare companion agent[11] . Fifteen conversations conducted by three participants were selected for reliability analysis. Two minute video segments were extracted from the beginning, middle, and end of each conversation, resulting in a total of 41 video clips for analysis (4 conversations were too short to use all three time points). Three research assistants were asked to view each of the 41 video clips and rate each for arousal and valence using the Affect Grid[12], a self-report instrument that assess arousal and valence on a 2-dimensional grid, where arousal ranges from unpleasant (1) to pleasant (9) and valence ranges from sleepiness (1) to high arousal (9). Judges were also asked to specify a single English word that best described user mood. Video clips were provided for judges to view in any order, and as frequently as they liked.

Results: Arousal scores assigned by judges ranged from 3 to 9 (mean 6.57, SD 1.15) and valence scores ranged from 4 to 9 (mean 6.59, SD 1.17). Judges used 26 English words to describe the moods they observed. The most commonly used words were: "happy" (42 instances), "content" (16), "good" (12), "neutral" (9), and "calm" (8). Ratings of arousal and valence were significantly correlated among the three judges, with intraclass correlation coefficients of 0.662 for arousal ($p < .001$) and 0.646 for valence ($p < .001$). Of the 41 video clips, judges only agreed on English mood labels 12 times: 11 of these were pairs of judges, and only once did all 3 judges volunteer the same label (in all of these cases the label was "happy").

User Mood Variability Analysis: We next sought to characterize the amount of variance in user affect, both within and between conversations in order to determine the amount of variance that was due to momentary affect (within a conversation), the amount due to mood (between convresation), and the eamount due to personality (between subjects). 145 clips from 42 videotaped interactions (described above) was used in this study, with two minute video segments extracted from the beginning, middle, and end of each conversation as before.

For each of valence and arousal, we performed a restricted maximum likelihood fit (using lme4 [13] in R [14]) of a 3-level variance components model:

$$\begin{aligned} &y_{ijk} = \beta_0 + P_i + M_{ij} + \epsilon_{ijk} \\ &P_i \sim N(0, \tau_P^2), \quad M_{ij} \sim N(0, \tau_M^2), \quad \epsilon_{ijk} \sim N(0, \sigma^2) \end{aligned}$$

where y_{ijk} is the average of the judges' ratings for participant i, conversation j, and videotape segment k. We tested for significant intraclass correlation at the level of conversations with a restricted likelihood ratio test [15] that compared this model against a 2-level model which omitted the M_{ij} term.

Results: In the full corpus, valence was observed to range from 2 to 9 (overall mean 5.82, SD 1.32) and arousal ranged from 3 to 9 (mean 5.50, SD 1.40). The estimated variance of valence and arousal within and between conversations is shown in Table 1. For both arousal and valence, most variance was accounted

Table 1. Variance of valence and arousal between and within conversations

	Valence		Arousal	
	Variance	% Total	Variance	% Total
Participants(trait) τ_P^2	0.55	0.50	0.82	0.70
Conversations(mood) τ_M^2	0.16	0.15	0.11	0.09
Segments(affect) σ^2	0.39	0.35	0.36	0.28

for at the level of participants, followed by segments and conversations. There was significant intraclass correlation found at the level of conversations, both for valence ($RLR = 9.44$, $p = 0.001$) and for arousal ($RLR = 5.89$, $p = 0.007$).

Discussion: We demonstrates that user affect can be reliably assessed by human judges using arousal and valence scores, on the basis of observed verbal and non-verbal conversational behavior during interactions with a virtual agent. The use of English words was not a reliable measure of affect as there was essentially no agreement among the judges on terms used. We also show that there is significant intraclass correlation in ratings at the level of conversations, while controlling for overall intraclass correlation. This demonstrates that these assessments partially captured mood: a phenomenon occurring on a larger time scale than a single conversation, yet distinct from an individual's overall baseline affective state. Thus, a longitudinal model of affective state should include both inter-subject (a subject specific baseline) and inter-conversation (mood) components.

4 The Effect of Form of Request and Mood on Persuasion

Following Aderman's work (described in section 2), we decided to adapt his methodology to investigate the effects of mood and persuasive request phrasing on exercise motivation. This specific area was chosen due to previous literature showing that agents are effective exercise counselors, and that they elicit similar effects from dialogue phrasing as found in human-human interactions [16][17].

The study was conducted in the context of the "Virtual Laboratory" system [18], in which a standing group of participants interact with a virtual exercise promotion agent up to once a day from their home computers. The agent encourages participants to walk every day and tracks their progress through a supplied pedometer that the agent discusses with them.

Our manipulation consists of the agent asking participants to exercise, phrased as either a favor to the agent or direct request. Our hypothesis was that participants will walk significantly more steps when they are in a negative mood and are told to walk using a favor dialogue, and when they are in a positive mood and are told to walk using a request dialogue.

Measures: The Affect Grid (Section 3) was used by participants to rate their mood. Finally participants upload the amount of steps they walked since their last session via a pedometer provided at the beginning of each session with the system.

Experimental Protocol: This study was divided into two separate interaction phases: a desensitization phase, and a collection phase. In the desensitization phase (5 days), participants did not interact with the agent, but instead were given an Affect Grid each session for five sessions. This was done to both reduce habituation effects from prior interactions with the agent, and to collect baseline valence and arousal measurements for each participant. This data was used to calculate the change in valence and arousal each day in the following phase.

In the collection phase (2 months), participants first filled out the Affect Grid at the beginning of each session, then conducted their usual counseling conversation with the agent but with the following change: instead of negotiating daily pedometer step count goals the agent asks participants to walk as either a favor or as a request. The exact language used was:

Favor: *I was wondering if you'd mind doing me a favor and take a walk before our next session.*

Request: *Would you take a walk before our next session.*
The manipulation was randomly selected every day for every participant (within-subjects).

Results: Twenty-one participants (mean age 61.5) interacted with the system over two months, resulting in 696 unique interactions (mean=33.1 per participant, SD = 16.2) with the agent, with one participant dropping out of the study. For each interaction, the number of steps the participant had walked since their last session along with their valence and arousal were recorded.

A linear mixed-effects regression model was used to fit the data. This model is an extension of linear regression models that allows for the linear predictors to contain both random and fixed effects. This model used the study condition of favor (Coded as 0) versus request (Coded as 1), the number of interactions, and the difference in participant's valence and arousal from their baseline to estimate the number of steps they walked since their last interaction. Baseline arousal and valence was estimated for each participant using their average valence and arousal recorded via the Affect Grid during the desensitization phase of the study. The average of these scores were used to model the participant specific baseline affect found in study 1. Steps were put on a logarithmic scale to restrict the range of outcomes to greater than 0 steps, and to account for the right tail skew of the measure. Since exact p values and confidence intervals cannot be calculated for mixed effect models analytically, a semi-parametric bootstrap was used, as described by Carpenter, et al [19]. All statistics were calculated using R-2.14.1 and the lme4 package [14][13].

As shown in Table 2, if the agent used the request dialogue while the participant was in a positive mood they walked significantly more steps, and if they agent used the favor dialogue while the participant was in a negative mood they walked significantly more steps ($p < .01$). Additionally, it was found that participants walked significantly more steps when their valence and arousal scores were opposite in sign ($p < .01$). Thus, when a participant is in a high arousal, low valence state, a favor message predicts more walking, whereas when a participant is in a low arousal, high valence state, request predicts more walking (Figure 1).

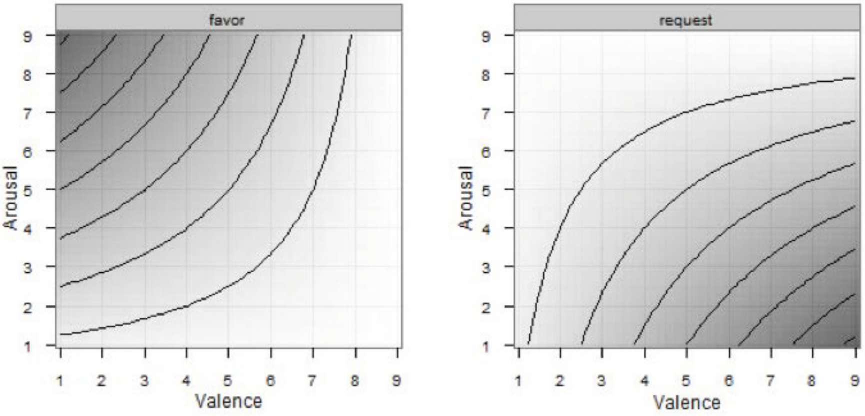


Fig. 1. Change in the number of steps walked based on mood and dialogue manipulation. Darker areas represent where each dialogue had the most positive effect on step count.

However, the effect of the manipulation decreased over time, as shown by the quadratic session terms in Table 2, such that it was no longer significant after a month. This habituation effect is consistent with previous research on affect [20][21], showing the decay of the manipulation through the course of the study.

Discussion: We found that the form of a persuasive message should be tailored based on user mood in order to be maximally effective. These results are contrary to our hypotheses and findings in the previous literature, but our experiment differs from the earlier work in three key aspects. In Aderman's original work, participants were asked to do a favor or request for the experimenter, whereas

Table 2. A Linear Mixed-Effect Regression Model Predicting Participant's Step Count (log-transformed). Inter-subject Variance: (Estimate: .258, 95% CI [.142, .343]), Residual Variance: (Estimate: .623, 95% CI [.566, .678]). Legend: V = Valence, A = Arousal, C = Condition (Favor coded as 0, Request coded as 1), S = Session, S^2 = Sessions×Sessions (To model habituation of affect over time)

Parameter	Est.	SE	P	Parameter	Est.	SE	P
Intercept	6.19e-03	1.96e-03	0.93	$V \times S^2$	2.62e-06	-2.99e-06	1.00
V	2.06e-01	1.79e-03	0.39	$A \times S^2$	-1.20e-05	6.54e-06	0.99
C	-2.08e-02	-1.15e-03	0.87	$V \times A \times C$	4.59e-01	1.55e-02	0.52
S	5.50e-03	-1.95e-04	0.57	$\boldsymbol{V \times A \times S}$	**1.02e-01**	**4.45e-04**	**0.03**
S^2	-1.04e-04	3.25e-06	0.59	$\boldsymbol{V \times C \times S}$	**-8.53e-02**	**-2.42e-06**	**0.03**
$\boldsymbol{V \times A}$	**-1.45**	**-4.22e-03**	**0.01**	$A \times C \times S$	2.73e-02	-1.17e-04	0.42
$\boldsymbol{V \times C}$	**1.07**	**-1.977e-03**	**0.01**	$V \times A \times S^2$	-1.65e-03	-1.13e-05	0.11
$A \times C$	-1.22e-01	2.60e-03	0.74	$V \times C \times S^2$	1.41e-03	4.13e-06	0.09
$V \times S$	-8.66e-03	1.29e-05	0.72	$A \times C \times S^2$	-8.29e-04	-7.96e-08	0.25
$A \times S$	-1.06e-03	-2.10e-04	0.96	$V \times A \times C \times S$	-2.44e-02	-1.23e-03	0.73
$C \times S$	1.27e-03	1.30e-04	0.91	$V \times A \times C \times S^2$	-3.98e-04	2.47e-05	0.81

in our experiment the participant is doing a favor for the agent. However, due to the virtual nature of the agent, the agent cannot benefit from this request; therefore the participants are indirectly doing a favor for themselves. This change in perspective could account for the reversal of the observed trend since the persuasive outcome of interest is self-efficacy instead of altruistic behavior. Additionally, the majority of studies on mood observed only a single session of affect while disregarding the longitudinal property of mood in the process.

5 Conclusion

We found that inter-conversation mood is a significant component of user affect, and that mood can be reliably assessed on the basis of user verbal and nonverbal behavior during interactions with a virtual agent. We also found that mood should be taken into account when selecting persuasive messages in order to maximize compliance, although the effectiveness of a simple (non-varying) mood-based manipulations decays over time. In future studies, we plan to explore the use of non-invasive sensors to automatically detect user mood and investigate the application of the results found in this paper to develop affectively tailored dialogue systems. We also plan to explore how well our findings on mood hold up during much longer time periods, such as year-long interactions with an agent.

Acknowledgments. We thank Connor Westfall, Colin Shaughnessy, Bridgette Collado, and Ashley Kline for their assistance with this study. This material is based upon work supported by the National Science Foundation under Grant No. 1012086 and ISS-0545932. 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.

References

1. D'Mello, S., Graesser, A., Picard, R.: Toward an affect-sensitive autotutor. IEEE Intelligent Systems 22(4), 53–61 (2007)
2. Burleson, W., Picard, R., Perlin, K., Lippincott, J.: A platform for affective agent research. In: 3rd International Conference on Autonomous Agents and Multi Agent Systems. ACM Press (2004)
3. Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration. In: CHI 1999, Extended Abstracts on Human Factors in Computing Systems, CHI EA 1999, pp. 242–243. ACM, New York (1999)
4. Larsen, R.J.: Toward a science of mood regulation. Psychological Inquiry 11(3), 129–141 (2000)
5. Batson, C.D., Shaw, L.L., Oleson, K.C.: Emotion. In: Differentiating Affect, Mood, and Emotion: Toward Functionally Based Conceptual Distinctions, pp. 294–326. Sage Publications, Inc., Thousand Oaks (1992)
6. Thayer, R.E.: The biopsychology of mood and arousal. Oxford University Press, New York (1989)

7. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Stanford University (1977)
8. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Cambridge University Press, Cambridge (1988)
9. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology 17(3), 715–734 (2005)
10. Aderman, D.: Elation, depression, and helping behavior. Journal of Personality and Social Psychology 24(1), 91–101 (1972)
11. Vardoulakis, L.P., Ring, L., Barry, B., Sidner, C., Bickmore, T.: Designing Relational Agents as Long Term Social Companions for Older Adults. In: Nakano, Y., et al. (eds.) IVA 2012. LNCS (LNAI), vol. 7502, pp. 289–302. Springer, Heidelberg (2012)
12. Russell, J.A., Weiss, A., Mendelsohn, G.A.: Affect grid: A single-item scale of pleasure and arousal. Journal of Personality and Social Psychology 57(3), 493–502 (1989)
13. Bates, D., Maechler, M., Bolker, B.: lme4: Linear mixed-effects models using s4 classes (2011)
14. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008) ISBN 3-900051-07-0
15. Scheipl, F., Greven, S., Küchenhoff, H.: Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Computational Statistics & Data Analysis 52(7), 3283–3299 (2008)
16. de Melo, C.M., Carnevale, P., Gratch, J.: The effect of expression of anger and happiness in computer agents on negotiations with humans. In: The 10th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011, pp. 937–944. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2011)
17. Schulman, D., Bickmore, T.: Persuading users through counseling dialogue with a conversational agent. In: Proceedings of the 4th International Conference on Persuasive Technology, Persuasive 2009, pp. 25:1–25:8. ACM, New York (2009)
18. Bickmore, T., Schulman, D.: A virtual laboratory for studying long-term relationships between humans and virtual agents. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009, vol. 1, pp. 297–304. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2009)
19. Carpenter, J.R., Goldstein, H., Rasbash, J.: A novel bootstrap procedure for assessing the relationship between class size and achievement 52(4), 431–443 (2003)
20. Titchener, E.B.: Lectures on the elementary psychology of feeling and attention. Arno Press, New York (1973)
21. Dijksterhuis, A., Smith, P.K.: Affective habituation: Subliminal exposure to extreme stimuli decreases their extremity. Emotion 2(3), 203–214 (2002)

Generating Norm-Related Emotions in Virtual Agents

Nuno Ferreira[1], Samuel Mascarenhas[1], Ana Paiva[1], Frank Dignum[2],
John Mc Breen[3], Nick Degens[3], and Gert Jan Hofstede[3]

[1] INESC-ID and Instituto Superior Técnico, Technical University of Lisbon,
Av. Prof. Cavaco Silva, Taguspark 2744-016, Porto Salvo, Portugal
[2] Utrecht University 3508TB, Utrecht, The Netherlands
[3] Wageningen University Hollandseweg 1, 6706 KN Wageningen, The Netherlands

Abstract. The increased believability provided by emotions in virtual characters is a valuable feature in a multi agent environment. Despite much research on how to model emotions based on events that affect a character's goals, the current emotional models usually do not take into account other sources of emotions, such as norms and standards. Moreover, current normative systems usually do not consider the role of emotions. Systems that include emotions and norms are too domain-specific or lack flexibility. We propose a model for the generation of emotions based on the appraisal of actions associated with norm-related events, such as the fulfilment or violation of a norm.

Keywords: Emotions, Norms, Appraisal, Standards, Believability.

1 Introduction

Virtual environments often try to simulate social situations where agents should follow existing norms. We argue that for social agents to be believable, they should have emotional reactions related to the importance of the norms which are fulfilled or violated (by choice or necessity). However, despite research on how to model emotions and how to model norms in virtual agents, there is no link yet established between norm-related actions and the emotions that can arise by witnessing such actions. We propose a model to generate emotions in virtual agents that result from the evaluation (appraisal) of actions that are perceived to cause fulfilment or violation of norms. The normative emotional agents were then integrated into an architecture for virtual agents, and tested in a scenario. The results of a preliminary user study indicate that the emotional responses produced by the model were perceived by users and correlated to how strongly the user believed that the norm was important for the agents.

In section 2, we present related work on normative systems and emotions. In section 3, we present our appraisal model of norms. In section 4, we present a case study and the evaluation. Finally, we draw conclusions and discuss future work.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 97–104, 2012.

2 Related Work

There has been considerable research on how to use emotions to increase the believability of synthetic characters. Traditional animators suggest that to properly portray the emotional reactions of a character, the emotions must affect the reasoning process and consequences should be noticeable in the actions of the characters [18]. This led the computer science community to use emotion theories from psychology that model the generation of emotions in humans [3] [14] [2] [8] [9] [11].

However, emotional characters must balance their personal goals with their social environment to be believable. Normative models were developed to solve this problem without imposing hard constraints. A well-known normative system in virtual agents is *Thespian* [16]. In Thespian, obligations are created when an agent performs a certain action towards another agent. To satisfy the obligation the target agent must choose a proper action in response. Another normative system is *culturally affected behaviour* (CAB) [17], which focuses on so-called cultural norms. In this system, norms are represented using graphs named sociocultural networks, where actions are linked to cultural norms with a value that indicates whether the action conforms to the norm.

Some emotional models try to model norm-related emotions without a representation of norms, by casting norm violations as goal violations [10] or include norms in very specific domains [8]. Some normative systems, including Thespian and CAB, were further extended with emotional models [15] [4]. But those models typically are too domain-specific or lack flexibility. We argue that not only do emotions play a fundamental part in norm-related decision processes, but that the norms themselves influence the emotional state. Thus, virtual agents that connect emotions and norms will be far more believable.

3 Linking Norms and Emotions

We aim to generate emotions in virtual agents by the appraisal of actions associated with the fulfilment or violation of a norm. Hence, our agents need to have a normative model so that they can recognize norms, and when they are fulfilled or violated. The agents must also have an emotional model that evaluates the actions of agents, and generates an emotional response based on their goals and standards.

Norms prescribe what behaviours are expected in a certain social context. In our model, a norm is specified by its activation conditions, which mark the norm as active, and its expiration conditions. The behaviour prescribed by a norm is represented by a set of conditions, called normative conditions. The agent should try to satisfy these conditions, if the norm is an obligation, and avoid them, if the norm is a prohibition. If the agent succeeds, the norm is fulfilled, when it fails, the norm is violated. Our norm model is based on the work presented in [19], [5], [7] and [12]. A norm contains the following components:

- *ID:* A unique identifier that is used to identify the norm.
- *Name:* A name that describes the norm.
- *Type:* A value that informs if the norm is an obligation or a prohibition.
- *Targets:* The agents that are expected to fulfil the norm (when active).
- *Activation Conditions:* Conditions that cause the activation of the norm.
- *Expiration Conditions:* Conditions that cause the expiration of the norm.
- *Normative Conditions:* Prescriptions for the behaviour of the targets of the norm.
- *Salience:* A value that *"indicates to an individual how operative and relevant a norm is within a group and in a given context"* [1]. The salience of a norm depends on several contextual, social and individual factors (cues), such as the level of compliance and the frequency of punishment.

In our model agents monitor their own norms that they (and others) should observe. Each agent has a Normative Environment to store information about norms, whether they are active, recently expired, or were fulfilled or violated. Obligations are fulfilled when the normative conditions become true and violated if they expire without being fulfilled. Prohibitions are fulfilled as long as their normative conditions remain true, and violated when they become false.

Our emotional model follows the OCC Appraisal theory of emotions (named after its creators Ortony, Clore and Collins) [13]. According to OCC, the appraisals focused on how actions conform or not with internalized standards will trigger *"Attribution Emotions"* (pride, shame, admiration and reproach). Pride and shame occur when the agent is appraising its own actions as praiseworthy or blameworthy, respectively, while admiration and reproach arises from appraising the actions of others as praiseworthy or blameworthy.

According to OCC, the praiseworthiness of an action is often assessed in terms of its (perceived) social value. So, in our model, actions that cause the fulfilment of a norm are considered praiseworthy while actions that violate norms are blameworthy. Four factors determine the value for the praiseworthiness or blameworthiness: the salience S of the norm ($S \in [0,1]$), the estimated cost C of the action ($C \in [0,1]$), if the action was intentional I or not ($I \in [0,1]$), and if the agent is responsible R for the action ($R \in [0,1]$).

The praiseworthiness P is given by $RI\,(SW_s + CW_c)$, with W_s and W_c being a weight for the salience and for the cost, respectively. An action is only praiseworthy when the agent is perceived as having the intention and the responsibility for it. If so, this value is proportional to the salience of the norm and the cost of the action. The blameworthiness B is given by R [(SWs + CWc) Wi (1 I)], where the factor Wi reduces the blameworthiness of less intentional actions. It is also related to the salience and the cost, and zero if the agent is not perceived as responsible.

Another appraisal variable that can influence the intensity of the attribution emotions is the expectation-deviation D. For instance, the admiration we would feel for a fire-fighter saving the life of a child is likely to be less intense than the admiration that we would feel if it was the child who saved the fire-fighter's life, because the latter deviates more from what is expected. In our model, the expectation-deviation is $1-S$ if the norm is fulfilled and S if the norm is violated.

The intensity of the attribution emotion is given by $PW_p + DW_d$ if the norm is fulfilled and $BW_b + DW_d$ if the norm is violated, where W_p, W_b and W_d are weights for the praiseworthiness, blameworthiness and expectation-deviation, respectively.

4 Case Study and Evaluation

We implemented our model in an agent architecture called FAtiMA [6], a BDI architecture that endows agents with the ability to generate emotional reactions to events, based on the OCC model but in which there was no explicit notion of norms. With the addition of our model, agents constantly check if any norm becomes active or expires. Every time that an agent perceives a new event, it will check if it is an action of an agent that causes the fulfilment or violation of a norm. When a norm fulfilment is detected, the agent appraises that event and computes its praiseworthiness and expectation-deviation to determine the intensity of the resulting emotion.

Using the extended architecture, two versions of a simple social scenario were created. The scenario occurs in a bar where the user plays the role of a character that is sitting at a table with two friends (a smoker and a non-smoker) and there is a prohibition to smoke, as described in an introductory text. We made two versions of this scenario where the only difference was the salience (all weights were set to 0.5, intentionality and responsibility to 1 and cost to 0) of the non-smoking norm (see Figure 1). In the low-salient version, the salience of the norm is set as 0.1. A friend starts smoking, the the non-smoker character perceives that as a norm violation and appraises the event as blameworthy. However, the blameworthiness is so low that it is not enough to exceed the threshold for triggering a reproach emotion, thus no emotional expression is made. In the high-salient version, the salience of the norm is set as 0.9. The smoker friend still smokes since the norm, while important, is still not as important as its goal to smoke. When the non-smoker friend perceives this norm violation, it appraises the action as very blameworthy, feels a strong reproach emotion, and reacts with a frown expression and the background character gestures his annoyance.

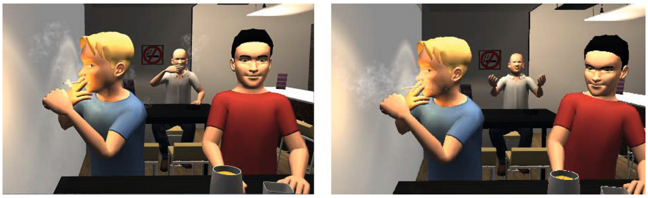


Fig. 1. In the low-salience version (left image) the non-smoker does not react emotionally, while in the high-salient version (right image) the non-smoker reacts with a frown expression

The bar scenario that was previously described was used to conduct a small pilot study. The aim of the study was to investigate if users would perceive differences in the emotional response of the agents and if those differences would relate to the specified salience of the smoking ban in the virtual environment.

Participants were randomly assigned to interact with one of the two versions of the virtual bar that were previously described. Besides the different value assigned to the salience of the smoking ban norm, all of the other parameters of the agents in the two versions (goals, relations, properties) are exactly the same.

After they interacted with one of the versions, subjects were asked about which emotions did they agreed (using a 7-point Likert scale) that the non-smoking character felt after witnessing his friend lighting a cigarette. The rationale for these questions was to check if the frowning expression of the non-smoking friend was being correctly interpreted as an emotional response.

Participants were then asked if they agreed that from the perspective of the characters the smoking ban was important and if it was acceptable to smoke inside the bar. A 7-point Likert scale was used for both questions as well. Finally, we asked participants if they smoke and also their gender, age and nationality.

In total, we had 17 Portuguese subjects (82% male), aged between 22 and 40, with the average age being 27. A total of 8 participants interacted with the low-salience version of the virtual bar and the other 9 with the high-salience version. Figure 2 shows the results obtained.

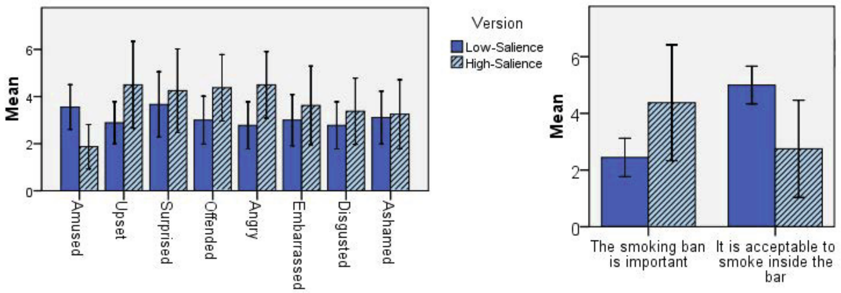


Fig. 2. The left side shows the results for the perception of the non-smoker's emotional state. The right side depicts the results obtained for the perceived relevance of the norm. (1 - Strongy Disagree, 7 - Strongly Agree)

Regarding the perception of the emotional state of the non-smoker character after the norm is violated, we found the following significant differences. In the high-salience version, the one in which the non-smoker frowns, participants agreed significantly more that the character was feeling upset $(U = 19, z = -1.7, p = .046, r = -.41)$, offended $(U = 16, z = -1.98, p = .024, r = -.48)$ and angry $(t(15) = -2.37, p = .016, r = .52)$. On the other hand, in the the low-salience version subjects perceived the character as more amused $(U = 12, z = -2.43, p = .008, r = -.59)$. There were no significant differences in the emotions of surprise, disgust, shame and embarrassment. Overall these results indicate that subjects, as expected, detected a significant change in the emotional state of the character after the norm is violated in the high salience version.

Concerning the questions about the perceived relevance of the smoking ban in the perspective of the characters, as shown in Figure 2 participants did in fact attribute a significantly higher importance ($t(8.6) = -2.114, p = .032, r = 0.58$) in the high-salience version. They also thought that it was more acceptable to smoke inside the bar in the low-salience version ($U = 11, z = -2.47, p = .0065, r = -.60$). To examine the link between these results and the non-smoker's emotional state, we run a Pearson's correlation test between the two. Concerning the user's perception of how important was the norm in the character's perspective it was significantly correlated with the perception of the non-smoker character being upset ($r = .42, p = .046$) and being angry ($r = .56, p = .01$). Similarly, the perception of how acceptable was for the characters to smoke inside the bar was significantly correlated with the non-smoker character being upset ($r = -.68, p = .001$), being angry($r = -.76, p < .0001$) and also being offended ($r = -.64, p = .003$). Although preliminary, the results obtained suggest that users were able to perceive a relationship between the emotions generated by our model and the specified salience of the norm in the scenario. This is an important result because it indicates that generating these kind of emotions from the specified norms of a multi agent environment can help users to better understand the social context the agents are simulating.

5 Conclusion

We argued that the link between norms and emotions is important to consider when modelling virtual agents, as norm-related events can be appraised and trigger emotions that will increase the character's believability. We proposed a normative model for agents to be able to recognize when norms are fulfilled or violated by actions, and an emotional model capable of generating emotions when agents witness such events. The proposed model was then integrated in an architecture for virtual agents to create two versions of a scenario where the user interacted with characters with different needs and goals, that reacted emotionally to the violation of a norm. In one version this norm had a low salience and in the second version, the salience was high. A small pilot study was conducted in which a group of participants interacted with one of the two versions created. The aim was to see how users interpreted differences in the agents emotional behaviour, with those differences being generated by our model. The results suggest that users did relate the differences in the versions to the importance of the norm. As future work we plan to extend the model by introducing enforcing mechanisms and to conduct further tests.

Acknowledgments. This work was partially supported by the Portuguese Fundação para a Ciência e a Tecnologia (INESC-ID multiannual funding) through the PIDDAC Program funds, by the European Community (EC) and by a scholarship (BD/62174/2009) granted by FCT. Additionally, it was funded by the SEMIRA (ERA-Compl/0002/2009), eCute (ICT-5-4.2 257666) and GALA (ICT-2009.4.2) projects. The authors are solely responsible for the content of this publication. It does not represent the opinion of EU or the FCT, which are not responsible for any use that might be made of data appearing therein.

References

1. Andrighetto, G., Villatoro, D.: Beyond the Carrot and Stick Approach to Enforcement: An Agent-Based Model. In: European Perspectives on Cognitive Science (2011)
2. Bates, J.: The nature of characters in interactive worlds and the Oz project. In: School of Computer Science, Carnegie Mellon University,Pittsburgh, PA Technical Report CMU-CS-92-200 (1992)
3. Bates, J.: Virtual Reality, Art and Entertainment. In: Presence 1.1, pp. 133–138 (1992)
4. Bulitko, V., et al.: Modeling Culturally and Emotionally Affected Behavior. In: Artificial Intelligence and Interactive Digital Entertainment Conference (2008)
5. Castelfranchi, C., Dignum, F., Jonker, C.M., Treur, J.: Deliberative Normative Agents: Principles and Architecture. In: Jennings, N.R. (ed.) ATAL 1999. LNCS, vol. 1757, pp. 364–378. Springer, Heidelberg (2000)
6. Dias, J., Paiva, A.C.R.: 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)
7. Dignum, F.: Autonomous agents with norms. In: Artificial Intelligence and Law 7.1, pp. 69–79 (1999)
8. Elliott, C.D.: The Affective Reasoner: A process model of emotions in a multi-agent system. PhDThesis Northwestern University 1992 (1992)
9. Gratch, J.: Émile: Marshalling passions in training and education. In: Proceedings of the Fourth International Conference on Autonomous Agents, pp. 325–332. ACM (2000)
10. Gratch, J., Mao, W., Marsella, S.: Modeling social emotions and socialattributions. Cambridge University Press (2006)
11. Marsella, S.C., Johnson, W.L., LaBore, C.: Interactive pedagogical drama. In: Proceedings of the Fourth International Conference on Autonomous Agents, pp. 301–308. ACM (2000)
12. Oren, N., Panagiotidi, S., Vázquez-Salceda, J., Modgil, S., Luck, M., Miles, S.: Towards a Formalisation of Electronic Contracting Environments. In: Hübner, J.F., Matson, E., Boissier, O., Dignum, V. (eds.) COIN 2008. LNCS, vol. 5428, pp. 156–171. Springer, Heidelberg (2009)
13. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press (1988)
14. Scott Neal Reilly, W.: Believable Social and Emotional Agents. PhD thesis. Citeseer, p. 288 (1996)
15. Si, M., Marsella, S.C., Pynadath, D.V.: Modeling appraisal in theory of mind reasoning. In: Autonomous Agents and Multi-Agent Systems 20.1, pp. 14–31 (2010)

16. Si, M., Marsella, S.C., Pynadath, D.V.: Thespian: Modeling Socially Normative Behavior in a Decision-Theoretic Framework. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 369–382. Springer, Heidelberg (2006)
17. Solomon, S., et al.: A language for modeling cultural norms, biases and stereotypes for human behavior models. Tech. rep. DTIC Document (2008)
18. Thomas, F., Johnston, O.: Disney Animation: The Illusion of Life. Abbeville Press (1981)
19. Villatoro, D., et al.: Dynamic Sanctioning for Robust and Cost-Efficient Norm Compliance". In: 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, July 16-22 (2011)

Virtual Agents in Conflict

Henrique Campos, Joana Campos, Carlos Martinho, and Ana Paiva

Instituto Superior Técnico - UTL and INESC-ID,
Av. Prof. Cavaco Silva, Taguspark 2744-016, Porto Salvo, Portugal
{henrique.t.campos,joana.campos,carlos.martinho}@ist.utl.pt,
ana.paiva@inesc-id.pt

Abstract. In this paper, we address a problem on how to model agents that engage in natural conflict situations. We propose that, in order to create such natural conflict situations, we need to rely on the agents' emotional reactions to situations. Emotional agents were created and embedded in a serious game, for helping children learning conflict resolution strategies. Agents have incompatible goals and respond emotionally to what happens in their environment. We conducted an evaluation to assess whether participants are able to perceive the conflict escalation process according to the agents' emotional behaviour. The results suggest that actions intensity, which changes due to emotional states, conveys the idea of conflict escalation and conflict is perceived.

Keywords: Pedagogical Environment, Intelligent Virtual Agents, Conflict, Emotion.

1 Introduction

Conflict is a normal part of everyone's life and it should be considered as a constructive process that makes our society move forward. Although conflict had been considered to be something to avoid due to the negative feelings and destructive behaviours associated to it, recent research acknowledge that conflict can yield beneficial aspects as well [19].

The work described in this paper is integrated into the SIREN[1] project, which aims at exploring games as a tool to teach conflict resolution skills to children. Games support learning in various forms as the virtual setting responds differently depending on the player's choices. In addition, people can take different roles, experience different perspectives and realise the consequences of their actions [10]. Yet, engaging players in learning-oriented games is a hard task. Over the years, the balance between learning and engagement has been approached, by using autonomous synthetic characters. These intelligent virtual agents are integrated into the game to affect the user's engagement and empathy towards the game's characters ([1] and [18]). It is by interacting with virtual characters and exploring the environment that the players will learn, by practice, to master skills they don't have [15]. Game learning environments have been developed

[1] http://sirenproject.eu

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 105–111, 2012.

for different purposes such as, raising awareness on general population, teaching students about a subject or army training and education [3][18]. In addition to that, using games for conveying children conflict resolution skills have also been explored, as it is the case of *FearNot!* [1] or *The Prom* [13].

These games deal with variations of the conflict topic, such as bullying or relationship/friendship management based on the characters personalities. In the particular case of *FearNot!*, the user takes responsibility for the victim of bullying and has to help her to make decisions. On the other hand, in *The Prom* environment, the user manages social relationships by taking actions that will balance the social world. Its focus is on the importance of characters' personalities to the social exchange.

In this paper we describe a prototype model of conflict that intends to convey conflict related aspects by the means of emotional agents. We explore how deep elements of conflict, such as one's emotions, convey aspects as conflict emergence and its escalation. The model is mainly focused on overt manifestations of conflict which are influenced by the emotional state of the agent. We believe that the cognitive appraisal is an important element to capture the essence of real conflict scenarios. This model was then embedded in a game scenario, which works as tool to support learning. Finally, we performed an evaluation in order to understand if the users could perceive conflict, due to the agent's behaviours, and whether they perceive the role of emotions in conflict escalation.

2 Background - Conflict

In the literature, there is not a reconciled definition for conflict. However, we may say that conflict varies along five dimensions: participants, causes, initiating action, participants' responses (one's attitudes, behaviours or strategies) and outcomes. Furthermore, conflict episodes have been compared to a plot [9], they have an initiating action (complication), a rising action (set of actions that contribute to conflict escalation), a climax and the outcome.

When the conflict gets worse, we say that it escalates. When it reaches the turning point (*climax*) and the magnitude of the situation decreases, we say that de-escalates. Escalation occurs when one or both parties engage in the conflict, moving it from a less severe stage to more contentious and heavy state [16]. We may say that escalation is driven by inner triggers [8], that is, emotions that weight one's current goals and assess the affective value of the situation [11].

To bring the situation to an end participants in the conflict may take several approaches according to Thomas' taxonomy [20] such as: *accommodation*, *avoidance*, *competition*, *collaboration* and *compromise*. These approaches are underlined by the dimensions of assertiveness and cooperativeness, which are phrased as intentional terms. Assertiveness refers to the extent to which protagonists try to achieve their own goal and cooperativeness refers to the extent of protagonists trying to satisfy the concerns of others.

3 Capturing Real Conflict Situations with Agents

Emotions are at the heart of social interaction and they play a relevant role triggering events such as conflict [14]. Therefore, the cognitive appraisal is an essential element for understanding conflict, where such situations emerge from one's subjective evaluations of the environment. Our model is illustrated in Figure 1. The conflict dynamics specified by our model tries to capture the essence of Thomas' definition [20] of the phenomenon, in which conflict is defined as "the process which begins when one party perceives that another has frustrated, or is about to frustrate, some concern of his". Further, we inspired in FAtiMA emotional model for agents [7] and Tessier et. al [19] conflict-handing model.

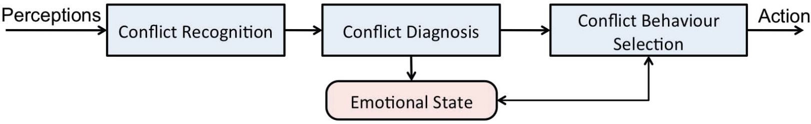


Fig. 1. Conflict handling model

The model consists in three main modules, where each works as follows.

First, in the **Conflict Recognition Module**, others' actions or events that affect (positively or negatively) a certain concern of the agent are perceived. These are checked, in order to evaluate whether they raise potential conflicts or contribute to the *escalation* of the current situation. It is specified the urgency of conflict, which determines how intense the situation is.

After that, in the **Conflict Diagnosis Module**, the diagnosis process is executed in two steps. First, a conflict description is generated, which depicts the *cause* (goal frustrated), *participants* involved, relationship between them and the importance of the conflict. With this description, emotional reactions are triggered and emotions are generated (this process is undertaken by FAtiMA [7]), where the intensity of the emotion reflects the urgency of the conflict [12].

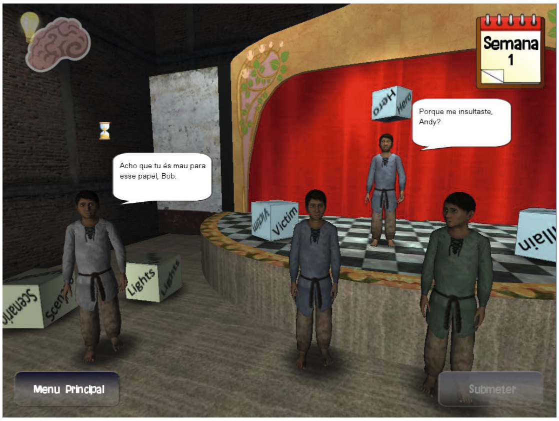
Finally, in the **Conflict Behaviour Selection Module** the behaviours for handling conflict range within the *assertiveness* and *cooperativeness* dimensions (see Section 3). For simplification reasons, we considered *Attacking* and *Evading* behaviours (from Raider's AEIOU model of communication in conflict [17]), which are associated to [*high* assertiveness, *low* cooperativeness] and [*low* assertiveness, *low* cooperativeness], respectively. The values of assertiveness and cooperativeness are balanced by the agent's emotional state. The reason behind this choice is based on the assumption that negative emotions are linked to less cooperative approaches [6], which will lead to more conflicts and their consequent escalation. For example, an agent becomes less cooperative as he gets more frustrated with the situation at hands. In this model, whether an agent is more prone to one of the aforementioned behaviours is determined by personality traits.

Our current investigation only aims to model simple behaviours, in order to demonstrate escalation. Therefore, we decided to focus on a set of behaviours that might lead to potential conflicts and their escalation to explore what believable conflict-oriented behaviour might be at the eyes of the human perceiver. In this way, we set that negative emotions will affect negatively the agent's actions towards conflict diminishing the possibility for cooperation [2]. A broader and more detailed set of behaviours will be developed in future work. The following Section describes the scenario where this model was applied and tested.

4 Case Study: My Dream Theatre

The *My Dream Theatre*[2] (see Figure 2) is an educational game that aims at teaching children, aged 9 to 11, some conflict resolution skills. The game setting is a theatre company and the user/child is challenged to be the director and to select the adequate cast for each performance.

Each virtual actor has a set of characteristics, such as: a proficiency level, preference for roles, interests and personality. As the player grants roles to the characters, conflict situations may emerge when characters themselves perceive an obstruction to their goals (e.g. hero role). How the agents appraise the situation will make their responses vary and consequently the interplay between characters will vary as well. The role of the child is to manage the conflict, advise the agents, and try to do so in a manner that the conflicts are resolved for a better performance in the end.



Porque me insultaste, Andy?

Acho que tu és mau para esse papel, Bob.

Fig. 2. *My Dream Theatre's* screenshot, showing two characters having a discussion about a role

[2] The assets of the game scenario were developed by Serious Games Interactive (http://www.seriousgames.dk/).

In this early prototype, we decided to model the *Attacking* and *Evading* behaviours, as we believe these are more likely to generate *escalation* as a result of what these behaviours bring to the social interaction. The conflict model (previously described) was implemented in FAtiMA emotional agents' architecture and integrated into the agents' minds. Inspired by natural conflict behaviours from the literature, agents with the tendency to *Attack* follow a destructive path to cope with the conflict [17]. These agents are prone to have high assertiveness and low cooperativeness. The actions, taken by agents with this tendency, range from a low level of aggressiveness to an extreme. For example, as the agents' emotional state worsens, their actions may progress as follows: lesser insult, criticise negatively, harsh insult, and threat. On the other hand, an agent with an *Evading* tendency may try to avoid conflict situations. Initially, this agent may want to cooperate [17]. However, the build up of negative emotions leads the agent to become less cooperative. Furthermore, as the emotional state gets worse, the actions performed by an agent with this kind of behaviour progresses as follows: ignore the situation, sacrifice own's goals to avoid further involvement and, finally, leave the scene.

In order to illustrate the agents' behaviours, consider the following scenario. In the first session with the *My Dream Theatre* game, the user has to direct a play where two characters, Andy and Bob, share the desire for the same role, the "Hero" role. For the user, the most rational choice for the part is Bob. Bob has higher proficiency and he is more cooperative comparatively to Andy. The complication starts when Bob receives the role. Andy appraises the situation as a negative interference to his self-interests what generates a negative emotional state. This is aggravated by the fact that Andy considers the role highly important to him. This trigger makes Andy upset enough to approach Bob aggressively, by verbally insulting him. With that, Bob who was initially happy, starts feeling upset, but as he was given his preferred role, he limits himself to only question the reason of the insult, trying to resolve this situation. Andy disapproves Bob's approach and gets even more upset. As the situation gets even more intense, as it escalates, Andy attacks Bob, who eventually reaches a high level of frustration. In the end, if the user doesn't intervene, Bob will end up giving up the Hero role, which is not good for the play.

5 Evaluation

In order to test the conflict model, we performed a preliminary evaluation phase, where we tried to assess whether people were able to recognise a conflict interaction by evaluating its participants' behaviours, contributions and outcomes. For that, we conducted a between-groups evaluation where participants were exposed to our model of conflict, full model (FM) condition, or to a control condition, a simplified model (SM) condition, where agents had no emotional affect on their behaviours.

A total of 80 participants (19 females, 61 males aged 14-48)[3] took part in the study, which was available through an online questionnaire that randomly assigned participants to one of the above test conditions. After watching a video of a user interacting with *My Dream Theatre*, which presents a situation similar to the one portrayed in Section 4, participants rated characters' behaviours and the situation process through 5-point likert scales.

The questions used were adapted from a self-serving questionnaire on conflict behaviour and escalation [5]. The data was analysed using the Mann-Whitney test. The first set of questions comprised characters' behaviours towards the conflict. Andy's behaviour, in the FM condition, was considered significantly ($p < 0.001$) more hostile, more competitive and evil-minded, compared to the control condition (SM). These results are consistent with its internal drives to follow a destructive path in a conflict interaction. On the other hand, Bob's attitude was considered to be more constructive. Participants rated Bob as significantly ($p < 0.001$) more friendly, collaborative and good-hearted in the FM condition, than in the SM condition. The second set of questions assesses the escalation process of the situation, in which participants reported that, in the FM condition, Andy significantly ($p < 0.001$) obstructed more Bob's goals. Further, only Bob was reported to significantly ($p < 0.001$) become more frustrated, in the FM condition. Nevertheless, the ambient was reported to worsen significantly ($p < 0.001$) more in the FM condition, compared to the control condition (SM). The detailed analysis of the results can be found in [4].

6 Conclusions

In this research, we aim to address the conflict phenomena by creating agents, which engage in natural situations of conflict, in the environment of an educative game to teach children conflict resolution skills. At the heart of this prototype is a model of conflict behaviour implemented in FAtiMAs emotional architecture, and integrated in the *Dream Theater* prototype. The modelled behaviours were inspired on natural conflict scenarios and agents' reactions are a result of their emotional state. To address whether the agents were effective in simulating a natural conflict scenario we performed an evaluation on our scenario. The results suggest that the agents' emotional behaviours are consistent with the process of conflict escalation, as well as, "attacking" and "evading" behaviours inspired by the literature on the subject.

Acknowledgements. The research leading to these results has received funding from European Community's FP7 ICT under grant agreement n°258453, FCT (INESC-ID multi-annual funding) through the PEst Program funds and a PhD scholarship (SFRH/BD/75342/2010) granted by FCT.

[3] Although the age range it is not the same as the target population for the learning game, we wanted to rapidly test the participants perception of conflict as a result of the agents behaviour before proceed.

References

1. Aylett, R.S., Vala, M., Sequeira, P., Paiva, A.C.R.: FearNot! – An Emergent Narrative Approach to Virtual Dramas for Anti-bullying Education. In: Cavazza, M., Donikian, S. (eds.) ICVS-VirtStory 2007. LNCS, vol. 4871, pp. 202–205. Springer, Heidelberg (2007)
2. Bell, C., Song, F.: Emotions in the conflict process: an application to the cognitive appraisal model of emotions to conflict management. International Journal of Conflict Management 16(1) (2005)
3. Buch, T., Egenfeldt-Nielsen, S.: The learning effect of global conflicts: Palestine. In: Conference Proceedings Media@Terra, Athens (2006)
4. Campos, H.: CONFLICT: Agents in Conflict Situations. Msc thesis, Instituto Superior Técnico (May 2012)
5. De Dreu, C.K.W., Nauta, A., Van de Vliert, E.: Self-serving evaluations of conflict behavior and escalation of the dispute. Journal of Applied Social Psychology 25 (1995)
6. Deutsch, M., Coleman, P.T. (eds.): Handbook of Conflict Resolution. Jossey-Bass Publishers (2006)
7. 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)
8. Kriesberg, L.: Escalation of Conflicts. In: Constructive Conflicts: From Escalation to Resolution, 3rd edn., ch.6, Rowman & Littlefield (2007)
9. Laursen, B., Pursell, G.: Conflict in Peer Relationships. In: Handbook of Peer Interactions, Relationships, and Groups. Guilford Press (2009)
10. Lieberman, D.A.: What Can We Learn From Playing Interactive Games, ch. 25, pp. 379–397. Lawrence Erlbaum Associates (2006)
11. Lindner, E.G.: Emotions and Conflict. In: Handbook of Conflict Resolution, 2nd edn., ch. 12, pp. 268–263. Jossey-Bass (2006)
12. Maiese, M.: Emotions. In: Burgess, G., Burgess, H. (eds.) Beyond Intractability, Conflict Information Consortium, University of Colorado, Boulder (July 2005), http://www.beyondintractability.org/bi-essay/emotion
13. McCoy, J., Treanor, M., Samuel, B., Tearse, B., Mateas, M., Wardrip-Fruin, N.: Comme il faut 2: a fully realized model for socially-oriented gameplay. In: INT3 2010: Proceedings of the Intelligent Narrative Technologies III Workshop. ACM (2010)
14. Nair, N.: Towards understanding the role of emotions in conflict: a review and future directions. International Journal of Conflict Management 19(4) (2008)
15. Oblinger, D.: Simulations, games and learning. Educause (2006)
16. Pruitt, D.G.: Some research frontiers in the study of conflict and its resolution. In: Handbook of Conflict Resolution, 2nd edn., ch. 37. Jossey-Bass (2006)
17. Raider, E., Coleman, S., Gerson, J.: Teaching conflict resolution skills in a workshop. In: Handbook of Conflict Resolution, 2nd edn. Jossey-Bass (2006)
18. Rowe, J., Mott, B., McQuiggan, S., Robison, J., Leea, S., Lester, J.: Crystal island: A narrative-centered learning environment for eighth grade microbiology. In: AIED 2009: Workshops Proceedings (2009)
19. Tessier, C., Müller, H.-J., Fiorino, H., Chaudron, L.: Agents conflicts: New issues. In: Conflicting Agents. Multiagent Systems, Artificial Societies, and Simulated Organizations, vol. 1, Springer, US (2002)
20. Thomas, K.W.: Conflict and conflict management: Reflections and update. Journal of Organizational Behavior 13 (1992)

How Do You Like Me in This: User Embodiment Preferences for Companion Agents

Elena Márquez Segura[1], Michael Kriegel[2], Ruth Aylett[2],
Amol Deshmukh[2], and Henriette Cramer[1]

[1] Mobile Life @ SICS
DSV, Forum 100, 164 40
Kista, Sweden
{elena,henriette}@mobilelifecentre.org
[2] School Of Mathematical and Computer Sciences
Heriot-Watt University,
EH14 4AS, Edinburgh, Scotland, UK
{m.kriegel,r.s.aylett,a.deshmukh}@hw.ac.uk

Abstract. We investigate the relationship between the embodiment of an artificial companion and user perception and interaction with it. In a Wizard of Oz study, 42 users interacted with one of two embodiments: a physical robot or a virtual agent on a screen through a role-play of secretarial tasks in an office, with the companion providing essential assistance. Findings showed that participants in both condition groups when given the choice would prefer to interact with the robot companion, mainly for its greater physical or social presence. Subjects also found the robot less annoying and talked to it more naturally. However, this preference for the robotic embodiment is not reflected in the users' actual rating of the companion or their interaction with it. We reflect on this contradiction and conclude that in a task-based context a user focuses much more on a companion's behaviour than its embodiment. This underlines the feasibility of our efforts in creating companions that migrate between embodiments while maintaining a consistent identity from the user's point of view.

Keywords: Embodiment, HRI, User Preferences, Scenario-based study.

1 Introduction

Embodiment is currently a prominent issue in agent research. The field of embodied cognition [5] argues strongly that in the human case, mind and body cannot be separated, but form an interlinked whole. The field of situated agents in turn has argued that embodiment is an important part of the situational coupling between an agent and its environment [23]. With the Uncanny Valley [17], in which near-human embodiments can cause significant negative reactions from interaction partners if appearance and behaviour are at all inconsistent, we see the vital role that embodiment also plays in interaction. Embodied Conversational Agents [4] are a practical demonstration of the importance of an embodiment in non-verbal communication, extending an agent's communicative bandwidth substantially beyond its explicit messages.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 112–125, 2012.

For these reasons, our investigation of long-term companions in the Lirec project[1] has focused on embodied agents, whether with robotic or graphical bodies. The project has also addressed migration of an agent or companion between different embodiments [13] so that for example the mobility limitations of a robotic body can be overcome by moving the software-generated ‘personality’ of the companion into a hand-held device.

This raises significant questions. Some related to functionality (a hand-held device cannot pick up a telephone but a robot can), and others to how users relate to such a companion in their interaction. Does the fact that a robot interacts in a shared physical space with a user, while a graphical character does not, make a difference? In a project studying long-term companionship, this is a question deserving investigation.

The study reported in this paper involves a companion known as Sarah. This can be embodied in a robot, on a large graphical screen or in a handheld device. The facial appearance was deliberately chosen so as to be robotic rather than naturalistic in order to avoid the possibility of Uncanny Valley effects. The social context for Sarah is a university work environment. The robot embodiment is designed to act as a Team Buddy for a group of researchers in a specific lab, issuing reminders about activities, and taking and delivering messages for absent team members from visitors to the lab as example activities. For this reason, the scenario developed for our study involved office tasks. The objective of the study was to establish within this context whether two different embodiments made a difference to interaction with the companion with the broader goal of contributing to design guidelines for embodied companions.

2 Related Work

Embodied agents, including social robots, have been designed to improve interaction for users, taking advantage of social cues that ease coordination between human user and agent and offering a more engaging experience [3]. The particular embodiment chosen has shown to affect how people respond to an agent [16]. For a physically embodied agent, additional social aspects come into play. Physical, expressive movement of the agent can shape perceptions of an agent’s behaviours and intentions [18]. Proxemics, relating to the physical distance between interacting humans [9], with different interpersonal distances considered appropriate for specific contexts, also play a more pronounced role.

A number of studies investigated the differential impact of embodiment both on users’ perceptions, and in some cases on task execution. In some of these studies [1, 2, 12], a real co-located robot was compared with a filmed version of the same robot. In others, a real robot was paired with a graphical version of it [6, 11, 15, 20, 21, 26]. A wide range of tasks and contexts can be found in these studies, however none focused on an office environment as this study does. In some, verbal communication pre-dominated. These included: taking care of a pet [6]; playing in a children’s game of chess [20]; a health interview [21]; the Japanese game of Shiritori [11]. A smaller number involved physical manipulation: interactive drumming [12], a Towers of Hanoi-type puzzle [26] and moving books about in a room [2].

[1] http://lirec.eu

These studies all showed users preferred the physical robot to either a video or a graphical character. In addition a variety of other effects were detected. Thus, playing chess against a real robot was classified as significantly more fun than with a graphical version [20]. In the health interview, subjects forgot more and disclosed least with the physical robot, but spent more time with it and had more positive attitudes towards it [21]. In the interactive drumming game with children, enjoyment of game with the physical robot was higher, and its perceived intelligence and appearance received higher scores. Looking at the task, in the physical robot case, more interaction, better drumming and turn taking were observed [12]. In the Towers of Hanoi type task, users had more positive feelings about a co-located real robot, but interestingly no significant increase in task performance was found [26]. In the book moving task, the most similar to the one in our study, subjects were more likely to agree to an unusual instruction (throwing a book into a bin) given by the physically-present robot and gave it more physical space, than to a video version of the robot [2].

Interesting issues emerge from these studies, relevant to our own. One is the impact of a physically-embodied robot on the user's feelings compared to the impact on their behaviour. A second is how far the study required interaction with the agent. In some cases this was continuous throughout the study, as in collaborative game playing, while in others it was a less-coupled collaboration more typical of the office environment. In the more collaborative tasks, users remained in a fixed position in relation to the agent rather than moving around in the physical space where the agent is located. Finally, some tasks were essentially information-exchange based, while others involved physical manipulation of objects.

While these studies show an overall preference for physical robots, a full understanding of the effects of embodiment has not yet been reached. We aim to further investigate the effect of the expectations raised by an agent's embodiment on the users' attitudes towards the agent and their behaviour during an interaction with an agent. Building on the previous work above, a scenario and agent behaviours were designed to explore these issues further.

3 The Study

The importance of agent appearance matching user expectations is apparent in [8]. Our agent was embodied in two different forms: one robot agent (r-agent) and one virtual agent (v-agent) displayed graphically on a large free-standing screen that is roughly equal in size to the r-agent. Both display the same expressive behaviour (speech, gaze, gestures, posture [22]), reactions, and have the same facial appearance. We hypothesise that: i) the perception and expectations the users have of the agent depend on the type of embodiment; ii) this will shape the kind of interaction they expect and accept from the agent; and iii) this in turn will affect the users' assessment of their interaction with the agent. Specifically:

- Participants will rate interaction with the r-agent more positively.
- They will have higher tolerance for, and acceptance of, potentially annoying behaviour from the agent, such as inappropriate interruptions and mistakes.

- Users will perceive the r-agent as more similar to a human being in terms of communicative capabilities compared to the v-agent.

In order to study the second hypothesis (ii)), we specifically designed the agent's behaviour of the agent to test the user's tolerance of, trust in and acceptance of the agent.

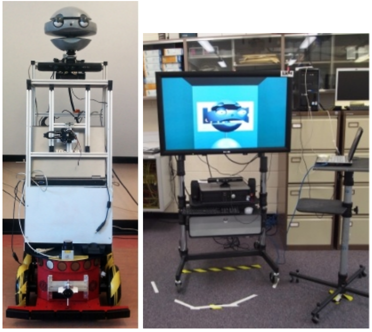


Fig. 1. Robotic (a) and virtual (b) embodiment of the agent

Due to the nature of our scenario, we focused on the communication capabilities expected in the agent. Not only office co-workers but also visitors will interact with the agent. It is therefore important to gain insight into how naive users naturally try to communicate and provide input in different tasks. As in [10], this study recruited subjects that had not had much contact with robots or virtual characters. Apart from helping with the design of agent communication behaviours, this may help to predict what communication behaviours agents in different embodiments are likely to meet. Our findings may suggest guidelines for communication models so that they better fit observed human tendencies in communication.

A scenario was designed in which the user performs activities associated with office tasks, such as finding a paper, working on it, and placing it back somewhere else. The agent assisted with these tasks.

The two embodiments of the companion called Sarah used for this study had the same capabilities and appearance. Each had a non-naturalistic head, modelled on that of a turtle, used a unit-selection text-to-speech system with a Scottish English female voice, and carried out body tracking of the user using a Kinect sensor installed on the embodiment. A user interface was developed to allow a wizard to remotely control speech acts and direct the gaze of the agent head to indicate different locations in the room. The r-agent consisted of a Pioneer P3AT robot with an enhanced superstructure and head, equipped with a laptop PC and Kinect sensor, as in Figure 1a. The graphical v-agent consisted of an animated 3d model of the same robot head displayed on a 42 inch LCD screen. A Kinect is installed underneath the screen as shown in Figure 1b.

3.1 The Experiment

We performed an experiment in which participants worked with Sarah as v-agent or r-agent in a role-play situation in which they had to interact to fulfil allocated tasks. Social talk, averted attention, and an occasional mistake were included in the behaviour of the agent so as to study the possible influence of the type of embodiment upon the users' responses to these behaviours and their perception of the agent. We chose a Between-Group rather than Within-Subject experimental design, since logistically it would have presented a challenge to switch the embodiment during each session. Moreover, it would also have been difficult to recruit subjects willing to invest such a significant amount of time. Finally, subjects would have been aware of our 2 experimental conditions had they interacted with both embodiments, which may have biased their answers.

We recruited 42 participants (16 female, 26 male), aged 23 to 56, from the university campus where the role-play was carried out. Eighteen were students (42.9%); most liked gadgets (95.2%); had no previous experience using virtual characters (62.5%) or robots (73.8%); and did not have technical knowledge about virtual characters (85.7%) or robots (81.0%). Participants engaged individually in a 30 to 45 minutes session, in which about 10 minutes were spent interacting with the agent. Before this, the participant filled in a closed questionnaire that took about 10 minutes. This questionnaire included demographic questions, questions related to the participant's use and knowledge of virtual agents and robots, an abridged 10-measure version of the Big-Five personality test [7], and a subset [24] of the Negative Attitude Towards Robots (NARS) questionnaire [19] mainly in the form of 7-point Likert scales. After this, the participant read the following brief:

"Bob and Paul are professors at this university who work together in the Lab you are entering. Bob is now on holiday and needs to mark some exams. He has forgotten one in the lab and has asked you to do that for him. If there's no one in the lab, don't hesitate to ask the Team Buddy for anything you need.

So, your tasks are to: 1) Get the exam paper that Bob has forgotten, 2) Mark that exam paper by comparing it with the answer sheet provided. Score every question and write the final score on the paper (correct answers give 1 point, incorrect answers -0.5 point and unanswered questions 0 points), 3) After you have successfully marked the paper give it to Paul or place it in his mail box if he is not in the office.
When you finish the task, you can come back to this room."

Afterwards, the participant answered a 5 minute closed questionnaire, mainly focused on their perception of the agent and their experience of interacting with it. Finally, a semi-structured interview about 10 minutes long was conducted to gain more insight in the participant's perception of the agent and its behaviour, with particular emphasis on the three behaviours of social talk, averted attention, and mistakes, further described below.

The initial questionnaire, the post-interaction questionnaire, and interview took place in a reception room. The wizard operated the companion's speech and behaviour (gaze) in the same room, in an area concealed by a partition, out of the sight of the participants. The interaction was conducted in a second room, where the participant was directed after the initial questionnaire and instructions. This room was unoccupied except for the participant and the embodied agent. The agent was placed so that it was visible from the entrance and four cameras were used to record the interaction. Both the agents were placed in the same position in the room, but only one agent was prepared for the experiment at a time, the other was hidden.

Two cameras recorded the user's facial expressions at two points during the interaction: 1) while the participant is marking the paper, moment in which the companion interrupts and tries to engage in small talk, and 2) at the end of the experiment when the participant has to leave the exam paper but the companion gives incorrect information about the container into which to drop it.

The two other cameras were used to record the interaction covering the whole room. The Wizard acted as the speech recognition system for the agent, with live audio and video streams from the role-play and a set of scripted sentences from which to choose the agent's response. Occasionally, if no scripted sentence matched the situation, sentences would be typed directly. This was only done for unanticipated but feasibly recognisable utterances from the user. The use of this approach was motivated by the substantial resources that would be required to develop a sufficiently functional speech recognition system to maintain a relatively natural interaction. However we consider the type of speech recognition simulated technically possible. We also expected this experiment to help establish the training data that would be needed to successfully recognise the speech used in this specific setting.
Four modes of agent behaviour were designed for the scenario:

1) **Diligent and cooperative.** The default mode. The agent is cooperative and diligent, focused on the participant and responsive to their requests. Except for the initial greeting when the participant first comes to the room, the companion takes a secondary role in the interaction, leaving the participant to take the initiative in the conversation. Its indications are accurate and help the fulfilment of the participant's tasks.

2) **Interruptions and social talk.** Initiated when the participant starts marking the paper. The agent engages in small talk by asking the participant a battery of 6 prepared questions like "*so... what did you have for breakfast?*". If a participant tried to take the initiative by asking the agent similar questions, they were cut short by a generic reply such as "*I'm not programmed to understand everything*". If the participant expressed an explicit or implicit wish to stop this interaction (e.g. "*I am trying to focus here*", "*Do you want me to finish my task?*") or ignored the agent, it would acknowledge the situation and announce its shift to the averted attention behaviour mode ("*You seem distracted, I'm going to leave you alone and watch some videos*"). Otherwise, after the last question in the battery, the agent would ask the first question again ("*What time is it?*") and then, without any announcement, shift to the averted attention mode.

   The aim of this behaviour is to study the influence of the embodiment in engaging the user in social talk in a context in which they are busy. It was discussed during the semi-structured interview, with a focus on whether the interruptions were annoying or not, on the appropriateness of this interaction, and overall whether the participant would like a companion that engaged in social talk during working time.

3) **Averted attention.** Initiated after the previous behaviour mode. It is explicitly announced if the participant expresses a wish to stop the social interaction with the agent, or unannounced in the case the participant keeps interacting with the agent until the last question is asked. The agent watches videos displayed on a laptop placed nearby. If the participant asked any questions, the agent would ignore it and keep its gaze fixed on the screen of the laptop. The agent stays in this mode unless the participant actively seeks the agent's attention, either by asking repeated questions (three times), or making a gesture such as waving.

The aim of this mode is to study how the participant would seek the agent's attention and whether this varies for the different embodiments of the agent.

4) **Mistake.** The final behaviour starts when the participant asks the Team Buddy for information about the location of Paul's mailbox. Instead of indicating where it is, the agent describes the location of a box labelled "Trash". This is inspired by [2]. The aim is to study the influence of the embodiment on the participant's trust. If the participant explicitly questions the agent's response by stating the location was incorrect or the object pointed is not the mailbox, but the trash, the agent would ask: "*Are you sure?*" If the user then continues questioning the agent (e.g. "*Yes, I am sure*"), the latter admits its mistake, apologizes for it, and gives the right location.

The behaviour of the companion also included "body language". We expected gaze would improve the interaction [25] and, inspired by [18], designed a set of agent gaze behaviours for the different situations:

i. **Pointing gaze** to indicate locations: the agent moves its head and eyes towards one location while making a mild surprise gesture with the forehead, a facial behaviour used to highlight or indicate something. This was used at the two moments when the agent indicates where to find and leave the exam paper.
ii. **Gaze fixed away** from the user, when the agent is engrossed in looking at videos displayed on a laptop placed nearby as part of the averted attention behaviour.
iii. **Tracking gaze**: the default gaze used in the diligent and cooperative mode. The eyes of the agent follow the user. This gesture is designed to invite interaction engagement by indicating that the user is the focus of the agent's attention.

3.2 Evaluation

Data was gathered by a closed questionnaire, an open semi-structured interview, and through the video recordings of the interactions. In this paper we only discuss the findings from the first two data sources. The perception and expectations of the users were evaluated via questions in the closed questionnaire and in the interview. Inspired by [14], we focused on the communicative and cognitive capabilities of the agent. The participants were asked to reflect on whether they changed and if so, how they changed, their manner of speaking.

The tolerance and acceptance of potentially annoying behaviour from the agent was evaluated through all three types of data source: video analysis gave an indication of how much the participants trusted the agent when there was an obvious mistake in its information (behaviour mode 4 above), the questionnaires covered this issue (e.g. raising the extent to which the agent is trustworthy, or honest), which were then further explored during the open interview. The questionnaire also contained items related to the participant's assessment of the whole experience and the agent, such as "*I enjoyed the interaction with the companion*".

4 Results

Out of the 42 interactions, 4 were discarded due to technical problems. All results reported consequently are based on the remaining 38 interactions, 19 for each of the two embodiments.

4.1 Questionnaires

Analysis of the pre-interaction questionnaires showed, as expected, no significant differences in either participant personality or negative attitude towards robots between the 2 conditions. We can therefore discard the possibility that any differences found between the 2 groups are due to certain personality trait or NARS clusters within the groups. Regarding post-questionnaires, we investigated items indicative of users' overall perception of the interaction and of the 4 agent behaviours listed above. Variables were collected via responses to statements such as "*The companion was very capable in helping me*" on 7-point Likert Items (1-Disagree Strongly, 2- Disagree Moderately, 3 - Disagree A Little, 4 – Neither Agree Nor Disagree, 5- Agree A Little, 6 – Agree Moderately, 7 – Agree Strongly).

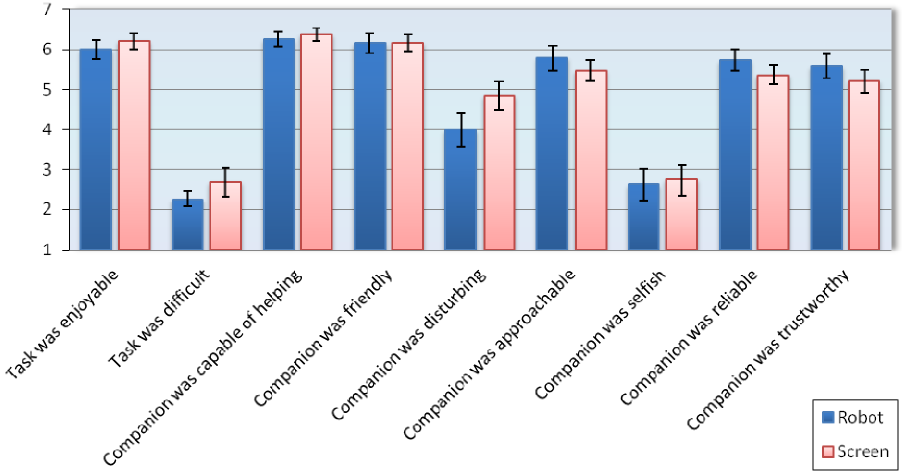


Fig. 2. Post Questionnaire Results for both participant groups

Users overall stated that they enjoyed the interaction with the companion. Looking at measures related to the cooperative agent behaviour, participants overall disagreed with the task being difficult and found the agent very capable of helping. While the agent was perceived as friendly, opinion on whether it is disturbing were more mixed; the agent on average being perceived as just a little disturbing. Both these measures are related to the agent's interruptions and social talk behaviour. Despite its averted attention behaviour the agent was perceived as approachable and not selfish. The agent was also seen as moderately reliable and trustworthy despite its mistake behaviour. Figure 2 summarizes these results.

Kolmogorov-Smirnov tests revealed that none of the dependent variables above are normally distributed. We therefore ran a series of Mann-Whitney U tests to determine whether there are differences in these variables between the group of participants that interacted with the r-agent and those that interacted with the v-agent. The tests revealed that none of the variables reported above is significantly influenced by the type of embodiment. This suggests that many aspects of perceived agent personality are dominated by behaviour and not by embodiment. One should however keep in mind the possibility that significant differences between the groups exist, that were simply not captured by the questions we asked and that would have been brought to light by different questions.

4.2 Open Interviews

The open semi-structured interview yielded interesting results concerning user preferences for one or the other agent embodiment. Participants were confronted with the idea of interacting with the other embodiment than the one they had seen. This was done verbally without showing the other embodiment. They were asked whether they thought the interaction would have been different, with which embodiment they would rather interact, and why.

Twenty six of 38 participants stated they would prefer to have interacted with the robot. Of those, 11 had interacted with v-agent and 15 with r-agent: Of those who interacted with the r-agent and would not choose it as a preferred embodiment (4 participants), 1 did not answer, 2 said the interaction would not have been different for any embodied agent is a human person, and 1 would rather have interacted with the screen, for she was more used to it and would found it easier.

Of those who interacted with the v-agent and would not choose the r-agent (8 participants), 5 said the interaction would not have been different, but then commented on advantages of interacting with the robot ("*Maybe the robot can turn the head. That's useful for directions*"), 1 did not choose any embodiment, but stated the interaction would have been better with the robot (better communication), 1 did not choose any embodiment, said that the interaction would have been different and pointed at advantages and disadvantages in each case, and finally 1 said the interaction would have been the same. Of the 38 participants, 34 motivated their choice, which gives us insight into what is important for the user in each embodiment:

Social Presence - The main reason for choosing the r-agent was its presence. Of the 34 who gave reasons in favour of one or another embodiment, 14 mentioned presence in one or another way (e.g. "*more social presence*", "*more physical presence*", "*more present*").

Movement - Movement was the second most discussed reason. Of the 34, 12 mentioned movement in one or another way: some mentioned the fact that the movement of its eyes and/or head would be more apparent and noticed compared to the movement of v-agent's head/eyes; others assumed the r-agent would have been able to pick things up or move from one location to another one. Anticipating that movement might be a reason to favour the r-agent, we specifically asked whether the participants would like the r-agent to move about. Of 20 responses, most of them (13) would find it useful. However, a few of them (7) would find it scary.

More Interactive, Better Communication - Nine of 34 mentioned the communication or the interaction would improve with the r-agent (e.g. "*better communication*", "*more interactive*", "*more approachable*", "*multimodal interaction*").

Comfort, Company, Relationship - Six participants mentioned the r-agent would make for a more approachable agent (response e.g.: "*more approachable*", "*more comfort*", "*company*", "*display of emotion*", "*relationship*".)

Engaging, Fun, Interesting - Six participants mentioned the r-agent would make for a more interesting experience (e.g. "*fun*", "*engaging*", "*interactive*", "*interesting*").

Real vs. Not Real - Four participants associated the r-agent with "something real": E.g. "*It seems more real, with more personality*".

On the other hand, five participants described the interaction with the v-agent as if talking to somebody on TV, or interacting with a TV, or a computer, or a tutorial: "*The robot is more real… the screen is like a tutorial […]. The robot is …. More verbal… interactive… more interesting*".

Surprise - Twenty six of the participants gave their first thoughts regarding their experience. Of those, about half of them (12) mentioned they were surprised by the agent, its capabilities, or the scenario: "*I thought it was going to be more mechanical*". It seems the capabilities of the agent surpassed the participants' expectations: *"I was surprised how well he understood me", "It's so funny! It does a lot more than I thought!", "I wasn't expecting it to be that smart!".*

Participants were asked about the point when they were marking the exam paper and the agent was in interruption/social talk mode, and whether the interruptions were annoying. Of seventeen participants that interacted with the v-agent, 7 did not find it annoying, 2 found it distracting and 6 found it annoying. In the case of the r-agent, 15 participants replied: 13 did not find it annoying, 2 found it distracting and only 1 of them found it annoying.

Participants were asked about the final mistake of the agent. Overall 23 participants pointed out the mistake to the agent ("*But this box says Trash*"), 10 for the r-agent and 13 for the v-agent. Eight participants put the paper in the indicated box without questioning the agent (5 r-agent, 3 v-agent). Unlike the response to interruptions, there were no significant differences. One person for each embodiment mentioned it was funny, some participants mentioned they trusted the agent (six participants for the v-agent and 7 for the r-agent): "*He knows better"*.

Some participants justified the agent's mistake, blaming themselves or others (5 participants for the v-agent and 3 for the r-agent), e.g.: "*somebody had put the wrong label to the box*", "*somebody might have changed the location of the mail*", "*Paul might call his mailbox Trash*". Two of those who found it annoying (one for each embodiment) saw a malicious intention: "*It lied to me to piss me off*".

The use of human characteristics of the participants when describing their interaction was also similar in the two embodiments: "*She was absorbed with the videos*", "*The voice sounded sincere*". Participants were specifically asked whether they changed the way they talked to the agent compared to how they talk to people. Of 19 interacting with the v-agent, 4 said they did not change their way of talking and 13 said they did. The most mentioned changes were that they spoke more clearly (4 participants), more slowly (4 participants) and with simpler sentences (4 participants). Only 7 participants interacting with the r-agent reported changing their speech.

5 Discussion

As reported above participants expressed a clear preference for the robotic embodiment when presented with the 2 different choices (v-agent and r-agent). It is at first sight curious that this clear preference for a physical embodiment is not reflected in any of the analysed questionnaire data. One possible explanation is that participants compared the agent with their initial expectation or mental model of it. About half of the participants who expressed their initial impressions commented they had been surprised by the agent, the scenario or the interaction; they didn't expect what they found. Many of them commented on the agent surpassing their initial expectations. This might have been reflected in the questionnaire results, very positive in general and independent of the embodiment.

However, when confronted with the idea of an alternative agent, they were able to picture what the interaction would have looked like with the new agent, compared to the experience of the agent they had interacted with. We can only speculate whether this preference for the robot would have held up if they could have actually seen and interacted with the alternative embodiment. It is possible that the fact that they did not actually see or experience the alternative embodiment biased them towards the robot, their imagination imbuing robots with more attractiveness compared to virtual agents.

Another possible explanation for this disparity is that the participants' strongest impressions were influenced by the immersive task and experience of having a dialogue with an artificial intelligence. Since neither the task nor the dialogue differed between embodiments, these factors might have simply overshadowed the effect of embodiment on the participants.

Significantly less annoyance was caused by the r-agent's interruptions compared to those of the v-agent, as expressed in the interviews. This corresponds with our hypothesis that users would tolerate more inappropriate behaviour by the r-agent. However, this is once again in contrast with our questionnaire results, which did not show significant differences across embodiments in the measures related to the interruptions (friendly and disturbing). It is possible that these adjectives were not ideal candidates in order to gauge participants' tolerance. After all, one can be friendly but still annoying and disturbing could have been interpreted in its alternative meaning of weird / creepy.

There is a significant difference in the number of participants who changed their speech when interacting with the agent, compared to talking to a human. Seven participants who interacted with the r-agent changed it, compared to 13 in the case of the v-agent. This evidence supports our hypothesis that the embodiment influences the user's mental model of the communicative capabilities of the agent. Moreover, as in [10], people tailored the way they communicate with the agent to account for performance of the agent so far. This is just the case for the r-agent group.

6 Conclusion

Previous studies in embodied agents have shown an overall preference for physical robots over virtual agents. In this paper, we have further investigated the effect of the embodiment in shaping the user's behaviour and attitude towards the agent. Taking this objective and previous work into account, a scenario and agent behaviours were designed and built. Social talk, averted attention, and an occasional mistake were included in the behaviour of the agent.

We chose a Between-Group experiment design. Results from the interview analysis show a clear preference for interacting with the physically embodied rather than on-screen agent if offered that possibility. Subjects also talked to the robot more naturally, closer to the way they would talk to a human being. Together with the physical and social presence being one of the most mentioned advantages of the robotic embodiment, the findings indicate that the physical existence of the agent makes for more natural interaction.

Otherwise, however, the differences found between the embodiments were relatively minor with the overwhelming number of factors analysed not differing across conditions. We suggest this is due to the participants' strong involvement in a task context. When their interaction with the companion has a reason and is not merely serving the satisfaction of curiosity about technology, embodiment factors seem to move to the background. Only when given time to reflect and taken out of the task context as in our post interviews embodiment differences come to the foreground. Based on this hypothesis we would expect to see much stronger differences in a similar experiment that is stripped of the task. For our work on migration this implies that it is important for a migrating companion to engage users in tasks in order to downplay embodiment differences and appear as a consistent identity in different bodies.

Acknowledgements. This work was partially supported by the European Commission (EC) and is currently funded by the EU FP7 ICT-215554 project LIREC (Living with Robots and Interactive Companions). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

References

1. Austermann, A., Yamada, S., Funakoshi, K., Nakano, K.: Similarities and Differences in Users' Interaction with a Humanoid and a Pet Robot. In: 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 73–74 (2010)

2. Bainbridge, W., Hart, J., Kim, E., Scassellati, B.: The effect of presence on human-robot interaction. In: The 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2008, Munich, Germany, pp. 701–706 (2008)
3. Breazeal, C., Scassellati., B.: How to build robots that make friends and influence people. In: Proc. IROS 1999, Kyongju, Korea, pp. 858–863 (1999)
4. Cassell, J.: Embodied Conversational Agents: Representation and Intelligence in User Interfaces. AI Magazine 22(4) (2001)
5. Damasio, A.: The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Houghton Mifflin Harcourt, New York (1999)
6. Gomes, P.F., Márquez Segura, E., Cramer, H., Paiva, T., Paiva, A., Holmquist, L.E.: ViPleo and PhyPleo: Artificial pet with two embodiments. In: Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology, ACE 2011, Lisbon, Portugal (2011)
7. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the Big-Five personality domains. Journal of Research in Personality 37, 504–528 (2003)
8. Goetz, J., Kiesler, S., Powers, A.: Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: Proc. ROMAN 2003, pp. 55–60 (2003)
9. Hall, E.T.: The hidden Dimension. Garden City, N.Y (1966)
10. Kim, E.S., Leyzberg, D., Tsui, K.M., Scassellati, B.: How People Talk When Teaching a Robot. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, HRI 2009, San Diego, USA, pp. 23–30 (2009)
11. Komatsu, T., Abe, Y.: Comparing an On-Screen Agent with a Robotic Agent in Non-Face-to-Face Interactions. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 498–504. Springer, Heidelberg (2008)
12. Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D.S., Nehaniv, C.L.: Effects of Embodiment and Gestures on Social Interaction in Drumming Games with a Humanoid Robot. Advanced Robotics 23(14), 1951–1996 (2009)
13. 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)
14. Kriz, S., Anderson, G., Trafton, J.G.: Robot-Directed Speech: Using Language to Assess First-Time users' conceptualizations of a robot. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2010, Osaka, Japan, pp. 267–274 (2010)
15. Lee, K., Jung, Y., Kim, J., Kim, S.: Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. Int. J. Human-Computer Studies 64(10), 962–973 (2006)
16. Lohse, M., Hegel, F., Swadzba, A., Rohlfing, K., Wachsmuth, S., Wrede, B.: What can I do for you? Appearance and application of robots. In: Proceedings of The Reign of Catz and Dogz? Symposium at AISB 2007 (2007)
17. Mori, M.: Bukimi no tani - The uncanny valley (K. F. MacDorman & T. Minato, Trans.). Energy 7(4), 33–35 (1970)
18. 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: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, HRI 2009, San Diego, USA, pp. 61–68 (2009)

19. Nomura, T., Suzuki, T., Kanda, T., Kato, K.: Measurement of Negative Attitudes toward Robots. Interaction Studies 7(3), 437–454 (2006)
20. Pereira, A., Martinho, C., Leite, I., Paiva, A.: iCat, The Chess Player: The Influence of Embodiment in the Enjoyment of a Game. In: Proceedings of the 7th International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2008, Estoril, Portugal, pp. 1253–1256 (2008)
21. Powers, A., Kiesler, S., Fussell, S., Torrey, C.: Comparing a Computer Agent with a Humanoid Robot. In: Proceedings of the ACM/IEEE International Conference on Human-robot Interaction, HRI 2007, Washington DC, USA, pp. 145–152 (2007)
22. Riek, L.D., Rabinowitch, T., Bremner, P., Pipe, A.G., Fraser, M., Robinson, P.: Cooperative Gestures: Effective Signaling for Humanoid Robots. In: Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction, HRI 2010, Osaka, Japan, pp. 61–68 (2010)
23. Steels, L., Brooks, R.: The artificial life route to artificial intelligence: Building situated embodied agents. Lawrence Erlbaum Associates, New Haven (1993)
24. Syrdal, D.S., Dautenhahn, K., Koay, K.L., Walters, M.L.: The Negative Attitudes towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study. In: Proc. New Frontiers in Human-Robot Interaction, AISB 2009 Convention (2009)
25. Thomaz, A.L., Cakmak, M.: Learning about Objects with Human Teachers. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, HRI 2009, San Diego, USA, pp. 15–22 (2009)
26. Wainer, J., Feil-seifer, D.J., Shell, D.A., Matarić, M.J.: Embodiment and human-robot interaction: A taskbased perspective. In: Proceedings of the 16th IEEE International Conference on Robot & Human Interactive Communication, RO-MAN 2007, Jeju, South Korea, pp. 872–877 (2007)

A Second Chance to Make a First Impression? How Appearance and Nonverbal Behavior Affect Perceived Warmth and Competence of Virtual Agents over Time

Kirsten Bergmann[1,2], Friederike Eyssel[1], and Stefan Kopp[1,2]

[1] Center of Excellence in "Cognitive Interaction Technology" (CITEC), Bielefeld University
[2] Collaborative Research Center 673 "Alignment in Communication", Bielefeld University
P.O. Box 10 01 31, 33501 Bielefeld, Germany
{kbergman,skopp}@techfak.uni-bielefeld.de,
friederike.eyssel@uni-bielefeld.de

Abstract. First impressions of others are fundamental for the further development of a relationship and are thus of major importance for the design of virtual agents, too. We addressed the question whether there is a second chance for first impressions with regard to the major dimensions of social cognition–warmth and competence. We employed a novel experimental set-up that combined agent appearance (robot-like vs. human-like) and agent behavior (gestures present vs. absent) of virtual agents as between-subject factors with a repeated measures design. Results indicate that ratings of warmth depend on interaction effects of time and agent appearance, while evaluations of competence seem to depend on the interaction of time and nonverbal behavior. Implications of these results for basic and applied research on intelligent virtual agents will be discussed .

Keywords: Evaluation, agent appearance, nonverbal behavior, warmth, competence.

1 Introduction

One of the major challenges in current research on intelligent virtual agents (IVAs) is the question of how to elicit positive affect, user acceptance and even long-term relationships between users and agents. Inspired by human-human interaction it seems to be of striking importance that a user's *first impression* of an IVA is positive: When two people meet for the first time, they immediately form initial ideas from each other. These early impressions have a major impact on how their relation further develops. People's behavior towards others is shaped depending on differences in first impressions such that people who have favorable impressions of someone tend to interact more with that person than others having unfavorable impressions [18]. First impressions are, therefore, an important basis for whether humans will build rich relations with others.

The tendency to form first impressions is absolutely fundamental with regard to person perception and social cognition [28]. Thus, the two universal dimensions of social

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 126–138, 2012.

cognition–*warmth* and *competence* [12,10]–are notably shaped by the first moments of contact, whereby the warmth dimension captures whether people are friendly and well-intentioned and the competence dimension captures whether people have the ability to deliver on those intentions. A number of studies have shown that warmth and competence assessments determine whether and how we intend to interact with others (cf. [10]): We seek the company of people who are assumed to be warm and and avoid those who appear less sociable (i.e. cold). With regard to competence, we prefer to cooperate with people we judge as competent, while incompetent people are disregarded. With regard to the relation of warmth and competence, Fiske et al. [12, p. 77] state that "warmth judgments are primary: warmth is judged before competence, and warmth judgments carry more weight in affective and behavioral reactions". In addition, "perceived warmth is more easily lost and harder to regain compared to perceived competence" [10, p. 17].

So, as early evaluations of others are a major concern for social evaluations–which cues do people take into account when making first impressions? How can we strengthen the chances for making a solid first impression? Empirical evidence from social psychology has demonstrated that initial impressions are formed rather quickly on the basis of minimal information with visual *appearance* and *nonverbal behavior* providing the major cues [25]. With regard to the latter, *co-speech gestures* are a particular kind of nonverbal behavior: Gestures convey semantic information, while the main function of other nonverbal behaviors like facial expressions or body posture is to communicate sympathy/antipathy and other affective/emotional signals. Accordingly, gestures are especially important in settings and tasks in which information is to be transferred and we will focus on this kind of nonverbal behavior here. Nevertheless, despite of all evidence for the fact that first impressions are lasting impressions in terms of setting the tone for a relationship, they do not define its boundaries or potential. Subsequent encounters still have the chance to modify real quality of any relationship. Especially competence traits are still subject to subsequent modifications (cf. [10]). So it seems that there is still a chance for what we call a 'second impression'.

All these issues of first and second impressions are, however, mostly unexplored for virtual agents. Although there is empirical evidence from IVA research that provides support for the fact that visual appearance and nonverbal behavior are cues of major importance for human's evaluation of virtual agents (cf. [21]), the relation of (1) agent-related cues like agent appearance and nonverbal behavior, (2) social evaluation in terms of the major dimensions of warmth and competence as well as (3) dynamic modifications of impressions has not been investigated for IVAs, yet. In this paper, we aim to investigate interaction effects of these variables. In particular, we address the question how warmth and competence ratings change from a first impression after a few seconds to a second impression after a longer period of human-agent interaction depending on manipulations of agents' visual appearance and nonverbal behavior. The following section gives an overview of related work and background literature. Section 3 describes the setting and procedure of the evaluation study. Results are presented in Section 4. Finally, we discuss the results and draw conclusions in Section 5.

2 Related Work and Background

2.1 Effects of Agent Appearance

A growing body of empirical evidence suggests that the appearance of virtual and robotic agents has an extensive influence on how humans evaluate them, as Goetz et al. [14, p. 1] put it aptly: “The book is judged by its cover. [...] A robot’s appearance and behavior provide cues that influence perceptions of the robot’s propensities, and assumptions about its capabilities”. A major dimension of agent appearance which is subject to a large amount of recent research activities (particularly in the field of robotics) is the degree of *human-likeness*. Among researchers it is a controversial issue whether machines should be endowed with a human-like interface or not. On the one hand, it is argued that humanoids provide a more intuitive interface because rules of human interaction can easily be transferred [7]. On the other hand there are also opponents who argue that a human-like appearance results in unrealistic expectations or even fear [11]. In a few studies from robotics it was examined how human-like vs. machine- or robot-like agents are evaluated by human users, but none of them directly investigated the variables of warmth and competence. Hinds et al. [16] found that machine-like interfaces tend to be treated less politely and less socially interactive than human-like interfaces in a joint task between humans and robots. Moreover, expectations regarding abilities and reliability were lower for the machine-like interfaces. Nishio & Ishiguro [24] found a strong effect of appearance on human evaluations: They reported that interlocutors tend to hold different impressions of robotic agents of different appearances, even if the agents were tele-operated by a single person. Goetz et al. [14] presented evidence for a task-dependent relation between a robot’s appearance and users’ acceptance of and cooperation with a robot. In fact, participants systematically preferred robots for jobs when the robot’s human-likeness matched the sociability and seriousness required in those tasks. Woods et al. [30] further provided evidence that children judge human-like robots as aggressive, but machine-like robots as friendly.

Often discussed with regard to effects of appearance and behavior is Mori’s ‘uncanny valley’ hypothesis [22] which states that the perceived familiarity of robots rises with increasing anthropomorphism until a point is reached beyond which ratings go into reverse and robots are perceived as eerie instead of familiar. According to Mori the uncanny valley effect is even stronger for animated and moving agents than it is for static agents. Although originally proposed for robotics this hypothesis has also been transferred to virtual agents where a couple of studies provided evidence for the effect in IVAs as well [13,2]. It turned out that the degree of realism does not necessarily result in positive evaluations. Instead, it is more important that the degree of realism is consistent with the agent’s behavior.

In another line of research it is investigated how virtual agents are perceived depending on manipulations of human-likeness in contrast to zoomorphic agents. Sträfling et al. [29] compared a cartoon-like rabbit and a realistic anthropomorphic agent in a teaching scenario. Results showed that the appearance of the agent mattered such that the rabbit-like agent was preferred. By contrast, Bailenson et al. [1] did not find significant differences between a human-like and a rabbit-like appearance with regard to perceived likeability, but both interfaces received significantly higher ratings than a third abstract IVA variant. In a different study Buisine & Martin [8] compared three different cartoon-like agents (two male, one female) and reported evidence for an influence of agents' appearance on likeability: one of the male agents was signicantly preferred over the others. It was, however, unclear by which aspect(s) of appearance this effect was caused. An analysis of participants' qualitative comments suggests that this effect was due to the wide smile of that particular character.

Overall, results of studies investigating how virtual agents are perceived depending on their visual appearance are inconclusive (see [21] for a more comprehensive overview) which may be due to the fact that appearance consists of many different aspects and cannot be broken down to a single variable. In addition, as some researchers pointed out, agent appearance should be consistent with the agent's behavior [13,2] and also with the task [14]. Accordingly, it makes sense to investigate the effect of agent appearance in interaction with other variables. Nevertheless, the conclusion can be drawn that agent appearance has the potential to modify ratings of warmth, likeability etc.

2.2 Effects of Nonverbal Behavior

Research investigating the impact of IVA's nonverbal behavior on human perception of warmth and competence is only sparse. So far, only Niewiadomski et al. [23] considered the role of verbal vs. nonverbal vs. multimodal emotional displays on warmth, competence and believability, whereby nonverbal behavior consisted of facial expressions accompanied by emotional gestures. It was found that all dependent measures (warmth, competence, and believability) increased with the number of modalities used by the agent. From other studies, although often not directly addressing the dimensions of warmth and competence, there is further evidence that endowing virtual agents with human-like, nonverbal behavior may lead to enhancements of the likeability of the agent, trust in the agent, satisfaction with the interaction, naturalness of interaction, ease of use, and efficiency of task completion [6,15].

With regard to the particular effects of co-speech gestures, Krämer and colleagues [20] found no effect on agent perception when comparing a gesturing agent with a non-gesturing one. The agent displaying gestures was perceived just as likeable, competent, and relaxed as the agent that did not produce gestures. In contrast, a number of other studies found beneficial effects of virtual agents using gestures. Cassell & Thorisson [9] reported that nonverbal behavior (including beat gestures) resulted in an increase of perceived language ability and life-likeness of the agent, as well as smoothness of interaction. A study by Rehm & André [26] investigated whether a gesturing agent would change the perceived politeness tone compared to that of the textual utterances and whether the subjective rating is influenced by the type of gestures (abstract vs. concrete). These studies revealed that the perception of politeness depends on the graphical quality of the employed gestures. In cases where iconic gestures were rated as being of higher quality than metaphoric gestures, a positive effect on the perception of the agent's willingness to co-operate was observed. In cases where iconic gestures were rated as being of lower quality than metaphoric gestures, a negative effect on the perception of the agent's willingness to co-operate was observed. Moreover, in the aforementioned study by Buisine & Martin [8], effects of different types of speech-gesture cooperation in an agent's behavior were found: redundant gestures increased ratings of explanation quality, expressiveness of the agent, likeability and positive perception of the agent's personality. Finally, in our own previous research we compared different kinds of gestural behavior in an IVA (individualized gesture models vs. common gesture models vs. control conditions) [5]. Especially gestural behavior generated with individualized models turned out to have significant benefits on human ratings in terms of likeability, competence and human-likeness.

In sum, previous research has shown that IVAs using gestures provide the potential to increase human judgements of likeability, competence and other variables. However, gesture use alone is not a guarantee for positive ratings. These are rather depending on further factors like gesture quality [26] or adequate gesture models.

2.3 Changing Impressions

Regarding the question whether user evaluations of robotic or virtual agents might change as a function over time research is very sparse. There is only work by Komatsu & Yamada [19] who investigated what they called 'adaptation gap'–the difference between users' expectations before starting their interactions with a robot and their evaluation after interacting. Participants were divided into two groups which received different information about the interface beforehand (lower expectation group vs. higher expectation group). The study revealed a significant difference between experimental conditions with regard to participants' post-interaction judgements.

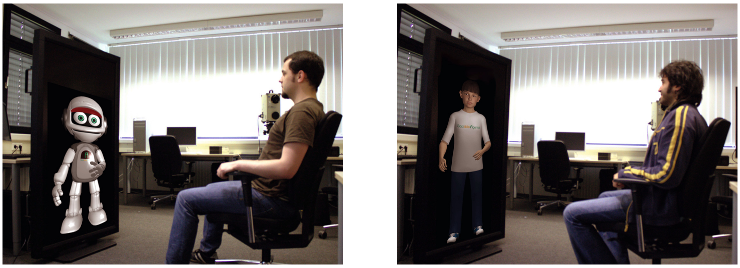
Thus, there is initial evidence that human users' judgements of agents might modify with ongoing interaction. There is, however, no detailed account of interaction effects between dynamics of evaluation and other variables such as agents' visual appearance and behavior. Our study presented in this paper aims to be a first step into the direction of closing this gap.

3 Study

The study employed a 2 x 2 between-subject factorial design with repeated measures to investigate the effects of agent appearance (robot-like vs. human-like) and angent behavior (gestures absent vs. gestures present) on the perception of virtual agents by human participants.

Participants. A total of 80 participants (20 in each condition), aged from 19 to 48 years (M = 25.11, SD = 5.45), took part in the study. 51 participants were female and 29 were male. All of them were recruited at Bielefeld University and received 3 Euros for participating.

Procedure. The experiment was conducted in two consecutive phases. In the first phase, participants were provided with a short introduction by the virtual agent which took approximately 15 seconds: "Hello, my name is Billie/Vince. In a moment you will have the have the possibility to getting to know me closer. But first you will be provided with a questionnaire concerning your first impression of myself". Subsequently, participants were asked to state their first impression regarding their perception of the virtual agent's personality in a questionnaire (*T1*, took approximately five minutes). In the second phase, the virtual agent described a building with six sentences which took approximately 45 seconds. Each sentence was followed by a pause of three seconds. Participants were instructed to carefully watch the presentation given by the virtual agent in order to be able to answer questions regarding content and subjective evaluation of the presentation afterwards. Immediately upon receiving the descriptions by the IVA, participants filled out a second questionnaire (*T2*) stating their perception of the virtual agent's personality at this point of time.



(a) Robot-like IVA 'Vince' (b) Human-like IVA 'Billie'

Fig. 1. Set-up of the study in which the IVAs provided users with explanations about a building

Figure 1 shows the setup used for stimulus presentation: The virtual agent was displayed on a 80 x 143 cm screen. Participants were seated 170 cm away from the screen and their heads were approximately leveled with the virtual agent's head. Participants have been left alone for the stimulus presentations, and after receiving the questionnaires to complete them, i.e., neither the experimenter nor the virtual agent were present.

Independent Variables. As a within-subject variable repeated measurements were taken at two *points of measurement* (T1, T2) in the experimental procedure. T1 was chosen to measure how users rate the agent in terms of a first impression after 15 seconds, while T2 was chosen to evaluate user ratings after a longer time of information presentation by the IVA (second impression). In addition, participants were randomly assigned to one out of four conditions which resulted from the manipulation of two independent between-subject variables: *agent appearance* and *agent behavior* (gesture use).

Agent appearance. Two different kinds of virtual agents were employed in the study with regard to the dimension of human-likeness (anthropomorphic–robotic IVAs). We employed the IVA 'Vince' as a robot-like character and 'Billie' as a human-like agent (see Figure 1). To hold conditions as constant as possible despite the intended manipulations, both agents were displayed with the same overall size, and no other nonverbal behaviors than gesture use (see below) were employed. Verbal utterances were also identical across conditions, although the two agents used different synthetic voices (both Mary TTS [27]) each of which matched with the agent's visual appearance. A child-like voice was used for Billie which is based on a female German Hidden semi-Markov model voice. It was modified by a slightly shortened vocal tract, F0 values were shifted by 50 Hz resulting in a higher-pitched voice, the range of f0 values was expanded and duration scaling was modified such that the synthesized speech output was slower (children speaking slower than adults). A machine-like voice for Vince was based on a male German Hidden semi-Markov model voice. This voice was also modified by a slightly shortened vocal tract, F0 values were shifted by 100 Hz and the range of f0 values was expanded.

Agent behavior. The gesturing behavior of the virtual agents was manipulated in the way that either no gestures were used at all ('gestures absent'), or the *Generation Network for Iconic Gestures* (GNetIc, for details, see [3]) was employed to generate gestures for the virtual agents ('gestures present'). The GNetIc model serves to simulate co-speech gesture use in the style of individual speakers accounting for obvious and important inter-individual differences in gestural behavior. For the purpose of the current study, we employed an individual speaker's GNetIc model which has been shown to increase the perceived quality of an object description given by a virtual human, and also resulted in a more positive rating of an IVA in terms of likeability, competence and human-likeness, compared to other GNetIc models as well as control conditions [5]. All descriptions given by the virtual agents were produced fully autonomous at runtime by using an integrated speech and gesture production architecture [4].

Dependent Variables. Participants' responses regarding the dependent variables were collected using 18 items [12,17] such as 'pleasant', 'friendly', 'helpful' which had to be assessed on seven-point Likert scales how well they apply to Billie/Vince (not appropriate–very appropriate). For subsequent data analyses, average scores were computed as indices for the scales of warmth and competence, whereby remaining ones were used as filler items.

WARMTH. The following eight items measured warmth-related traits: pleasant, sensitive, friendly, helpful, affable, likeable, approachable, sociable. The scale was highly reliable with Cronbach's α = .92 for T1 and α = .92 for T2.

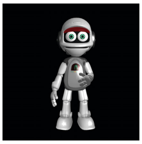
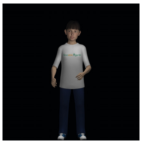
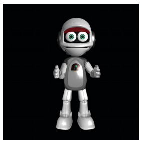
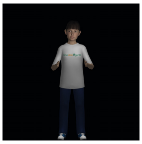
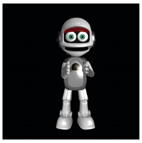
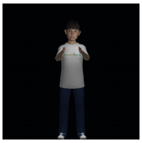
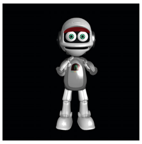
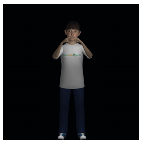
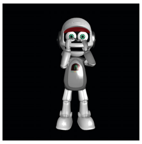
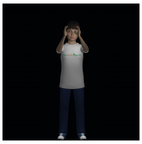
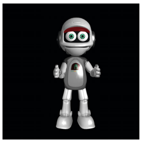
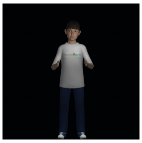
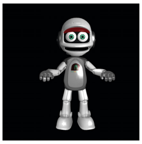
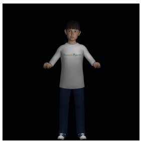
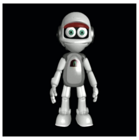
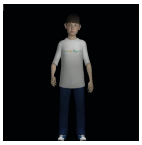
COMPETENCE. Competence-related traits were measured with four items (intelligent, organized, expert, thorough). The competence scale also showed high reliabilities of Cronbach's α = .84 for T1 and α = .83 for T2.

Data Analysis. First, descriptive data analyses were conducted to summarize the data. Second, a mixed-model design ANOVA was conducted to analyze the effect of the within-subject variable *point of measurement* (T1 vs. T2) and between-subject variables *agent appearance* (robot-like vs. human-like) and *agent behavior* (gestures absent vs. gestures present) on the dependent variables WARMTH and COMPETENCE.

4 Results

The descriptive data analysis revealed means and standard deviations as summarized in Table 2. Concerning WARMTH, the robot-like agent with gestures present had the highest mean score for both points of measurement. However, lowest WARMTH-ratings

Table 1. Stimuli presented in the ‘gestures present’ conditions: verbal description given in each condition (left column; translated to English; gesture positions labelled with squared brackets) and the different virtual agents either displaying gestural behavior

Agent appearance	***robot-like***	***human-like***
Introduction		
[The church is squared]...		
...and in the middle there is [a small spire.]		
The spire has [a tapered roof].		
And the spire has [a clock].		
There is [a door] in front.		
And in front of the church there is [a low, green hedge].		
There is [a large deciduous tree] to the right of the church.		

at measurement point T2 were observed for the robot-like agent as well, namely in the condition with gestures absent. For T1 the human-like agent with gestures received lowest WARMTH-ratings. With regard to competence the highest mean scores at both points of measurement were reached by the robot-like agent with gestures present as well as the human-like agent with gestures absent.

In the following we report (interaction) effects of between-subject and within-subject variables with regard to the two dependent measures of WARMTH and COMPETENCE. Only significant effects are reported at a level of $p < .05$.

Table 2. Means and standard deviations of WARMTH and COMPETENCE as a function of *agent appearance*, *agent behavior* and *point of measurement*

		Robot-like virtual agent		Human-like virtual agent	
		Gestures absent M (SD)	Gestures present M (SD)	Gestures absent M (SD)	Gestures present M (SD)
Warmth	T1	4.56 (1.09)	5.15 (1.21)	4.25 (1.14)	4.05 (1.18)
	T2	3.90 (1.32)	4.49 (1.36)	4.12 (1.09)	4.19 (1.18)
Competence	T1	4.50 (0.93)	4.75 (1.35)	4.85 (1.48)	3.98 (1.21)
	T2	4.38 (1.26)	4.93 (1.16)	4.79 (1.42)	4.55 (1.11)

WARMTH There was a significant main effect of the within-subject variable *point of measurement* on the degree of WARMTH perceived by participants ($F_{(1,76)} = 11.65$, $p = .001$) in the way that at measuring point T1 the agents were perceived as warmer than at measuring point T2.

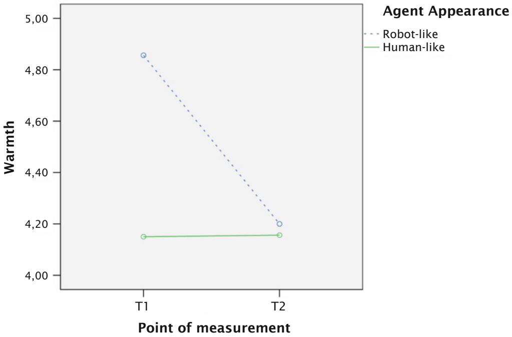


Fig. 2. WARMTH ratings as a function of *agent appearance* and *point of measurement*

Concerning interaction of independent variables, there was a significant interaction effect of *point of measurement* and *agent appearance* ($F_{(1,76)} = 12.10$, $p = .001$) indicating that the rating of WARMTH at the two points of measurement differed significantly. As visualized in Figure 3 ratings of WARMTH decreased for the robot-like agent between the two points of measurement, while ratings remained constant for the human-like agent. While participants rated the robot-like agent as being warmer than the human-like agent at measuring point T1, ratings of warmth did not differ between robot-like and and human-like agents at T2.

COMPETENCE. Regarding perceived COMPETENCE of the virtual agents, there was a significant effect for the interaction of the *point of measurement* and *agent behavior* ($F_{(1,76)} = 4.56$, $p = .04$). While gesture use was found to result in an increase of perceived competence of the virtual agents between the two points of measurement, ratings of agent competence slightly decreased when the agents did not use any gestures (see Figure 3).

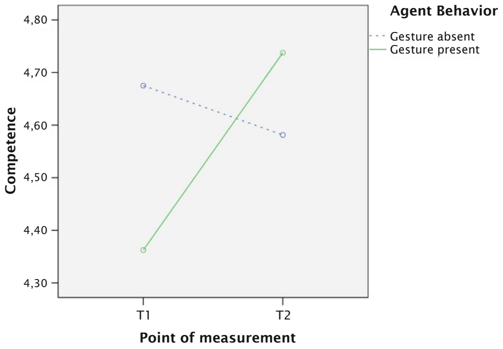


Fig. 3. COMPETENCE ratings as a function of *agent behavior* and *point of measurement*

5 Conclusion

The goal of this paper was to explore first and second impressions human users form of virtual agents depending on agent appearance and nonverbal (gestural) behavior. To investigate this objective, we conducted an experiment with a mixed design. Our results speak to three important issues. First of all, participants' rating with regard to the agents' warmth was sensitive to the point of measurement for the robot-like agent. According to participants' first impression the agents were perceived as warmer than according to second impressions after a longer phase of information presentation by the IVAs. By contrast, there were no significant effects of the point of measurement observed for competence. Second, with regard to warmth, there was a significant interaction effect of the point of measurement and agent appearance. While warmth-ratings between the two points of measurement decreased significantly for the robot-like agent, ratings remained constant for the human-like agent. Third, regarding competence, we found a significant interaction effect of the point of measurement and agent behavior. That is, the way that gesture use helped to increase competence ratings between measurement points, while the absence of gestures resulted in a decrease of competence ratings.

Compared to related work from robotics and virtual agents, our results provide further support for the fact that human users' evaluation are actually sensitive regarding the agents' appearance and nonverbal agent behavior. Specifically, our findings are generally in line with previous results from IVA research by Niewiadomski et al. [23] who showed that ratings of agents' warmth and competence increase with the number of modalities used by the agent. The particular setup of our study, however, allows for a more sophisticated view showing that the two dimensions of social cognition are not dependent on the agents' behavior (and agent appearance) in equal measures. Furthermore, going beyond previous work, we showed that IVA ratings are sensitive to the point of measurement.

How can we explain these findings? The overall decrease of warmth and increase of competence (depending on presence/absence of co-speech gestures) is in accordance with evidence from social psychology ([12,10], see Sect. 1), stating that warmth judgements are made rather quickly and that warmth is easily lost and hard to regain in contrast to competence. However, the interaction effect of the point of measurement and agent appearance is notable. Why do warmth rating decrease only for the robot-like agent? We suppose that this is a matter of expectations. The robot-like agent Vince with his large head and big eyes fulfills quite some characteristics of a child-like look which is typically associated with many positive attributes. In addition, in the graphical representation of Vince appears to be much closer and present than Billie. It seems, however, that being exposed to this interface for a longer time results in a decrease of this effect. We suggest, that the particular setting we employed in this study might have had an impact, too. In a different task, e.g. enforcing more interaction or personal communication, warmth ratings are likely to be different than in our setting of information providing by the IVA.

What can we learn from these findings for the design of (interactions with) virtual agents? First of all, our results clearly show that there is a second chance to make first impressions. However, with regard to warmth, overall ratings decreased between points of measurement. Interestingly, this was only due to a decrease of perceived warmth for the robot-like agent. For the human-like agent ratings remained constant. We can thus conclude to prefer human-like agents as they provide the potential for stable impressions of warmth. Moreover, with regard to competence, employing virtual agents with gestures helps to increase participants' ratings–independent of the agents' appearance. So we can advise to endow virtual agents with gestural behavior to improve their perceived competence. Notably, it should be kept in mind that the gestures employed here were generated from an individualized gesture model which has been rated positively in a previous evaluation study [5]. So there is still the challenge to choose an adequate gesture profile to generate gestures from. Nevertheless, our findings are in line with results from a previous study in which a different virtual agent was employed.

We started with the question whether there is a second chance for first impressions and presented a study that was novel in the way that it combined agent appearance and agent behavior of IVAs as between-subject factors with a repeated measures design to investigate warmth and competence evaluations by human raters. Overall, we aimed to control the experimental setting as much as possible. Nevertheless, one should be aware of the fact that our results are obtained from employing specific IVAs in a specific setting and domain of application (task-related monologue, visual descriptions). The current research points to important issues that need to be studied in future research. These should, one the one hand, isolate further variables like the agents' voice or other nonverbal behaviors, and on the other hand, broaden the scope by investigating different tasks or settings like 'real' interactive dialogues or long-term development. Nevertheless, we are confident that our findings provide an important step for further IVA research in at least two ways. First, we showed that timing actually matters. Results clearly indicate participants' impressions of IVAs cahnge over time providing interesting and detailed information about how we can improve our virtual agents. Second, our methodology of examining specific variables (i.e. agent appearance and agent behavior) allowed for, instead of taking the agent with all its characteristics as a whole, has been shown to be an adequate means to gain detailed insight into the way IVAs are judged by humans.

Acknowledgements. This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence 277 in "Cognitive Interaction Technology" (CITEC). The authors further acknowledge Francoise Therese Rasoloarinala for collecting the data analyzed in this paper.

References

1. Bailenson, J.N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., Blascovich, J.: The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. Presence 14(4), 379–393 (2005)
2. Bente, G., Krämer, N.C., Jan Peter de Ruiter, A.P.: Computer animated movement and person perception: Methodological advances in nonverbal behavior research. Journal of Nonverbal Behavior 25(3), 151–166 (2001)
3. 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)
4. Bergmann, K., Kopp, S.: Increasing expressiveness for virtual agents–Autonomous generation of speech and gesture in spatial description tasks. In: Decker, K., Sichman, J., Sierra, C., Castelfranchi, C. (eds.) Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, Budapest, Hungary, pp. 361–368 (2009)
5. 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)
6. Bickmore, T., Cassell, J.: Social dialogue with embodied conversational agents. In: van Kuppevelt, J., Dybkjaer, L., Bernsen, N. (eds.) Advances in Natural, Multimodal Dialogue Systems. Kluwer Academic Publishers, Dordrecht (2005)
7. Brooks, R.: Humanoid robots. Communications of the ACM 45(3), 33–38 (2002)
8. Buisine, S., Martin, J.C.: The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. Interacting with Computers 19, 484–493 (2007)
9. Cassell, J., Thórisson, K.: The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. Applied Artificial Intelligence 13, 519–538 (1999)
10. Cuddy, A.J., Glick, P., Beninger, A.: The dynamics of warmth and competence judgments, and their outcomes in organizations. Research in Organizational Behavior 31, 73–98 (2011)

11. Dautenhahn, K.: Robots as social actors: Aurora and the case of autism. In: Proceedings of the Third International Cognitive Technology Conference (1999)
12. Fiske, S.T., Cuddy, A.J., Glick, P.: Universal dimensions of social cognition: Warmth and competence. Trends in Cognitive Science 11(2), 77–83 (2006)
13. Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., Sasse, M.A.: The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In: Proceedings of ACM Conference on Human Factors in Computing Systems, pp. 529–536 (2003)
14. Goetz, J., Kiesler, S., Powers, A.: Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: Proceedingsof the 12th IEEE Workshop om Robot and Human Interactive Communication (2003)
15. Heylen, D., van Es, I., Nijholt, A., van Dijk, B.: Experimenting with the gaze of a conversational agent. In: Proceedings International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, pp. 93–100 (2002)
16. Hinds, P.J., Roberts, T.L., Jones, H.: Whose job is it anyway? a study of human-robot interaction in a collaborative task. Human Computer Interaction 19(1), 151–181 (2004)
17. Hoffmann, L., Krämer, N.C., Lam-chi, A., Kopp, S.: Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 159–165. Springer, Heidelberg (2009)
18. Kelley, H.: The warm-cold variable in first impressions of persons. Journal of Personality 18, 431–439 (1950)
19. Komatsu, T., Kurosawa, R., Yamada, S.: Difference between users' expectations and perceptions about a robotic agent (adaptation gap) affect their behaviors. In: Proceedings of the HRI 2011 Workshop Expectations in Intuitive Human-robot Interaction (2011)
20. Krämer, N.C., Tietz, B., Bente, G.: Effects of Embodied Interface Agents and Their Gestural Activity. In: Rist, T., Aylett, R.S., Ballin, D., Rickel, J. (eds.) IVA 2003. LNCS (LNAI), vol. 2792, pp. 292–300. Springer, Heidelberg (2003)
21. Krämer, N.C.: Soziale Wirkungen virtueller Helfer–Gestaltung und Evaluation von Mensch-Computer Interaktion. Verlag W. Kohlhammer, Stuttgart (2008)
22. Mori, M.: The Buddha in the robot. Charles E. Tuttle Co., Tokyo (1982)
23. Niewiadomski, R., Demeure, V., Pelachaud, C.: Warmth, competence, believability and virtual agents. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 272–285. Springer, Heidelberg (2010)
24. Nishio, S., Ishiguro, H.: Attitude change induced by different appearances of interaction agents. International Journal of Machine Consciousness 3(1), 115–126 (2011)
25. Peplau, L., Taylor, S., Sears, D.: Social Psychology. Prentice Hall (2005)
26. Rehm, M., André, E.: Informing the design of agents by corpus analysis. In: Nishida, T., Nakano, Y. (eds.) Conversational Informatics. John Wiley & Sons, Chichester (2007)
27. 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)
28. Skowronski, J., Amady, N. (eds.): First Impressions. The Guilford Press (2008)
29. Sträfling, N., Fleischer, I., Polzer, C., Leutner, D., Krämer, N.C.: Teaching learning strategies with a pedagogical agent: The effects of a virtual tutor and its appearance on learning and motivation. Journal of Media Psychology: Theories, Methods, and Applications 22(2), 73–83 (2010)
30. Woods, S., Dautenhahn, K., Schulz, J.: The design space of robots: Investigating children's views. In: Proceedings of the IEEE Workshop on Robot and Human Interactive Communication, pp. 47–52 (2004)

Spatial Misregistration of Virtual Human Audio: Implications of the Precedence Effect

David M. Krum, Evan A. Suma, and Mark Bolas

Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista, CA 90094, USA
{krum,suma,bolas}@ict.usc.edu

Abstract. Virtual humans are often presented as mixed reality characters projected onto screens that are blended into a physical setting. Stereo loudspeakers to the left and right of the screen are typically used for virtual human audio. Unfortunately, stereo loudspeakers can produce an effect known as precedence, which causes users standing close to a particular loudspeaker to perceive a collapse of the stereo sound to that singular loudspeaker. We studied if this effect might degrade the presentation of a virtual character, or if this would be prevented by the ventriloquism effect. Our results demonstrate that from viewing distances common to virtual human scenarios, a movement equivalent to a single stride can induce a stereo collapse, creating conflicting perceived locations of the virtual human's voice. Users also expressed a preference for a sound source collocated with the virtual human's mouth rather than a stereo pair. These results provide several design implications for virtual human display systems.

Keywords: virtual human audio, spatial sound, stereo audio, precedence effect, ventriloquism effect, mixed reality.

1 Introduction

Computer controlled virtual humans are an increasingly important part of applications in entertainment, training, therapy, novel human-computer interfaces, and social research. Often a 3D mixed reality presentation is preferred, where life-sized virtual characters are blended into a staged physical setting. Digital projectors are often the display of choice, since they are relatively inexpensive, can be used without head tracking, do not require users to don any display gear, and can be seen by multiple users at a number of angles and distances.

While a single loudspeaker located near a character's mouth and chest can portray the voice of a single virtual human character, this placement can be problematic. With a rear projected screen configuration, the loudspeaker would likely block the video image. Furthermore, placing the loudspeaker behind the screen would result in muffled audio. While there are perforated screens that allow sounds to pass, these screens require front projection. Rear projection is more desirable since it prevents users from accidentally blocking the projection

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 139–145, 2012.

and casting shadows across the character. A rear projection display combined with stereo loudspeakers is thus a common compromise in many installations.

As with any stereo pair, this configuration is subject to the precedence effect [15,19], which can interfere with stereo spatialization. The interference occurs when a listener is standing much closer to one of the two loudspeakers in a stereo pair. At this location, the wavefront from the nearby loudspeaker arrives sooner than, or precedes, the other loudspeaker's wavefront. The human perceptual system has an echo cancellation process, which ignores the second wavefront. Only the initial wavefront is perceived, causing the perceived sound location to collapse to the nearby loudspeaker, breaking the stereo spatialization. Our concern was that the precedence effect might cause a virtual human's voice to shift to the left or right as the listener moved around. Adding to our uncertainty was a second phenomenon, the ventriloquism effect, which might counteract the precedence effect. The ventriloquism effect can create the perception that a voice or sound, generated elsewhere, is emanating from the visual image of a temporally related source [5,7,13,18]. A previous study of interactions between ventriloquism and precedence effects demonstrated that they can work in concert, i.e. strengthening the perceived locality of a sound with a visual image that is coincident with a preceding sound source [11]. However, that study did not examine how the two effects might work in opposition.

With the goal of greater versimilitude in mixed reality training, it is problematic if a character's voice emanates from a point that is perceptibly offset from the character's mouth and body. Furthermore, a breakdown in spatialization can have negative effects on conversational interactions. Studies have shown that spatialization of multiple voices can increase speech comprehension, voice identification, and understanding [9,12,2].

Our goal was to examine the impact of precedence effect in a mixed reality virtual human presentation. Would it negatively impact a user's perception of the virtual human, or would it be masked by the ventriloquism effect working in opposition?

2 Related Work

A number of researchers believe that spatialized audio is an important sensory cue and have worked to improve 3D spatialized audio for users of mixed, augmented, and virtual reality [14,16,17,8]. Beyond stereophonic sound and its variants, current spatial audio reproduction systems include headphone based techniques using binaural audio and head related transfer functions [6] as well as techniques using arrays of loudspeakers like Ambisonics [10] and wavefield synthesis [3,4]. Headphone based techniques are less appealing since users must wear an additional device, preventing a simple "walk up and interact" experience. Wavefield synthesis requires large numbers of loudspeakers, perhaps hundreds or more, increasing cost and complexity. The Ambisonic technique typically requires four or more loudspeakers, as well as decoding hardware, and can have sound reproduction issues in large spaces without sound treatment to control echo and reverberation. Furthermore, wavefield synthesis and Ambisonic techniques may be unnecessary for virtual humans displayed on projected screens. The virtual characters appear on a screen in front of users, so the ability to present spatial sound from any point surrounding the users is simply unnecessary.

3 Methods and Apparatus

To determine if the precedence effect can alter the perception of a projected virtual human, we designed and conducted a mixed design study where participants listened to a virtual human reading literary passages. Participants were divided into three equal size groups for placement at one of three physical locations in front of the virtual human, representing the between subjects condition. Audio presentation was the within subjects condition.

Thirty-six participants, over the age of 18, with 20/20 corrected vision and self-identifying as having hearing in both ears were enrolled through email and the Craigslist website. The gender ratio was evenly balanced and participants ranged from 20 to 67 years of age (*M*=37.6, *SD*=13.6).

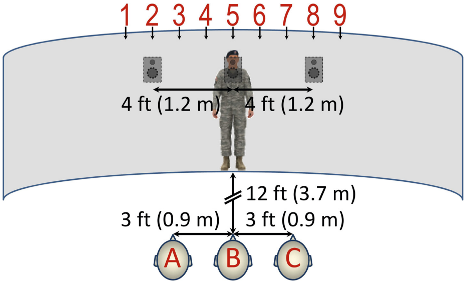


Fig. 1. Three loudspeakers, located behind a perforated (sound transparent) screen, provided audio to the left of the projected virtual human, at the center, or in stereo. Participants stood at positions: A, B, or C.

Participants were placed at one of three positions approximately 12 feet (3.7 m) in front of the virtual character. Position A was 3 feet (0.9 m) to the left of the screen mid-line, position B was on the mid-line, and position C was 3 feet (0.9 m) to the right of the mid-line (see Figure 1). This lateral distance was chosen, based on our experience with mixed reality virtual human installations for museums and military training, to represent the distance of a single stride that a single user might make or a comfortable distance between two participants.

The virtual human used throughout the study was a male soldier using a Cepstal LLC text-to-speech voice. The character was rendered using the Panda3D graphics library and animated to provide appropriate visemes, eye blinks, and breathing motions. Passages were selected from Herman Melville's *Moby Dick* for variety, duration, and good delivery by the character.

The character was projected onto a curved screen approximately 8.5 feet (2.6 m) tall and 31.3 feet (9.5 m) in width along the curve (see Figure 1). Three loudspeakers were placed behind the screen at the center, 4 feet (1.2 m) to the left of center, and 4 feet (1.2 m) to the right of center. The loudspeakers were Mackie HR824 high resolution studio monitors, mounted 57 inches (145 cm) high to bring them to the height of the virtual character's mouth and chest. A perforated screen was used, which allowed sound be heard through the screen. As previously mentioned, this perforated screen requires front projection and is thus not optimal for normal interaction with virtual characters. For this study, the perforated screen was satisfactory as user movement was restricted. Sound pressure levels were calibrated to provide matched values between audio presentations. Loudspeakers were also swapped halfway through the study to help counterbalance any tonal differences between loudspeakers. Some sound treatment was also applied behind the loudspeakers and screen to limit reverberations.

In the first phase of the study, each participant listened to three passages read by the virtual human and presented once each by either the left loudspeaker, the center loudspeaker, or a stereo pair. Participants were not told which audio presentation was being used. The order of the audio presentation was fully randomized. After each passage, participants were given two survey questions related to virtual human co-presence, based on the Bailenson et al. social presence questionaire [1]. ("I perceived the virtual human as being only a computerized image, not a real person." and "I perceived that the virtual human was present in the room with me.") Participants were asked to respond on 7 point scales. Participants then indicated the apparent horizontal location of the voice by referencing a set of markers from 1 to 9 placed along the top of the screen. The markers were spaced approximately 16 inches (41 cm) apart, with the 5 marker located at the center, above the center loudspeaker and the character.

In the second phase, participants listened to the virtual human's delivery of four sentence pairs. Each sentence pair consisted of the same sentence, repeated twice, but using different loudspeakers. Two of the sentence pairs were presented first in stereo and then by the center loudspeaker. The other two pairs were presented first by the center loudspeaker and then in stereo. Order of presentation was randomized. Participants were asked, for each sentence pair, "Which line was delivered more like a real person?", and asked to select either the first or second sentence. Participants were not told which audio configuration was used.

4 Results and Discussion

A mixed ANOVA statistical test was performed to determine if the within-subjects condition of audio presentation as well as the between-subjects condition of listener position created significant differences in the perceived location of the sound source at the $\alpha = .05$ level. Since Mauchly's Test of Sphericity indicated a possible violation of sphericity for the within-subject effects of audio presentation, we performed a Greenhouse-Geisser correction.

Data from the first phase of the study is listed in Table 1. A significant main effect of audio presentation (left, center, or stereo) was observed in perceived location, $F(2, 66) = 32.40, p < .001, \eta_p^2 = .50$. The effect of audio presentation was expected as the sound source changed position between audio presentations. This effect thus helps to confirm the validity of the experimental configuration. An examination of the 95% confidence intervals reveals that the perceived location of speech produced by the left loudspeaker, 95% CI [3.27, 4.45], is well separated and clearly different from the center [5.92, 6.35] and stereo [5.24, 6.09] presentations (see Figure 2a).

Table 1. Perceived Location by Audio Presentation and Listener Position. (Location values signify: 1=Left, 5=Middle, 9=Right.)

Audio Presentation	Listener Position	Mean Perceived Location	Standard Deviation
Left	A:Left	3.42	1.505
	B:Middle	4.33	2.270
	C:Right	3.83	1.267
	Total	3.86	1.726
Center	A:Left	6.67	0.778
	B:Middle	5.83	0.577
	C:Right	5.92	0.515
	Total	6.14	0.723
Stereo	A:Left	4.33	1.155
	B:Middle	5.00	1.279
	C:Right	7.67	1.303
	Total	5.67	1.897

The between subjects variable, listener position (A:Left, B:Middle, or C:Right) was observed to create a significant difference in perceived location $F(2, 33) = 5.56, p = .008, \eta_p^2 = .25$. More importantly, the crossing of the trendlines (see Figure 2b) shows a significant interaction effect for the stereo condition in combination with listener position $F(4, 66) = 10.16, p < .001, \eta_p^2 = .38$. For listeners at the rightmost position, the stereo sound source is perceived at the right side of the screen (larger numbers). For listeners at the leftmost position, the stereo sound source is perceived towards the left side (smaller numbers). This crossover is evidence of the precedence effect occuring in the stereo loudspeaker condition. There does appear a slight systemic shift to the right, perhaps due to room acoustics, as well as some pull towards the virtual human's central visual image, possibly due to the ventriloquism effect. However, the magnitudes of these effects do not obscure the interaction which suggests the precedence effect.

We did not observe any significant effect of audio presentation on the two questions concerning co-presence at the $\alpha = .05$ level with a mixed ANOVA. Several sources of variance may have affected these measures. Many participants may have been unfamiliar with virtual humans and had little common reference for co-presence. We also observed some possible confusion concerning the direction of the scales for the two questions. A scale reversal is present in the original social presence questionnaire from which these questions were adapted. Furthermore, the baseline realism of the virtual human's voice and behavior were limited, possibly overwhelming any contribution of varying audio presentation.

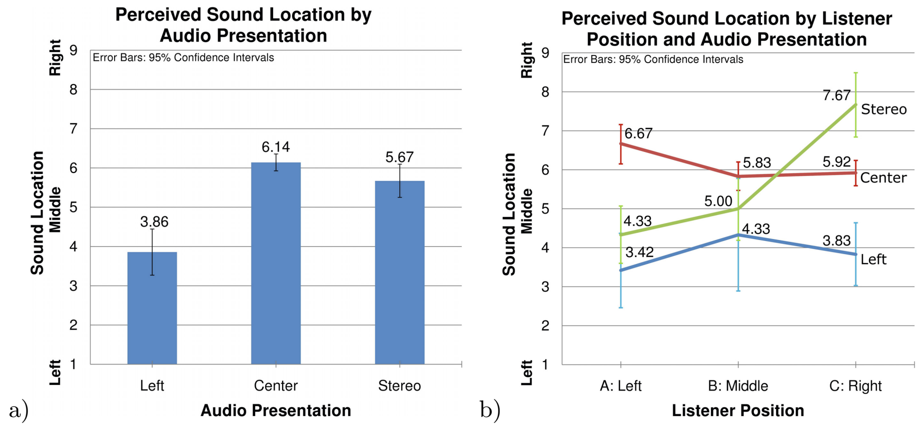


Fig. 2. a) Significant differences in perceived location were observed for audio conditions. Localization of speech from the left loudspeaker clearly differed from center and stereo presentations. b) A significant interaction effect on perceived location was observed for listener position and the stereo audio condition. Listeners on the left localized the stereo sound to the left, while listeners on the right localized it to the right.

For the second phase of the study, the sentence pair trials, loudspeaker preferences for each of the four trials were recoded numerically (0=center, 1=stereo) and then summed to provide an overall preference score. A One-Sample Wilcoxon Signed Ranks Test was conducted to compare these scores against an expected median value of 2.0, corresponding to random chance. The observed median of 1.0 indicated a significant preference for the center speaker location, $Z = -3.71, p < .001$, suggesting that a single loudspeaker delivered better realism.

5 Conclusion and Future Work

This study demonstrates that the precedence effect can occur in a typical stereo configuration for virtual human audio, causing misperception of the audio source. The offset distance tested can easily occur with a single mobile participant or multiple participants. These results demonstrate limitations of stereo loudspeaker pairs in supporting user movement and multiple users around virtual humans. While adjustments to stereo phase and panning can compensate for motion of a single user, user tracking is required, and these adjustments cannot scale to multiple users. Designers should examine the range of motion and number of users required and select complementary audio/visual components that can robustly collocate virtual human audio and visual imagery for the given installation. Consideration of perforated screens and individual loudspeakers assigned to each virtual character may be warranted. We expect that these results will also inform development of new technologies for presenting virtual human audio. To be of interest to virtual human installation designers, these approaches should be compatible with projected displays and attempt to better replicate the proximal sound field of a human voice.

References

1. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence-Teleop. Virt. 10, 583–598 (2001)
2. Baldis, J.J.: Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In: ACM CHI, pp. 166–173 (2001)
3. Berkhout, A.J.: A holographic approach to acoustic control. J. Audio Eng. Soc. 36(12), 977–995 (1988)
4. Berkhout, A.J., De Vries, D., Vogel, P.: Acoustic control by wave field synthesis. J. Acoust. Soc. Am. 93, 2764–2778 (1993)
5. Bertelson, P.: Chapter 14 ventriloquism: A case of crossmodal perceptual grouping. In: Gisa Aschersleben, T.B., Msseler, J. (eds.) Cognitive Contributions to the Perception of Spatial and Temporal Events, Advances in Psychology, vol. 129, pp. 347–362. North-Holland (1999)
6. Blauert, J.: Räumliches Hören (Spatial Hearing). S. Hirzel-Verlag, Stuttgart (1974)
7. Choe, C., Welch, R., Gilford, R., Juola, J.: The ventriloquist effect: Visual dominance or response bias? Atten. Percept. Psycho. 18, 55–60 (1975)
8. Courgeon, M., Rebillat, M., Katz, B., Clavel, C., Martin, J.C.: Life-sized audiovisual spatial social scenes with multiple characters: MARC & SMART-I2. In: Meeting of the French Association for Virtual Reality (2010)
9. Ericson, M.A., Brungart, D.S., Simpson, B.D.: Factors that influence intelligibility in multitalker speech displays. Int. J. Aviat. Psychol. 14, 313–334 (2004)
10. Fellget, P.: Ambisonics. part one: General system description. Studio Sound 17, 20–22, 40 (August 1975)
11. Harima, T., Abe, K., Takane, S., Sato, S., Sone, T.: Influence of visual stimulus on the precedence effect in sound localization. Acoust. Sci. Tech. 30(4), 240–248 (2009)
12. Ihlefeld, A., Sarwar, S.J., Shinn-Cunningham, B.G.: Spatial uncertainty reduces the benefit of spatial separation in selective and divided listening. J. Acoust. Soc. Am. 119(5), 3417–3417 (2006)
13. Jack, C.E., Thurlow, W.R.: Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Percept. Motor Skill. 37, 967–979 (1973)
14. Li, Z., Duraiswami, R., Davis, L.: Recording and reproducing high order surround auditory scenes for mixed and augmented reality. In: IEEE and ACM ISMAR, pp. 240–249 (November 2004)
15. Litovsky, R.Y., Colburn, H.S., Yost, W.A., Guzman, S.J.: The precedence effect. J. Acoust. Soc. Am. 106, 1633–1654 (1999)
16. Sodnik, J., Tomazic, S., Grasset, R., Duenser, A., Billinghurst, M.: Spatial sound localization in an augmented reality environment. In: OZCHI, pp. 111–118 (2006)
17. Sundareswaran, V., Wang, K., Chen, S., Behringer, R., McGee, J., Tam, C., Zahorik, P.: 3D audio augmented reality: implementation and experiments. In: IEEE and ACM ISMAR, pp. 296–297 (October 2003)
18. Thomas, G.: Experimental study of the influence of vision on sound localization. J. Exp. Psychol. 28(2), 163–177 (1941)
19. Wallach, H., Newman, E.B., Rosenzweig, M.R.: The precedence effect in sound localization. Am. J. Psychol. 62(3), 315–336 (1949)

Virtual Human Personality Masks: A Human Computation Approach to Modeling Verbal Personalities in Virtual Humans

Vaishnavi Krishnan[1], Adriana Foster[2], Regis Kopper[1], and Benjamin Lok[1]

[1] Computer and Information Sciences Engineering, University of Florida,
Gainesville, FL 32611, USA
{vakrishn,kopper,lok}@cise.ufl.edu

[2] Psychiatry and Health Behavior, Georgia Health Sciences University
Augusta, GA 30912, USA
afoster@georgiahealth.edu

Abstract. Modeling virtual humans that can exhibit realistic personalities is becoming increasingly important as virtual humans are being widely used for inter-personal skills education. We present Virtual Human Personality Masks, a system that combines human computation with the idea of using existing virtual humans to bootstrap the creation of other virtual humans to enable quick and easy generation of perceivable verbal personalities in virtual humans.

To evaluate this system, we created high and low verbosity-level variants of a virtual patient with symptoms of depression and conducted a user study with medical students split between two groups, each interacting with one of the two variants of the virtual patient. The participants' perceived verbosity levels of the virtual patients indicated that not only did the virtual patients created using our system exhibit the intended personality in a perceivable manner, but also exhibited other related personality attributes in a manner that is consistent with the human personality theory analogs of verbosity.

Keywords: Virtual Humans, Personalities, Verbosity, Human Computation, Conversational Agents, Virtual Patients.

1 Introduction

Virtual Humans (VHs) are increasingly being used as conversational partners in inter-personal skills education [1, 2]. To increase usefulness, inter-personal skills education should provide interactions with VHs that exhibit realistic personalities. In this paper, we propose a method for rapidly creating variants of existing VHs that can exhibit perceivable personalities using a human computation based approach[3].

Existing models of personality for conversational agents often use fixed sets of personality variables in combination with a decision algorithm to add verbal (the corpus) or non-verbal (gestures and emotions) personalities to the agent [4-9]. One such approach models the personality variables as states in a finite state machine [4]. Another approach uses a layered, Bayesian Belief Network to model personalities, moods and emotions [8]. Yet another approach uses multiple dimensions to represent a given personality state [9]. While these can be used to represent most generic personality states, the

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 146–152, 2012.

amount of time required to apply such models to specific scenarios depends on the extensiveness of the model creation process and the amount of refining required to apply the model to the scenario. Our approach aims at reducing these dependencies using *personality masks* created out of real humans. The *personality masks* are created by gathering different human responses to a subset of the stimuli from an existing stimulus-response corpus and can then be applied onto the original corpus, resulting in different personality variants of the original VH. We built our system on top of Virtual People Factory (VPF), an online application for users to create and interact with VHs [11].

To demonstrate the generic concept of using the Personality Masks system to generate perceivable personalities, we created high and low *verbosity masks* and used the resulting *verbosity-masked* VHs to evaluate if the intended verbosity levels were perceivable to users. The results indicated that the VHs created using our system exhibit the intended verbosity levels in a manner that was perceivable for the interactants and consistent with the human personality theory analogs of verbosity.

2 Virtual Human Personality Masks

2.1 Verbosity and Patients with Depression

Patients' talkativeness(verbosity) levels in medical interviews can be influenced by factors such as their level of comfort with the interviewer, the environment, their age or the presence of personality disorders [12] and is often associated with extraversion, one of the five factors in the Big Five theory of personalities [13]. Extraverted people are perceived as easier to interview, more willing to disclose information and more cooperative and extraversion in patients is directly related to their perception of social support [14], suggesting that it may influence the patients' prognosis. This relevance of verbosity to the depression scenario and the fact that verbosity is easily perceivable and quantifiable, led us to choose it as the personality trait to be modeled for a 21-year old depressed VP, Cynthia Young (created using VPF) [15-18] using our system.

The two phases involved in the system are described in the following sections.

2.2 Mask Creation Phase

In this phase, human responses to a subset of the stimuli in the corpus are used to create different personality masks for the intended personality. To achieve this, we apply the findings of Rossen et al. [10] by using an existing question-answering VH (a VP) to create a question-asking VH (a virtual doctor), using the set of stimuli (what the users can say to the VH) and responses (what the VH will respond to the user) from the original corpus. The question-asking VH (virtual doctor) is then used to gather the personality-specific responses that will form the personality masks for the intended personality. Below, we describe the steps for the Mask Creation phase.

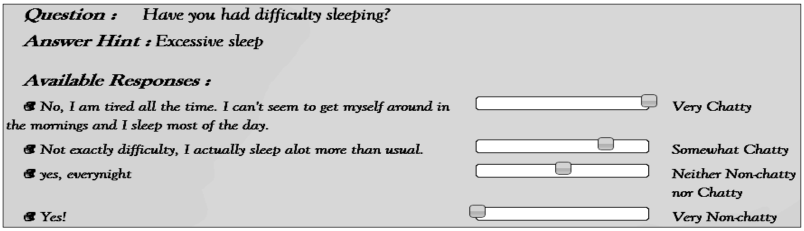
Step 1: Analysis of Interaction logs. Interaction logs of the existing corpus are analyzed to find the most frequently triggered stimuli. Since these stimuli are most likely to appear in future interactions with the VH, the ability of the VH to exhibit the intended personality while responding to these stimuli is pivotal to the design of our system. The threshold value for extracting such stimuli defaults to 25% usage, and is customizable by the domain expert. For the depression scenario, a set of approximately 100 stimulus-response pairs were filtered out from the original corpus with about 3500 stimuli and 400 responses.

Step 2: Review and refinement by domain expert. A domain expert reviews and refines the set of stimulus-response pairs filtered out in Step 1, using his/her expertise to determine which of the filtered stimuli can be used to elicit the intended personality trait. For example, the analysis of the interaction logs for our depressed VP corpus resulted in "What is your name?" as one of most frequently used stimuli, but is not useful in eliciting specifically long or short responses, not contributing to the creation of varied verbosity levels. For our example, 62 of the 100 most-frequently used stimulus-response pairs were selected by the domain expert for creating the virtual doctor.

Step 3: Response hint generation by domain expert. Once the set of interview questions are finalized, the domain expert fills-in response hints to guide the human interviewee to phrase their answers in a manner consistent with the original scenario. For example, the question "Have you had difficulty sleeping?" is open-ended and could be answered with excessive, normal or reduced sleep. However, in the original case history, the patient complained of excessive sleep and changing this key fact could completely change the diagnosis. The domain expert therefore added "Excessive sleep" as a response hint for that stimulus.

Step 4: Gathering responses from human interviewees. After the virtual doctor is created, an online chat-style interview link is sent out to several human interviewees who present varied verbosity levels. Before starting the interview, a description of the role that the interviewee will play during the interview is provided. In our example, this description asks the interviewee to role-play as a 21-year old student with symptoms of depression and talk to the virtual doctor as they would under such circumstances. During the interview, the virtual doctor asks questions to the interviewee while providing response hints for each question.

Step 5: Creation of Masks. Finally, the responses gathered from the human interviewees are cast into different *personality masks* by the domain expert using personality sliders that can be used to assign a "level" of the intended personality trait for each response. Fig. 1 shows an example of the verbosity slider for one of the 62 stimuli.



Question : Have you had difficulty sleeping?

Answer Hint : Excessive sleep

Available Responses :

- No, I am tired all the time. I can't seem to get myself around in the mornings and I sleep most of the day. — Very Chatty
- Not exactly difficulty, I actually sleep alot more than usual. — Somewhat Chatty
- yes, everynight — Neither Non-chatty nor Chatty
- Yes! — Very Non-chatty

Fig. 1. Verbosity sliders in the Personality Masks system for the depression example

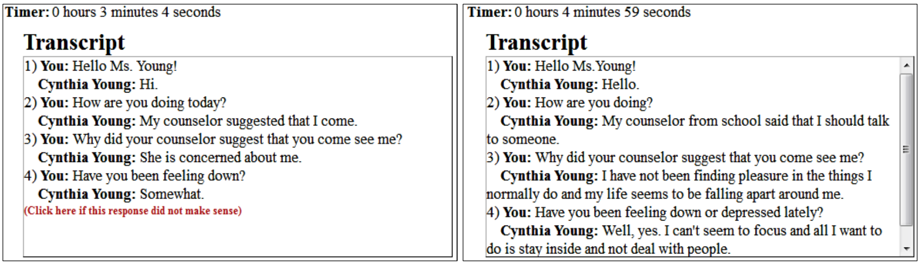
2.3 Mask Application Phase

In this phase, the generated *personality masks* are applied onto the VH in the original corpus, resulting in a corpus that has *personality-masked* responses for some of the stimuli and default responses from the original corpus for the remaining stimuli.

3 Evaluation Study

We conducted a between-subjects user study with verbosity level of the VP as the independent factor. The participants were medical students (n=31, mean age = 24.2 years, 19 female and 12 male). They were randomly split into two groups and were assigned to interview either the high (n=16) or low (n=15) verbosity-level variant of the depressed VP in an online chat-style interaction. At the end of the interview, they completed a post-survey. Fig. 2 shows interaction transcripts for these conditions.

The two verbosity masks were created from the responses provided by 11 female staff and residents from the medical school for the virtual doctor's questions in online chat-style interviews that lasted about 20 minutes each.



Timer: 0 hours 3 minutes 4 seconds

Transcript

1) **You:** Hello Ms. Young!
Cynthia Young: Hi.
2) **You:** How are you doing today?
Cynthia Young: My counselor suggested that I come.
3) **You:** Why did your counselor suggest that you come see me?
Cynthia Young: She is concerned about me.
4) **You:** Have you been feeling down?
Cynthia Young: Somewhat.
(Click here if this response did not make sense)

Timer: 0 hours 4 minutes 59 seconds

Transcript

1) **You:** Hello Ms.Young!
Cynthia Young: Hello.
2) **You:** How are you doing?
Cynthia Young: My counselor from school said that I should talk to someone.
3) **You:** Why did your counselor suggest that you come see me?
Cynthia Young: I have not been finding pleasure in the things I normally do and my life seems to be falling apart around me.
4) **You:** Have you been feeling down or depressed lately?
Cynthia Young: Well, yes. I can't seem to focus and all I want to do is stay inside and not deal with people.

Fig. 2. Interaction transcripts of low (left) and high (right) verbosity VP

Primary Hypothesis. Participants in the low-verbosity group will perceive a lower level of verbosity in their VP than participants in the high-verbosity group.

Secondary Hypothesis. Participants in the low-verbosity group will perceive their VP as less concerned, less cooperating, less willing to disclose information and less comfortable to deal with than the participants in the high-verbosity group.

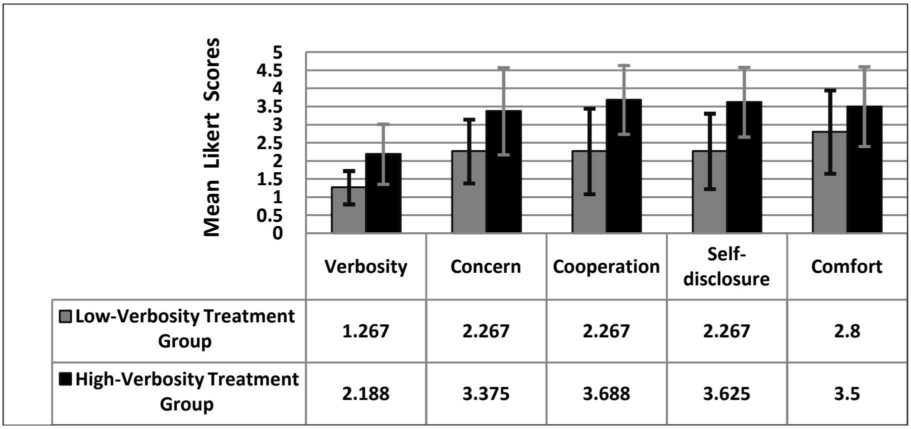
The metrics used in the post-survey for this study are summarized in Table 1.

Table 1. Likert scale rating (1-Strongly Disagree/5-Strongly Agree) statements for all metrics

Metric	Likert scale rating statements
Verbosity (Primary)	"The patient I spoke to gave detailed or long answers to the questions that I asked her in the interview"
Concern (Secondary)	"I think the patient was quite concerned about her problem"
Cooperation (Secondary)	"I think the patient cooperated well with me in answering my questions"
Self-disclosure (Secondary)	"I think the patient was willing to disclose information about herself."
Comfort (Secondary)	"I liked talking to the patient and felt like I was able to help her."

4 Results

The mean Likert scores for each metric across the two treatment groups are shown in Fig. 3. Table 2 summarizes the results of the Mann-Whitney U tests on the Likert scores. Results show that the scores for perceived levels of *verbosity*, *concern*, *cooperation*, *self-disclosure* and *comfort* from participants in the low-verbosity treatment group were significantly lower than the scores for perceived verbosity-level from participants in the high-verbosity treatment group. All differences were significant at $p < 0.05$ or less. Thus, we **accept both our Hypotheses**.



	Verbosity	Concern	Cooperation	Self-disclosure	Comfort
▨ Low-Verbosity Treatment Group	1.267	2.267	2.267	2.267	2.8
■ High-Verbosity Treatment Group	2.188	3.375	3.688	3.625	3.5

Fig. 3. Mean Likert Scores for all metrics (Error bars represent standard errors)

Table 2. Mann-Whitney U test statistics for all metrics

Metric	*U*	*Z*	|*r*|	*p (one-tailed)*
Verbosity	44.5	-3.24	0.58	0.001
Concern	59.0	-2.554	0.002	0.006
Cooperation	63.5	-2.325	0.42	0.010
Self-disclosure	61.5	-2.435	0.44	0.007
Comfort	79.5	-1.733	0.31	0.047

Table 3 summarizes the user comments, reinforcing the results from the post-survey.

Table 3. Examples of participants‘ responses to open-ended questions

Question Topic	High Verbosity Group	Low Verbosity Group
Overall experience	*“It was a very accurate experience. I enjoyed the activity.”*	*“It was very hard to ask further questions on a subject when she gave short or avoiding answers.”*
	“I felt like I wanted to speak more personally with her”	*“Real patients are more forthcoming when you ask them questions such as ‘why did you come here today?’.”*
Personality of the VP	*“Very typical depressed mood, but eager to get back to feeling like herself.”*	*“She seemed aware that there was a change in her behavior, but was hesitant to delve into the root of her issues.”*
	“She seemed very open & willing to talk to me even though she was depressed.”	*“Dry- perhaps she was simply depressed and down and having a hard time getting up the motivation to explain herself”*

5 Discussion

Analysis of the results of our study indicates that participants were able to perceive the intended verbosity levels in the VPs they interviewed. The results for *concern*, *cooperation*, *self-disclosure* and *comfort* were also congruent with the Big Five theory of personality characteristics in which talkativeness and sociability are frequently associated with extraversion [13, 16]. These results also showed that our system has been successful in shifting the effort from modeling *standardized* personalities to modeling *perceivable* personalities.

However, we also observed that the participants' perceived verbosity-level of the high-verbosity VP were not as high as we expected them to be, even though the difference was statistically significant. One potential reason for this is the observation that the high-verbosity VP might not have been a considerable deviation from regular patients, as inferred from some of the previous work, that assert the fact that humans respond better to computer-based agents that are highly exaggerated [19].

6 Conclusions and Future Work

We have presented Virtual Human Personality Masks, a system that combines the concepts of human computation [3] with the idea of using existing VHs to bootstrap the creation of other VHs [10], to rapidly generate perceivable personalities in VHs. Below, we present the future directions for this research as identified by our experiences in designing the Personality Masks system and our understanding of potential applications for this system.

We aim to extend our work to model more complex personalities, such as the personality traits listed in the Big Five Personality model (OCEAN model) [15]. From our experiences with modeling verbosity as a personality trait for our depressed VP, we identified that modeling complex personalities would need inputs from other sources as well. In the case of our verbosity example, both the review and refinement of stimuli for the role-reversed VH and the response hint generation were done by the domain expert. However, for more complex personalities, psychology experts might provide guidelines on selecting the stimuli that can elicit the intended personality, while the domain experts provide response hints for each of those stimuli.

We also intend to explore the learning effects of using the *personality-masked* VHs in interpersonal skills education by running studies that measure the experience gained by interacting with VHs of different personalities.

We will also run follow-up studies to compare the extent of personality perception across different scenarios – for example, comparing the perception of personalities in the depression scenario (which needs a high-level of emotional connection between the patient and the doctor) with those in a Cranial Nerve disorder scenario (which mostly involves physical exams and less verbal interaction between the patient and the doctor). Through this study, we will aim to establish a relationship between the type of the scenario and the extent to which virtual human personalities would affect the learning of interpersonal skills in the scenario.

Acknowledgements. Thanks to Dr. Carolyn Stalvey, UF College of Medicine, for helping us find study participants, all participants for their time, reviewers for constructive feedback and members/alumni of the Virtual Experiences Research Group, UF for their support.

References

1. Johnsen, K., et al.: The validity of a virtual human experience for interpersonal skills education. In: Proc. CHI 2007, pp. 1049–1058. ACM, San Jose (2007)
2. Randall, W.H., et al.: Virtual Humans in the Mission Rehearsal Exercise System (2003)
3. von Ahn, L.: Human computation. In: Proc. DAC 2009 (2009)
4. Badler, N.I., Reich, B.D., Webber, B.L.: Towards Personalities for Animated Agents with Reactive and Planning Behaviors. In: Creating Personalities for Synthetic Actors, Towards Autonomous Personality Agents, pp. 43–57. Springer (1997)
5. Ball, G., Breese, J.: Emotion and personality in a Conversational Character. In: Workshop on Embodied Conversational Characters, Tahoe City, CA (1998)
6. Doce, T., Dias, J., Prada, R., Paiva, A.: Creating Individual Agents through Personality Traits. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 257–264. Springer, Heidelberg (2010)
7. Kasap, Z., Magnenat-Thalmann, N.: Intelligent virtual humans with autonomy and personality: state of the art. In: Intelligent Decision Technologies, vol. 1, pp. 3–15 (2007)
8. Kshirsagar, S., Magnenat-Thalmann, N.: Virtual humans personified. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1, Bologna, Italy, pp. 356–357. ACM (2002)
9. Rushforth, M., Gandhe, S., Artstein, R., Roque, A., Ali, S., Whitman, N., Traum, D.: Varying Personality in Spoken Dialogue with a Virtual Human. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 541–542. Springer, Heidelberg (2009)
10. 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)
11. Rossen, B., Lind, S., Lok, B.: Human-Centered Distributed Conversational Modeling: Efficient Modeling of Robust Virtual Human Conversations. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 474–481. Springer, Heidelberg (2009)
12. Sadock, B.J., Sadock, V.A.: Psychiatric History and Mental Status. In: Kaplan and Sadock's Synopsis of Psychiatry (2007)
13. Fleeson, W., Gallagher, P.: The implications of Big Five standing for the distribution of trait manifestation in behavior: Fifteen experience-sampling studies and a meta-analysis. J. Pers Soc. Psychol. 97(6), 1097–1114 (2009)
14. Leskelä, U., et al.: The Influence of Major Depressive Disorder on Objective and Subjective Social Support: A Prospective Study. J. Nerv. Ment. Dis. 196(12), 876–883 (2008)
15. Goldberg, L.R.: An alternative description of personality: The big-five factor structure. Journal of Personality and Social Psychology 59, 1216–1229 (1990)
16. Nakao, K., et al.: The influences of family environment on personality traits. Psychiatry and Clinical Neurosciences 54(1), 91–95 (2000)
17. Hevil Shah, M.P.H., et al.: Interactive Virtual-Patient Scenarios: An Evolving Tool in Psychiatric Education. Academic Psychiatry 36, 146–150 (2012)
18. Nuzzarello, A.: C Birndorf, An Interviewing course for a psychiatry clerkship. Academic Psychiatry 28, 66–70 (2004)
19. Mateas, M.: An Oz-Centric Review of Interactive Drama and Believable Agents. In: Wooldridge, M.J., Veloso, M. (eds.) Artificial Intelligence Today. LNCS (LNAI), vol. 1600, pp. 297–328. Springer, Heidelberg (1999)

The Effect of Visual Gender on Abuse in Conversation with ECAs

Annika Silvervarg[1], Kristin Raukola[1], Magnus Haake[2], and Agneta Gulz[1]

[1] Dept. of Computer and Information Science, Linköping University, Linköping, Sweden
{annika.silvervarg,agneta.gulz}@liu.se
[2] Cognitive Science, Lund University, Kungshuset, Lundagård, Lund, Sweden
magnus.haake@lucs.lu.se

Abstract. Previous studies have shown that female ECAs are more likely to be abused than male agents, which may cement gender stereotypes. In the study reported in this paper a visually androgynous ECA in the form of a teachable agent in an educational math game was compared with a female and male agent. The results confirm that female agents are more prone to be verbally abused than male agents, but also show that the visually androgynous agent was less abused than the female although more than the male agent. A surprising finding was that very few students asked the visually androgynous agent whether it was a boy or a girl. These results suggest that androgyny may be a way to keep both genders represented, which is especially important in pedagogical settings, simultaneously lowering the abusive behavior and perhaps most important, loosen the connection between gender and abuse.

Keywords: Embodied conversational agent, pedagogical agent, teachable agent, social conversation, off-task conversation, visual gender, abuse.

1 Introduction

Embodied conversational agents – computer software that interacts with a user in natural language via text or voice and is represented by an anthropomorphic body or part of a body – are becoming increasingly common. We find ECAs in a variety of contexts, where they provide assistance, offer information, offer company, serve as coaches or instructors, etc.

Research on the conversational and social interactions between humans and ECAs has focused on their potential beneficial effects: boosting and increasing learning, trust, engagement, etc. [1, 2]. However, an early call to also consider “a darker side” of these interactions was provided by De Angeli, Brahnam & Wallis [3]. In a study of conversations between humans and ECAs, De Angeli & Brahnam [4] made the observation that approximately 11% of the conversations were concerned with hardcore sex. Another observation was that the female agent suffered considerably more abuse than the male counterpart or the agent depicted as a robot. Not only were there quantitative differences, but also differences in their nature with considerably more threats, violence and coarse proposals in relation to the female ECA. A more recent study by

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 153–160, 2012.

De Angeli [5], analyzing interaction logs of a chatbot that assumed different embodiments, replicated the findings. The female agents were considerably more prone to be sexualized and verbally abused than the male agents. As the author remarks, the findings are not surprising given the research in social psychology on the function and impact of gender stereotypes – but this does not imply that they should be ignored.

One domain where phenomena such as abusive conversation and reinforcement of cultural stereotypes are of particular importance is education – a domain with an ethical dimension constantly present [6]. But in spite of ECAs becoming more frequent in educational contexts, for instance in educational software, there is very little exploration of issues surrounding agent abuse in the domain of education technology. An exception is a study by Veletsianos, Scharber & Doering [7] with an informative title: “When Sex, Drugs, and Violence Enter the Classroom – Conversations between Adolescents and a Female Pedagogical Agent”. Their analysis of the conversation between 59 14-15 year olds and the agent Joan, designed to assist the students in a social studies assignment, revealed that the prevailing type of conversation was *off-task* rather than *on-task* and that a large proportion was abusive. Veletsianos et al. [7] bring up the visual representation of the agent, pointing out that previous work has shown that user reactions are influenced by the visual representation of virtual agents. They discuss whether Joan, presented as in her twenties or early thirties, blond and attractive, would have received less abuse if she had been portrayed in a more professional and teacher-like manner, e.g. more authoritative, older, wearing glasses, etc.

The study to be presented in this text has similarities with that of Veletsianos et al. [7]. It includes 35 13-14 year olds, an educational setting, and an ECA that engages in both on-task and off-task conversation. However, our ECA does not assume an assistant, teacher or instructor role. Instead it is a *teachable agent* (*TA*), assuming the role of the student’s tutee – a somewhat younger, not so knowledgeable peer. Our question was how to handle the visual design of the TA given that: (i) our target group according to previous studies is relatively prone to use abusive language, and (ii) we do not want to give fuel to gender stereotypes according to which females are a much more common target for this abuse than males. Age was not a factor that we could manipulate. To exclude girl characters and limit ourselves to boy TAs did not seem an attractive solution – just as we deem it an unattractive solution for pedagogical software in general. Choosing a non-human – e.g. an animal character – would diminish some of the affordances in a human peer as sharing the experiences of going to school, having lessons in various subjects and being interested in support from an older student.

Our choice was hence to set up a study to explore what would happen in terms of verbal abuse with a *visually androgynous* TA – that could be seen as both a boy and a girl – compared to a boy TA and a girl TA. Next we briefly describe the agent, its pedagogical setting, conversational abilities, and three different visual representations.

2 A Teachable Agent with Social Conversational Skills

The ECA in the study to be presented is situated in an educational game for basic mathematics. For technical details of the system see Pareto, Haake, Lindström, Sjödén, Gulz [8]. The ECA is a *teachable agent* which, in brief, means that the agent is taught or trained by the student, and AI techniques guide the agent's behavior based on what it is taught [9, 10]. The Learning-by-teaching pedagogy is well documented as being powerful, both in general and as implemented via teachable agents [11, 12]. In several other studies on the TA-game used in the present study, learning outcomes in relation to game and agent characteristics have been researched [8, 13]. The present study has a focus on attitudes. But, importantly, attitudes to mathematics can have a significant impact on mathematics learning outcomes [14].

The TA in the math game can engage in two conversational modes. One is a multiple-choice guided conversation where the TA asks game related mathematical questions and the student explains by choosing one out of several alternative answers. The other mode is social conversation where users write freely by means of the keyboard and bring up basically any topic in a chat-like manner. The topics can be both *on-domain* such as school, math, the math game and *off-domain* like music and movies. The system uses the metaphor of regular breaks between lessons in school for switching between on-task activities (i.e. playing the game and on-task dialogue) and off-task activities (i.e. social conversation). A recent study of the game by Gulz, Haake & Silvervarg [13] indicates that the social conversation can: (i) improve students' game experience, (ii) improve learning outcomes, and (iii) engage learners in voluntary on-domain chats.

The off-task conversation is implemented using AIML (*Artificial Intelligence Markup Language*) [15]. It is implemented as a mixed-initiative dialogue strategy, which allows both the agent and the user to direct the dialogue. The agent is able to handle greetings, ask and respond to questions and statements on various topics, ask and respond to follow up questions, deal with utterances outside its domain of knowledge, handle abuse, know when and how to change topic, and to tell mini narratives. For further technical details see [16].

The verbal abuse towards the agent takes the form of comments and questions regarding the agent's appearance, intellectual capacities, and sexual orientation and activities. To fend this off, the agent identifies such utterance using a list containing words classified as abusive. The list has been compiled during iterative development and testing with the target user group. The agent employs a strategy where the two first responses are general and acknowledge the abusive utterance but without any strong reaction (since back talk in this user population seems to encourage continued abuse), followed by an attempt to change the subject, and finally a strong marking that abuse is not tolerated and that the agent will not respond at all as long as the abuse continues. This is exemplified in the dialogue excerpt below:

User: you are ugly
Agent: eh, let's talk about something else
User: you smell
Agent: come on, let's talk about something else
User: ugly
Agent: maybe we could talk about hobbies instead
User: ugly
Agent: knock it off, or I won't talk to you anymore

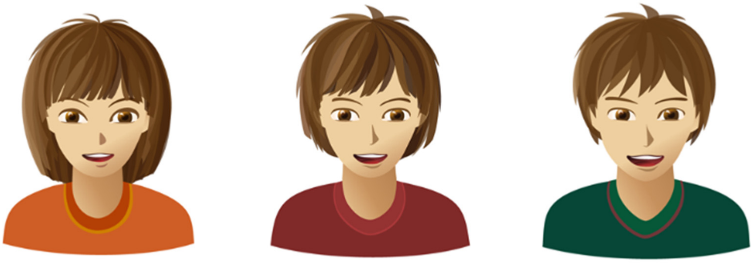


Fig. 1. The agents' visual representation: female, androgynous, and male

Three agent representations were used in the study, differing in their visual gender (female, androgynous and male), see Fig. 1. Interests, conversational style, etc., were the same and designed to be gender neutral. The visual design of the characters used a common basic set of graphical elements and exploited a number of visual strategies.

- *Female:* Manipulated with regard to visual feminine attributes such as rounded shapes, big eyes, curved eye browses, small nose and mouth (cf. the *baby face* scheme) as well as pronounced eye lashes, long hair, and narrow shoulders.
- *Male:* Manipulated with regard to visual male attributes such as more angular, broad, and/or pronounced shapes (head, eyes, nose, mouth, etc.), straight eye brows, short hair, and broader shoulders.
- *Androgynous:* The visual androgynous attributes were manipulated to be somewhere between the female and male attributes (e.g. shapes, eye brows, and mouth). Especially the length of the hair was iterated several times.

Before the final design of the androgynous character, a test was carried out with 38 students. The test showed a tendency for perceiving the androgynous character as female. After this there was a last round of graphical fine tuning and evaluation with 47 13-14 year olds involved. 31 out of these students perceived the androgynous character as looking neither as a boy nor a girl. Eight students (3 girls, 5 boys) perceived it as a girl, and 8 students (4 girls, 4 boys) perceived it as a boy.

3 Study

Twenty-three female and 20 male 13-14 year olds from two classes in a Swedish school participated in the study. The students played the math game and interacted with the ECAs during two subsequent lessons, once with the girl or the boy agent and once with the visually androgynous agent. Half of the participants encountered the androgynous agent during the first lesson and the other half the boy or the girl agent. During each lesson they played the game, and chatted with the agent during two five minutes "breaks". Three female and 4 male students could not participate at both occasions. Three additional male students were removed as outliers (demonstrating extreme cases of verbal abuse). With this the final analysis included 20 females and 13 males, still balanced as to the presentation order of the agents.

User data was recorded for gender, math skill, agent representations, game data, chats, and questionnaire data. For this paper, the reported results are based on evaluations of verbal abuse within the chatlogs with regard to the visual gender of the agent. The chatlogs were annotated using the following coding schema:

1. *Light abuse:* noob, nerd, shut up, don't care, ugly (hair, clothes), shit, etc.
2. *Medium abuse:* gay/fag, stupid, idiot, moron, eat shit, shit yourself, creep, etc.
3. *Coarse abuse:* degrading expressions with swear words and/or sexual content.

Often the level of abuse of an expression is culture and language specific. Some abusive expressions are also hard to translate at all. The actual scheme is developed for Swedish expressions. A straightforward translation into another language (as above) may appear non-intuitive. Furthermore, the abusive weight in expressions may differ substantially between age groups. Notably Swedish teenagers have been involved in the construction of this classification scheme. The amount of abuse was calculated for each agent in relation to the number of users that interacted with the agent. In Fig. 2 the amount of different types of abusive utterances is presented, both the total number of abuses and a weighted total (*medium* with weight = 2, and *coarse* with weight = 3) are included. The study replicated earlier findings showing significantly (two tailed *t*-test: $p = 0.040$) more verbal abuse towards the female agent in contrast to the male agent. Though not significant, Fig. 2 indicates that for both the number of abusive comments and the weighted level of abuse, the results for the androgynous agent lie between the results for the female and male agent.

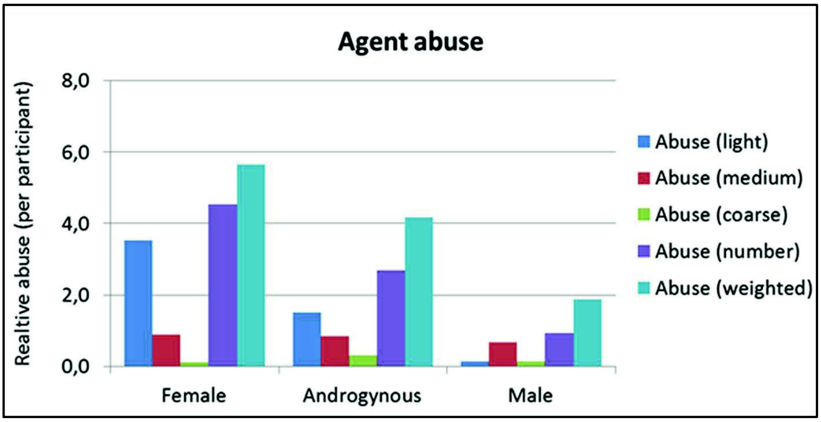


Fig. 2. The number and types of abuses (per participant) with regard to the visual gender (female, androgynous, or male) of the agent

Remarkably, the 13 male participants were by far more responsible for abusive comments (two-tailed t-test on Abuse (number & weighted): $p < 0.001$), see Fig. 3. Notably, there was no coarse abuse from the female participants, although they were more numerous (20 of 33).

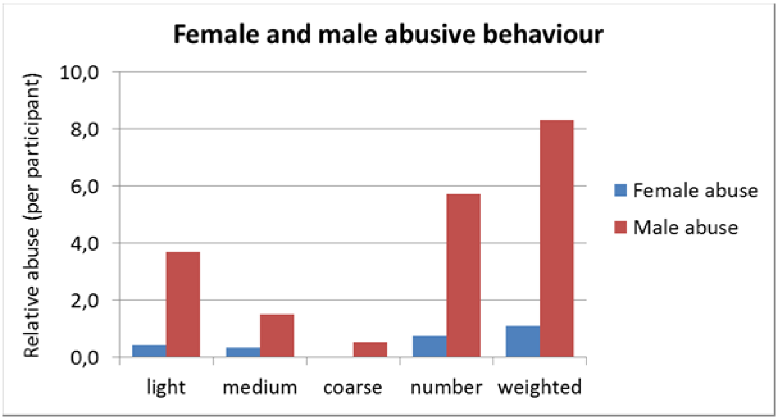


Fig. 3. Abusive behavior (per participant) for females and males respectively

A surprising finding was also that very few students asked the visually androgynous agent whether it was a boy or a girl whereas significantly more (one-tailed *t*-test: $p = 0.042$) students posed this question to the male and female agent. Particularly, participants did not ask for the agent's gender if it had the same visual gender as themselves.

4 Discussion

Our results replicated previous findings that female agents are more prone to be verbally abused than male agents. We also saw that the agent with an androgynous visual appearance placed itself between the female and male with respect to verbal abuse. Accordingly we suggest that this kind of agent is an interesting design choice if a concern is to reduce verbal abuse. Most researchers in the field, as well as pedagogues, agree that a mixture of male and female agents in pedagogical software is desirable. But it remains that female agents receive much more abuse than male agents. Notably, to perceive an agent as androgynous-looking does not mean that one does not assign a gender to the agent. Basically all students in the study did so (e.g. reflected in the use of "he" or "she" in the questionnaires). However, it is not the visual image as such that "decides" for them but they are given more freedom to assign a gender on other grounds. In other words, we do not remove gender, nor the possibility of identification via gender [17, 18] – rather the visually androgynous agent introduces more freedom for students to ascribe gender. Furthermore, the androgynous representation does not concur to an idea of *categorical* gender attributes and an associated reinforcement of gender stereotypes, loosening the connection between gender and abuse.

Finally, the students in the study also seemed to be comfortable with the "double-gendered" agent. We are however aware that the results from the present study might be culturally dependent and we are currently conducting cross-cultural studies to see if the results hold in countries like the USA, Korea and Finland, as well as Sweden where the initial study was performed.

References

1. Hoffmann, L., Krämer, N., Lam-chi, A., Kopp, S.: Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 159–165. Springer, Heidelberg (2009)
2. Bickmore, T., Pfeifer, L., Schulman, D.: Relational Agents Improve Engagement and Learning in Science Museum Visitors. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 55–67. Springer, Heidelberg (2011)
3. De Angeli, A., Brahman, S., Wallis, P.: Proceedings of Abuse: The darker side of human-computer interaction. In: Workshop at Interact (2005), `http://agentabuse.org/Abuse_Workshop_WS5.pdf`
4. De Angeli, A., Brahnam, S.: Sex Stereotypes and Conversational Agents. In: Proc. of the Workshop on Gender and Interaction – Real and Virtual Women in a Male World, 8th Int. Conf. on Advanced Visual Interfaces (2006), `http://sherylbrahnam.com/papers/EN2033.pdf`
5. De Angeli, A.: Gender Affordances of Conversational Agents. Seminar held at FBK, November 29 (2011), `http://gosh.fbk.eu/en/node/98`
6. Haake, M., Gulz, A.: Visual Stereotypes and Virtual Pedagogical Agents. Educational Technology and Society 11(4), 1–15 (2008)
7. Veletsianos, G., Scharber, C., Doering, A.: When Sex, Drugs, and Violence Enter the Classroom – Conversations between Adolescents and a Female Pedagogical Agent. Interacting with Computers 20(3), 292–301 (2008)
8. Pareto, L., Haake, M., Lindström, P., Sjödén, B., Gulz, A.: A Teachable Agent Based Game Affording Collaboration and Competition: evaluating math comprehension and motivation. Educational Technology Research and Development (2012)
9. Brophy, S., Biswas, G., Katzlberger, T., Bransford, J., Schwartz, D.: Teachable agents: Combining insights from learning theory and computer science. In: Lajoie, S.P., Vivet, M. (eds.) Artificial Intelligence in Education, pp. 21–28. IOS Press, Amsterdam (1999)
10. Blair, K., Schwartz, D., Biswas, G., Leelawong, K.: Pedagogical agents for learning by teaching: Teachable agents. Educational Technology & Society, Special Issue on Pedagogical Agents (2006)
11. Roscoe, R., Chi, M.: Understanding Tutor Learning: Knowledge-Building and Knowledge-Telling in Peer Tutors' Explanations and Questions. Review of Educational Research 77, 534–574 (2007)
12. Chase, C., Chin, D., Oppezzo, M., Schwartz, D.: Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. J. of Science Education and Technology 18(4), 334–352 (2009)
13. Gulz, A., Haake, M., Silvervarg, A.: Extending a Teachable Agent with a Social Conversation Module – Effects on Student Experiences and Learning. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 106–114. Springer, Heidelberg (2011)
14. Hackett, G., Betz, N.E.: An Exploration of the Mathematics Self-Efficacy/Mathematics Performance Correspondence. J. for Research in Mathematics Education 20(3), 261–273 (1989)
15. Wallace, R.S.: Artificial intelligence markup language (2010), `http://www.alicebot.org/documentation/`

16. Silvervarg, A., Jönsson, A.: Subjective and Objective Evaluation of Conversational Agents. In: Proc. of the 7th Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Barcelona, Spain (2011)
17. Baylor, A., Plant, E.: Pedagogical Agents as Social Models for Engineering: The Influence of Appearance on Female Choice. In: Proc. of Artificial Intelligence in Education (AIED 2005), pp. 65–72. IOS Press, Amsterdam (2005)
18. Rosenberg-Kima, R., Baylor, A., Plant, E., Doerr, C.: Interface Agents as Social Models for Female Students: The Effects of Agent Visual Presence and Appearance on Female Students' Attitudes and Beliefs. Computers in Human Behavior 24(6), 2741–2756 (2008)

Modeling Speaker Behavior: A Comparison of Two Approaches

Jina Lee and Stacy Marsella

University of Southern California,
Institute for Creative Technologies,
12015 Waterfront Drive, Playa Vista, CA 90094, USA
jinal@usc.edu, marsella@ict.usc.edu

Abstract. Virtual agents are autonomous software characters that support social interactions with human users. With the emergence of better graphical representation and control over the virtual agent's embodiment, communication through nonverbal behaviors has become an active research area. Researchers have taken different approaches to author the behaviors of virtual agents. In this work, we present our machine learning-based approach to model nonverbal behaviors, in which we explore several different learning techniques (HMM, CRF, LDCRF) to predict speaker's head nods and eyebrow movements. Quantitative measurements show that LDCRF yields the best learning rate for both head nod and eyebrow movements. An evaluation study was also conducted to compare the behaviors generated by the Machine Learning-based models described in this paper to a Literature-based model.

1 Introduction

Virtual agents are autonomous software characters that support social interactions with human users. One of the main goals in virtual agent research is to emulate how humans interact face-to-face. Virtual agents use natural speech and gestures to convey intentions, express emotions, and interact with human users much as humans use speech and gesture to interact with each other. Communication through a virtual agent's nonverbal behaviors has become an active research area of increasing importance, especially as better graphical representation and control over the agent's body has supported richer and subtler expression.

To realize this, researchers have taken different approaches to author the behaviors of virtual agents that are adaptable and appropriate to the context of the interaction. One of the foremost approaches in modeling nonverbal behaviors is the *Literature-based* approach, including the Nonverbal Behavior Generator (NVBG) [9] and the BEAT system [4]. This approach uses findings from nonverbal behavior research obtained through observation and interpretation of human interactions and builds computational models that operationalize those findings. In many of these studies, researchers analyze the recordings of human interactions and try to manually find regularities in various behaviors including head movements, posture shifts, or gaze movements. One of the major challenges with

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 161–174, 2012.

this approach lies in the fact that the full complexity of the mapping between various behaviors and communicative functions conveyed through the behaviors is not described in the research literature. There are many factors that potentially affect our nonverbal behaviors such as emotion, personality, gender, physiological state, or social context, and revealing the impacts of these factors and the interdependency among them is an extremely challenging task.

Recently, there have been growing efforts to use machine learning techniques as tools to learn patterns of behaviors [1,2,7,10,11,12,14]. In this *Machine Learning* approach, instead of manually trying to find associations between various factors and nonverbal behaviors, automated processes are used to find features that are strongly associated with particular behaviors. Then, one can use those features to train models that will predict the occurrences of the behavior. One advantage of this approach is that the learning is flexible and can be customized to find patterns in specific context. For example, to learn behavior patterns of different cultures, we may train separate models on each culture's data. Another advantage is that since the learning process is automated, we can process vastly larger amount of data in a given amount of time compared to manual analysis. However, obtaining good annotated data is often the greatest challenge in this approach.

The goal of the work described in this paper is two-fold. First, we extend our prior work [10] [12] of modeling speaker head nods using a machine learning approach to investigate whether we can improve the learning. Previously we built hidden Markov models to predict when speaker head nods occur, regardless of their magnitudes. Here we explore additional machine learning techniques and feature sets to predict not only the uniform head nods but also the different nod magnitudes and eyebrow movements. Secondly, we conduct an evaluation study to investigate how the different modeling approaches (Literature-based vs. Machine learning) compare with each other by asking human subjects to rate the perception of a virtual agent through its behaviors.

The following section describes the research on modeling nonverbal behaviors for virtual agents including our previous work using the two different approaches. We then describe the extension to the machine learning approach by exploring different learning techniques and features to learn patterns of uniform nods, nods with different magnitudes, and eyebrow movements. Finally, we present the evaluation study that shows how the behaviors from different modeling approaches are perceived by human users.

2 Related Work

Research on virtual agent has taken different approaches to realize nonverbal behaviors of the virtual agent. Our previous work on the Nonverbal Behavior Generator (NVBG) [9] employs a literature-based approach to generate behaviors according to the communicative functions. The system incorporates a set of nonverbal behavior rules that map from the agents' communicative intentions to various nonverbal behaviors. The communicative intentions are derived from a range of sources, from the agent's cognitive reasoning, dialog processing and emotional state, to linguistic features of the utterance text. For instance, the *Intensification Rule* in NVBG is triggered when the surface text includes intensifying expressions, suggested by words like 'very' or 'quite,' which then generates lowered eyebrow movements and a head nod. However, multiple rules could apply to one text segment, leading to rule conflicts. To resolve those conflicts, rules were prioritized using the frequency counts of feature/behavior correspondence extracted from corpora of human nonverbal behavior.

Others have employed a machine learning approach by using corpora of nonverbal behavior more extensively and developing probabilistic models that find the behavior patterns from data directly. Below we list a number of different machine learning techniques used for modeling nonverbal behaviors, which we also use for the work described in this paper.

Hidden Markov model (HMM) [19] is a statistical model that has been widely used in problems with temporal dynamics such as speech recognition, handwriting recognition, and natural language problems including part-of-speech tagging. A number of work in gesture modeling is based on hidden Markov models. Busso et al. [2] used features of actual speech to synthesize emotional head motion patterns for avatars. Our previous work used various linguistic features to predict speaker head nods and investigates the impact of using affective information during learning [10,11,12].

Conditional Random Fields (CRF) [8] relax HMM's conditional independence assumption and can learn long-range dependencies between input features, making it more suitable for various real-world problems. Morency et al. [14] trained CRFs to predict listener's head movements by using various multi-modal features of the human speaker (e.g. prosody, spoken words, eye gaze) and exploring different feature encoding methods (e.g. binary, step function, ramp function). Sminchisescu et al. [20] also used CRFs to classify human motion activities such as walking, jumping, or running using 2D image features and 3D human joint angle features.

Latent-Dynamic Conditional Random Fields (LDCRF) [15] incorporate hidden state variables, which allows them to learn the substructure of gestures, however, while requiring more time and data to train the models. Morency et al. [15] trained a LDCRF for recognizing head and eye gestures using rotational velocity of the head or eye gaze at a specific time frame. They have also trained support vector machines (SVM), HMM, and CRF and showed that LDCRF models yielded the best performance on visual gesture recognition task.

3 Modeling Head Nods and Eyebrow Movements

In this section, we explore multiple learning techniques and feature sets for learning models of speaker head nods and eyebrow movements. We first present the gesture corpus and features used for training, then describe the probabilistic models we learned and the learning results.

3.1 Gesture Corpus

AMI Meeting Corpus [3] was used in this work, which includes annotations of speaker transcript, dialog acts of utterances and timings of people's head movements in addition to other data (e.g. video data) not used in this work. The original corpus was also extended by our own annotations; first, we manually annotated the dynamics of the nods (small, medium, big) and eyebrow movements (inner brow raise, outer brow raise, brow lowerer). Secondly, we processed the speaker transcript through a text processor to obtain additional features. The following describes the features that were used to train the probabilistic models in detail.

Syntactic Features: Syntactic features include part of speech tags (18 total), phrase boundaries (sentence start, noun phrase start, verb phrase start), and key lexical entities (7 cases). Key lexical entities consists keywords that are shown to have strong correlations with head movements [13]. Some examples include ‘yes’ for affirmation and ‘very’ for intensified expressions.

Dialogue Acts: Dialogue acts describe the communicative functions of each utterance and are extracted from the AMI Meeting Corpus (15 total).

Paralinguistic Features: These features are also obtained from the gesture corpus and includes *gaps*, *disfluencies*, and *vocal sounds*. Gaps are speech gaps during speaking turns, disfluency markers are discontinuity or disfluencies while uttering, and vocal sounds are nonverbal sounds such as laughing, throat noises, or other nonverbal vocalizations.

Semantic Category: The speech transcription was processed through the Linguistic Inquiry and Word Count (LIWC) [18] to obtain the various semantic categories of each word. These categories include psychological construct categories (e.g., affect, cognition, biological processes), personal concern categories (e.g., work, home, leisure activities), paralinguistic dimensions (e.g. assents, fillers, nonfluencies), and punctuation categories (periods, commas, etc.). There are a total of 75 such categories.

Here we define the ***Basic Feature Set*** to include syntactic features, dialogue acts, and paralinguistic features, which are features that are obtained through a shallow parsing of the utterance. To study the impact of word semantics on learning the speaker behavior, we also define the ***Extended Feature Set*** to include the semantic categories in addition to the basic feature set.

3.2 Training Process

The HCRF Library [5] was used to train HMM, CRF and LDCRF models. For each learning technique, separate models were learned for different nods and eyebrow movements (e.g. separate CRF models for general, small, and medium nods). For HMMs and LDCRFs, the number of hidden states tried out are 2-6. Approximately 70% of the annotation data were used as training set and about 30% were used as test set, keeping the annotations of a particular person in either the training set or the test set. The training set was further split to set aside a validation set through a 3-fold cross validation. The training set, test set, and validation set were constructed with a sampling rate of 10Hz.

Table 1. Performance of the nod models

Model Type	*Feature Set*	*F-score*	*Precision*	*Recall*
GENERAL NODS				
HMM	Feature Basic	0.1421	0.0834	0.4788
HMM	Feature Extended	0.1357	0.0763	0.6101
CRF	Feature Basic	0.2575	0.2489	0.2667
CRF	Feature Extended	0.2525	0.2177	0.3004
LDCRF	Feature Basic	0.3002	0.2489	0.3781
LDCRF	Feature Extended	0.2665	0.2196	0.3391
SMALL NODS				
CRF	Feature Basic	0.1148	0.0745	0.2507
CRF	Feature Extended	0.0937	0.0517	0.4991
LDCRF	Feature Basic	0.1404	0.0996	0.2375
LDCRF	Feature Extended	0.1273	0.0750	0.4198
MEDIUM NODS				
CRF	Feature Basic	0.0654	0.0679	0.0631
CRF	Feature Extended	0.0368	0.0189	0.7109
LDCRF	Feature Basic	0.0479	0.0297	0.1248
LDCRF	Feature Extended	0.0423	0.0223	0.4135

3.3 Results

A number of models with different combinations of learning algorithms and feature sets were learned to predict the speaker head nod and eyebrow movements. First the results of predicting general head nods are described. We define *general head nods* as nods regardless of their magnitudes, thus combining all the nods in the original data. Next we present the results of learning nods with different magnitudes and various eyebrow movements.

General Head Nods

Given the differences between HMM, CRF, and LDCRF as described above, the underlying assumption was that CRF models will achieve better performances than HMMs, and LDCRF models will perform better than CRF models. The performance of the learned models were measured through F-score, precision, and recall (see the top entries of Table 1). The HMM model using the basic feature set yielded an F-score of 0.1421. As expected, the CRF models performed better than HMMs, and the LDCRF models performed better than the CRF models. This implies that learning the extrinsic dynamics between nods is important. Furthermore, the results of CRF and LDCRF models using basic or extended feature sets show that the LDCRF models (best F-score: 0.3002) perform better than the CRF models (best F-score: 0.2575), emphasizing the importance of learning the hidden substructure of nodding patterns. However, using the extended features did not seem to have a strong impact in all three learning algorithms.

Since the learning results of HMMs were noticeably lower than those of CRFs and LDCRFs, in the subsequent learning of nod magnitudes and eyebrow movements, we only report the results of CRFs and LDCRFs.

Head Nod Magnitudes

In addition to the general head nods, separate models for different dynamics of the nods (small, medium) were also learned. Among all the nod instances in the data, 53.4% were small nods, 40.5% were medium nods, and 6.1% were big nods. Overall, due to the reduced size of sample points, the performances of the models are not as high as those of the general nod models. For small nod models, the LDCRF model using basic feature set achieved the best F-score (0.1404) similar to the general nod model. The CRF model with basic feature set achieved the best performance for medium nods (F-score 0.0654) with a marginal improvement over the LDCRF model with basic feature set. Learning models for big nods failed due to lack of enough data.

Eyebrow Movements

4 different types of eyebrow models were learned: inner brow raise (AU1) models, outer brow raise (AU2) models, eyebrow raise models (combining AU1 and AU2) and brow lowerer (AU4) models. Results of the learned models are shown in

Table 2. Performance of the eyebrow models

Model	Type	F-score	Precision	Recall
INNER EYEBROW RAISE (AU1)				
CRF	Feature Basic	0.1871	0.1645	0.2170
CRF	Feature Extended	0.1661	0.1690	0.1633
LDCRF	Feature Basic	0.2749	0.1761	0.6265
LDCRF	Feature Extended	0.2066	0.1756	0.2509
OUTER EYEBROW RAISE (AU2)				
CRF	Feature Basic	0.1019	0.0562	0.5452
CRF	Feature Extended	0.1015	0.0566	0.4914
LDCRF	Feature Basic	0.1079	0.0695	0.2411
LDCRF	Feature Extended	0.0977	0.0568	0.3505
EYEBROW RAISE (AU1 or AU2)				
CRF	Feature Basic	0.3280	0.2155	0.6863
CRF	Feature Extended	0.3281	0.2109	0.7389
LDCRF	Feature Basic	0.3421	0.2438	0.5734
LDCRF	Feature Extended	0.3270	0.2345	0.5402
BROW LOWERER (AU4)				
CRF	Feature Basic	0.0133	0.0286	0.0087
CRF	Feature Extended	0.0874	0.0603	0.1585
LDCRF	Feature Basic	0.0770	0.0425	0.4120
LDCRF	Feature Extended	0.0733	0.0416	0.3053

Table 2. The inner brow raise (AU1) models yielded better results than the outer brow raise (AU2) or brow lowerer models (AU4). Given that the inner brow raises and the outer brow raises are hard to distinguish from facial expressions and that the two are strongly correlated, the two data were combined and an eyebrow raise model was learned. This eyebrow raise resulted in an improved performance with an F-score of 0.3421. Except for the brow lowerer model, in all the other cases the LDCRF models resulted in better performances than the CRF models, revealing the importance of learning the hidden substructure of eyebrow movements. Similar to the nod models, the extended feature set did not improve the learning over the basic feature set. The outer brow raise models and brow lowerer models had relatively poorer performances, due to lack of enough data samples in the corpus. The inner brow raise (AU1) movements consisted of 70.7% of all the eyebrow movements in the data, compared to 17.1% and 12.2% for outer brow raise (AU2) and eyebrow lowerer (AU4) movements, respectively.

3.4 Discussion

This section presented extensions to our previous machine learning approach on modeling speaker head nods [10]. Here the focus was on comparing different learning algorithms (HMM, CRF, LDCRF), exploring new features, and learning the dynamics of head nods and expanding the learning to eyebrow movements. As expected the LDCRF models tended to outperform other learning techniques. Specifically, considering models with higher F-scores (above 0.25), the LDCRF models had better results than the CRF models, implying the importance of learning the hidden sub-structure of the nods. HMMs had poorer performance than CRF or LDCRF models, suggesting that there may be long-range dependencies between the input features in modeling head movements and eyebrow movements. However, the extended feature set including additional semantic category did not improve the learning. Possible explanations for this could be that either the extended feature set was not very selective in learning these behaviors or that we need more data for the models to learn the relationships among different features.

4 Evaluation Study

This section presents an evaluation study on how the behaviors generated by different models are perceived by human users. Instead of comparing the behaviors from different probabilistic models, we take a broader view and address the issue of comparing different modeling approaches, namely the Literature-based approach and the Machine Learning approach. More specifically, we ask whether the different modeling approaches have an impact on the user's perception of the virtual agent. However, how to measure people's perception of the agent remains a challenge; simply asking subjects about naturalness, precision, and recall and asking them to make broad judgements about the behaviors, as done in an earlier study [11] raises questions. Recent work has moved to using instruments based on social psychology research [1,6]. In line with this research, here we use a modification of measurements developed by Nass et al. on politeness [17] and Osgood et al. on semantic differential [17].

4.1 Study Design

Hypothesis

The main interest of this study is to investigate how the behaviors generated by the different models impact the perception of the virtual agent. Our previous study [11] comparing the speaker head nods from a Machine Learning-based model to a Literature-based model showed that the general nods from simpler HMMs were perceived to be higher in precision, recall, and more natural. Based on this, we hypothesize that the behaviors from the Machine Learning-based model will receive higher perceptual ratings than the behaviors from the Literature-based model.

Independent Variables

The independent variables are the different modeling approaches from which the behaviors are generated:

- *C1: Probabilistic models described in Section 3*
- *C2: Literature-based model (NVBG) [9]*

Stimuli

Each participant was assigned to one condition (i.e. between-subject design). As stimuli, video clips were generated of a virtual agent displaying head nods and eyebrow movements while speaking an utterance. 23 utterances were initially selected from the ICT's Virtual Human Toolkit [21] utterance set, in which a virtual agent answers questions about the concept of virtual humans, what the Virtual Human Toolkit is, and about himself. To generate the head nods and eyebrow movements for each conditions, there were four different models to choose from (CRF-Basic Feature Set, CRF-Extended Feature Set, LDCRF-Basic Feature Set, LDCRF-Extended Feature Set) for each behavior. Based on past observations of behavior generation methods, human users preferred virtual agents with more behaviors mainly because they made the agent look more alive and less robotic. This suggested using the model that generates the most behaviors, namely the model with the best recall rate among the four different choices. To validate this, a preliminary evaluation was conducted and confirmed that human users indeed preferred the behaviors generated from models with high recall rate[1].

[1] To validate the choice of using models with high recall rate, a preliminary evaluation study was conducted. Fifteen utterances were selected from the utterance set mentioned above and two versions of the videos were created displaying behaviors generated from 1) models with the highest F-score and 2) models with the highest recall rate. 12 participants were recruited and asked to choose the video they preferred in a forced-choice manner. In 13 videos sets out of 15, the video displaying behaviors from models with high recall rate was preferred.

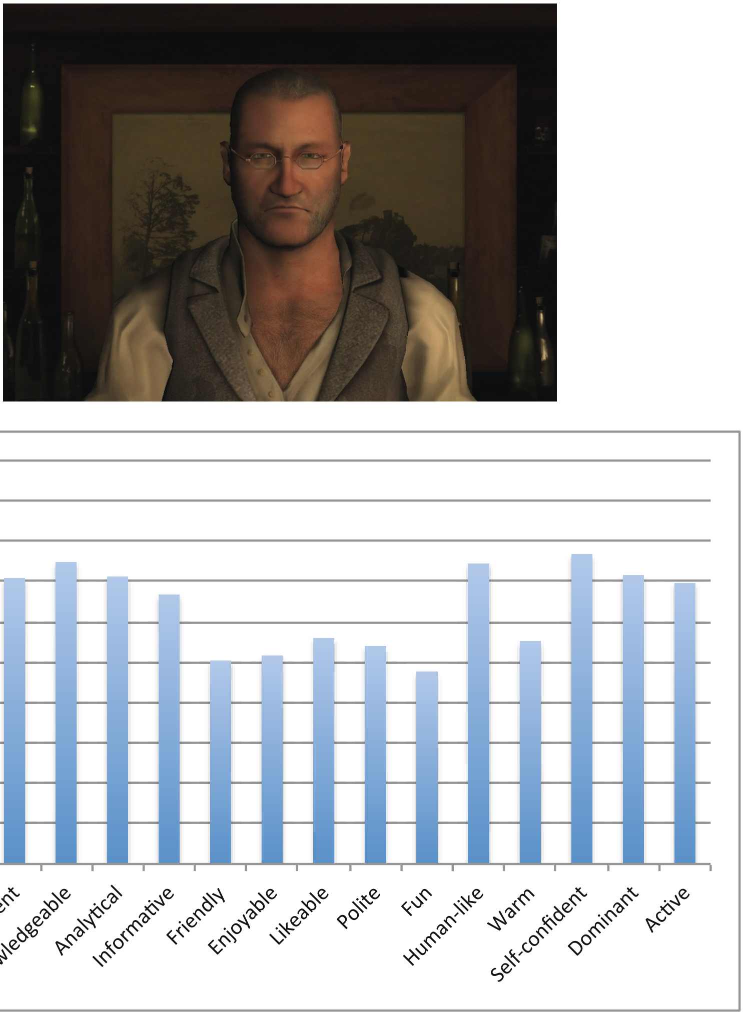


Fig. 1. Virtual agent Utah and the ratings on Utah's personality dimensions from a static image

Among the 23 utterances, 7 utterances with the greatest number of behavioral differences between the probabilistic model and NVBG were selected. In total, there were 14 video clips (7 utterances x 2 conditions) and each participant watched 7 video clips from the experimental condition they were assigned to.

Dependent Variables

To measure participants' perception of the agent through its behaviors, participants were asked to rate the agent on 16 personality dimensions (see Fig. 1) based on the studies of [6] and [1] using a 5-point Likert scale. These dimensions are adopted from the study of Nass et al. [16] used in their politeness study and the semantic differential established by Osgood, Suci, and Tannenbaum [17].

Baseline

To measure the first impression of the virtual agent 'Utah' used in this study, a preliminary study was conducted. A separate set of participants were recruited online (30 males, 20 females, ages ranging from 18 to 65) and were asked to rate him on the 16 dimensions described above on a 5-point Likert scale after seeing a static image of Utah. Fig. 1 shows the mean values of the ratings for each personality dimension. Utah was rated high on *competent, knowledgeable, analytical, informative, human-like, self-confident, dominant*, and *active* but low on *friendly, enjoyable, likeable, polite, fun*, and *warm.* The evaluation results presented in the next section are based on the differences between these means of Utah's initial impression and the ratings of the human subjects to measure the effects of only the modeling conditions and not the appearance of Utah.

Procedure

90 participants were recruited online with 46 males and 44 females and ages ranging from 18 to 65. Participants first filled out a demographic questionnaire asking for their age, gender, education level, ethnicity, and occupation. They were assigned to one modeling condition and watched 7 video clips of the agent speaking a sentence while making head nods and eyebrow movements. Each video clip lasted about 10 seconds. After watching each video, participants were asked to rate the agent on the 16 personality dimensions using a 5-point Likert scale. The order of the video clips and the 16 dimension ratings were randomized for each participant.

4.2 Results

First the means of the personality dimensions obtained in the preliminary study (using Utah's static image) were subtracted from the participants' ratings to provide a more accurate measurement of the perception of the agent due to the behaviors generated by different models.

To measure the reliability between the dimensions, the 16 dimensions were grouped into several factors by conducting a factor analysis. Three factors were extracted, explaining 68.75% of the total variance. Calculating Cronbach's alpha showed that the alpha values for all three factors were above 0.7, justifying combining the dimensions in the same group as a single value (see Table 3). These factors were labeled as Competence, Likeability, and Power.

Independent-samples T-tests were conducted to study the main effect of the modeling approach. The mean values and standard deviations are shown and plotted in Fig. 2. Between the two modeling approaches, there was a significant difference in Power, where the Literature-based model was rated significantly

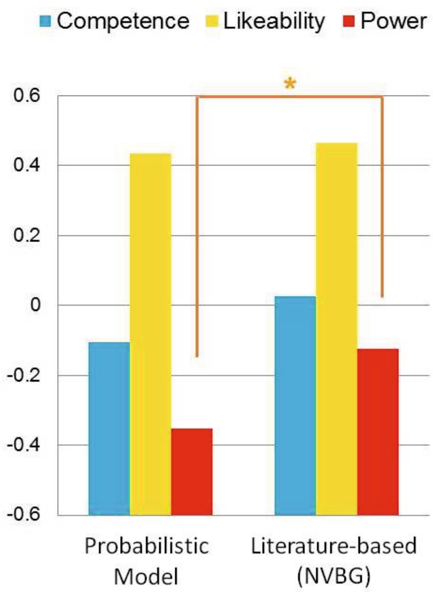
Table 3. Factor analysis for the perception of Utah's personality (principal component analysis with varimax rotation)

Item	Competence	Likeability	Power
Helpful	.805		
Useful	.804		
Competent	.774		
Knowledgeable	.811		
Analytical	.705		
Informative	.764		
Friendly		.826	
Enjoyable		.843	
Likeable		.851	
Polite		.670	
Fun		.765	
Human-like		.502	
Warm		.710	
Self-confident			.764
Dominant			.863
Active			.592
Cronbachs α	*.901*	*.906*	*.790*

higher (t(675.406) = -3.442, p<.01). However, there was no significant difference in Competence and Likeability. Therefore, the hypothesis stating that the Machine learning-based model will outperform the Literature-based model was *not* supported. In fact there was a trend that the Literature-based was rated higher than the Machine learning-based model in Competence, Likability, and Power.

4.3 Discussion

The evaluation result contradicts that of the earlier study [11] that contrasted the Literature-based model with the simpler HMM model, with the HMM being judged superior. There are several possible explanations. The earlier HMM study was attempting to learn just the head nods regardless of their magnitudes, whereas the current study involves head nods of different magnitudes and eyebrow movements as well. These finer behavioral distinctions perhaps made it harder for the models to learn the behavior patterns since there are fewer consistent behaviors across different subjects. In addition, showing more behaviors adds complexity to the evaluation task since there are more factors to consider when judging the perception of the virtual agent. The hypothesis was also setting a tough criteria for the Machine learning-based model. First, it required the Machine learning-based model to be rated higher in all three factors. Perhaps what the hypothesis should have focused on was the Likeability factor, which groups dimensions such as *Friendly, Human-like*, and *Warm*. Comparing the ratings for Likeability between the Machine learning model and the Literature-based model, the two models received similar ratings with a marginal difference. Second, the mappings and behaviors of the Literature-based model (NVBG) were



	Competence	Likeability	Power
C1 Probabilistic Model	-.104 (.913)	.434 (.797)	-.353 (.808)
C2 Literature based Model	.027 (.851)	.465 (.878)	-.125 (.924)

Fig. 2. Mean values of the agent perception factors. The means reflect differences from the agent's initial impression, shown in Fig. 1. Significant differences are identified by * (P<.05) and the exact values shown in the table (standard deviations shown in parenthesis).

distilled from years of social psychology research. Furthermore, it has been constantly modified and refined over the years as NVBG has been incorporated into numerous virtual agent projects, altogether setting a high bar for the machine learning model to beat.

5 Conclusion

In this paper we presented the work on learning probabilistic models to predict speaker head nods and eyebrow movements. We explored different learning algorithms (HMM, CRF, LDCRF) and feature sets to learn when speaker nods occur, as well as to learn the dynamics of head nods and the eyebrow movements. Consistent with our expectations, quantitative results (e.g. F-score) show that the LDCRF models had the best results, implying the importance of learning the dynamics between different gesture classes and the hidden sub-structure of the gestures. However, the extended feature set including additional semantic categories did not improve the learning, perhaps due to the fact that it requires more data to learn the impacts of the additional features.

The evaluation study conducted with human subjects focused on investigating how the behaviors generated by the different models affect the perception of the agent. Contrary to our expectation, the Machine learning-based model did not receive higher ratings than the Literature-based model; the complexity of the behavior sets in the video and the tough criteria to support the hypothesis may explain this result. This demands a follow-up study.

This work could be extended in several ways. Similar probabilistic approaches could be taken to learn patterns of additional behaviors or mappings of different communicative functions. For example, we may customize the learning by training models on data from specific groups of people that convey their status, individual traits, or cultural background. The evaluation study could also be improved by letting the human users *interact* with the virtual agents rather than showing videos of them. In addition, a more comprehensive evaluation is necessary to study the implications of each type of behavior generated under different conditions and what the users infer from each of those behaviors.

References

1. 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)
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. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Language Resources and Evaluation Journal 41(2), 181–190 (2007)
4. Cassell, J., Vilhjálmsson, H.H., Bickmore, T.: BEAT: the behavior expression animation toolkit. In: SIGGRAPH 2001: Proc. of the 28th Annual Conf. on Computer Graphics and Interactive Techniques, pp. 477–486 (2001)
5. HCRF library (including CRF and LDCRF) (2012), http://sourceforge.net/projects/hcrf/
6. Hoffmann, L., Krämer, N.C., Lam-chi, A., Kopp, S.: Media Equation Revisited: Do Users Show Polite Reactions towards an Embodied Agent? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 159–165. Springer, Heidelberg (2009)
7. Kipp, M., Neff, M., Kipp, K.H., Albrecht, I.: Towards Natural Gesture Synthesis: Evaluating Gesture Units in a Data-Driven Approach to Gesture Synthesis. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 15–28. Springer, Heidelberg (2007)
8. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the Eighteenth Int. Conf. on Machine Learning, pp. 282–289 (2001)
9. 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)

10. Lee, J., Marsella, S.: Learning a model of speaker head nods using gesture corpora. In: Proc. of the 8th Int. Joint Conf. on Autonomous Agents and Multiagent Systems (2009)
11. Lee, J., Marsella, S.C.: Predicting speaker head nods and the effects of affective information. IEEE Transactions on Multimedia 12(6), 552–562 (2010)
12. Lee, J., Neviarouskaya, A., Prendinger, H., Marsella, S.: Learning models of speaker head nods with affective information. In: Proc. of the 3rd Int. Conf. on Affective Computing and Intelligent Interaction (2009)
13. McClave, E.Z.: Linguistic functions of head movements in the context of speech. Journal of Pragmatics 32, 855–878 (2000)
14. 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)
15. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition (2007)
16. Nass, C., Moon, Y., Carney, P.: Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. Journal of Applied Social Psychology 29(5), 1093–1109 (1999)
17. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The measurement of meaning, p. 197. University of Illinois Press (1957)
18. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count: LIWC 2001. Word Journal of the International Linguistic Association (2001)
19. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
20. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: Int. Conf. on Computer Vision, pp. 1808–1815 (2005)
21. ICT Virtual Human Toolkit (2012), http://vhtoolkit.ict.usc.edu

An Incremental Multimodal Realizer for Behavior Co-Articulation and Coordination*

Herwin van Welbergen[1], Dennis Reidsma[2], and Stefan Kopp[1]

[1] Sociable Agents Group, CITEC, Fac. of Technology, Bielefeld University
[2] Human Media Interaction, University of Twente
{hvanwelbergen,skopp}@techfak.uni-bielefeld.de, d.reidsma@utwente.nl

Abstract. Human conversations are highly dynamic, responsive interactions. To enter into flexible interactions with humans, a conversational agent must be capable of fluent incremental behavior generation. New utterance content must be integrated seamlessly with ongoing behavior, requiring dynamic application of co-articulation. The timing and shape of the agent's behavior must be adapted on-the-fly to the interlocutor, resulting in natural interpersonal coordination. We present AsapRealizer, a BML 1.0 behavior realizer that achieves these capabilities by building upon, and extending, two state of the art existing realizers, as the result of a collaboration between two research groups.

1 Introduction

Human conversations are highly dynamic, responsive interactions. In such settings, extended utterances are not pre-planned and then executed, but are produced by a speaker incrementally. This incremental delivery enables the speaker, first, to decompose a complicated intended message into a series of fragments and, second, to perceive the addressee's behavior and to adapt a running contribution in order to ensure the success of the communicative activity. Notably, such utterances still exhibit a high degree of synchrony between multimodal behaviors like speech, gestures, or facial expressions [1]. One mechanism with which incremental production of utterances is facilitated is co-articulation between adjacent behaviors. Co-articulation, for example, occurs when the retraction phase of one gesture and the preparation phase of the next one become fused into a direct transition [2]. In this way, co-articulation helps to create (or is a consequence of) fluently connected utterances and increases the flexibility for creating multimodal synchrony.

At the same time, conversation between humans is characterized by interpersonal coordinations on several levels. Bernieri et al. [3] define interpersonal coordination as the degree to which behaviors in an interaction are non random, patterned or synchronized in both timing and shape. They categorize interpersonal coordination in behavior matching or similarity and interactional synchrony. Behavior matching includes mimicry such as interlocutors adapting

* This work was partially supported by the DFG in the Center of Excellence "Cognitive Interaction Technology".

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 175–188, 2012.

similar poses. Interactional synchrony includes alignment of movement rhythm (e.g. alignment of postural sway or breathing patterns), synchronization of behavior (e.g. simultaneous posture changes by a listener and speaker) and smooth meshing/intertwining of behavior (in conversation these are, e.g., smooth turn-taking and backchannel feedback such as head nods or uttering "uh huh").

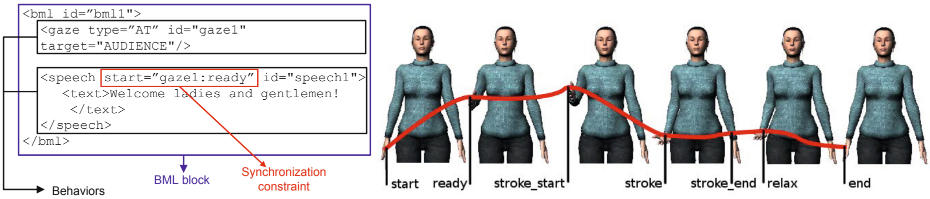
We aim to develop conversational agents that can make use of all these features of human conversational behavior: multimodal synchrony, flexible inter-personal coordination, and fluently connected, incrementally produced utterance with natural co-articulation. Thus far, no single virtual human generation platform features all of these capabilities. Therefore, we have developed AsapRealizer, a BML behavior realizer for virtual humans that builds on two existing realizers that have focused on either incremental multimodal utterance construction [4] or interactional coordination [5] as isolated problems. By combining the realization capabilities for incremental multimodal behavior construction and interactional coordination in a single realizer, we can enable interaction scenarios that go beyond the capabilities of these individual realizers.

A key feature of AsapRealizer is its ability to continuously and automatically adapt ongoing behavior while retaining its original specification constraints. This affords incremental processing in two ways: First, the behavior planner can issue behavior requests early on and then send detailed parameters later on. Second, the realizer can rely on predicted events and then adapt as new, possibly altered status information about these events arrive. An example of the former is a behavior that needs to be fluently connected to an ongoing behavior and might partly co-articulate with it, or when a sudden parameter change (e.g. increase in gesture amplitude) results in changed shape and timing of the ongoing and subsequent gesture phases. An example of the latter is when a behavior is synchronized to an external event (e.g. the interlocutor's head nod), or when behavior execution cannot be reliably predicted (as with a robotic body [6]). In case of updates to the timing of this behavior, continuous adjustment of other, synchronized, modalities may be needed to achieve specified time constraints. The AsapRealizer provides a way to cope with these challenges by interleaving scheduling and execution freely. This paper presents AsapRealizer's design and implementation, explains how it allows for incremental adaptive scheduling of behavior, and demonstrates how its implementation is applied specifically to achieve gestural co-articulation.

2 Related Work

It is increasingly acknowledged that, to achieve a more natural dialog, social agents require incremental (dialog) processing [7]. Such incremental processing enables social agents to exhibit interpersonal coordination strategies such as backchannel feedback, smooth turn-taking, or an alignment of movement rhythm between interlocutors. One prerequisite for this is a flexible behavior generation algorithm that is able to change the timing and shape of ongoing behavior on the fly and in *anticipation* to the movement of an interlocutor [8].

Most existing behavior realizers comply with the SAIBA Framework [9], in which the BML markup language provides a general, realizer-independent description of multimodal behavior that can be used to control a virtual human. BML expressions (see Fig. 1 for a short example) describe the occurrence of certain types of behavior (facial expressions, gestures, speech, and other types) as well the relative timing of the involved actions. Synchronization among behaviors is done through BML constraints, included within a BML block, that link synchronization points in one behavior (like “start”, “end”, “stroke”, etc; see also Fig. 1) to similar synchronization points in other behaviors.



```
<bml id=”bml1">
 <gaze type=”AT” id="gaze1"
 target="AUDIENCE"/>

 <speech start=”gaze1:ready” id="speech1">
    <text>Welcome ladies and gentlemen!
     </text>
 </speech>
</bml>
```

Fig. 1. Left: an example of a BML block. Right: the standard synchronization points of a gesture.

Several BML realizers have been implemented by different research groups [10,11,12,13], while other research groups have expressed their intent to join the SAIBA effort (e.g. the authors of [4,14]). Of these realizers, ACE [4] and Elckerlyc [13] are especially relevant for AsapRealizer, because unlike the other realizers, they are specifically designed to allow incremental behavior construction and interactional coordination, respectively.

For the present work, it is necessary to fluently stream the execution of behaviors from different BML blocks. However, a specification for co-articulation constraints between subsequent BML blocks is currently lacking: BML blocks can be only specified to start instantly, merging with ongoing behavior, or to start after all currently ongoing behavior blocks are completely finished. Furthermore, while BML provides a clear-cut specification of the internal multimodal synchronization of the behavior of a virtual human, it lacks the expressivity to specify the interaction of this behavior with other (virtual) humans. AsapRealizer provides a BML extension *BMLa* to address these shortcomings. To this end, BMLa adopts Elckerlyc’s interactional coordination specification mechanisms (see below). In addition to that, BMLa provides novel mechanisms to specify the combination of BML blocks. These mechanisms provide a Behavior Planner or other authors of a BML stream with specific control over whether or not co-articulation between gestures or other behaviors in two or more BML blocks may occur.

The ACE (Articulated Communicator Engine) [4] realizer was the first behavior generation system that simulated the *mutual* adaptations between the timing of gesture and speech that humans employ to achieve synchrony between co-expressive elements in those two modalities. It also pioneered the incremental scheduling of multimodal behavior for virtual humans. ACE’s incremental speech-gesture production model is based on McNeill’s segmentation hypothesis [2]: speech and gesture are produced in successive *chunks*. Each chunk contains one prosodic phrase in speech and one co-expressive gesture phrase.

Gesture movement between the strokes of two successive gestures (in two successive chunks) depends on their relative timing, ranging from retracting to an in-between rest position, to a direct transition movement. A flexible silent pause in speech was inserted to create enough time for the preparation phase of the second gesture. To achieve this production flexibility, ACE uses an incremental scheduling algorithm that plans part of the chunk in advance and refines it when the chunk is actually started, by setting up each chunk's inter-chunk synchrony with its predecessor. This starting time is decided in a bottom-up process, where the current hand and body positions influence the actual duration and trajectory of the preparation and retraction phases of two adjacent gestures. There is no continuous adaptation in ACE after the initiation of a chunk.

Elckerlyc [5], a more recent realizer, pioneered new ways of BML specification (using its BML extension BML^T) and an implementation of several behavior generation mechanisms that are essential for interactional coordination. These mechanisms include graceful interruption, re-parameterization of ongoing behavior and synchronization to predicted interlocutor behavior. For this, Elckerlyc provides a flexible behavior plan representation that can continuously be modified, while retaining the constraints specified in BML. The construction and representation of Elckerlyc's multimodal plan is discussed in detail in [5]. Thus far, updates to Elckerlyc's multimodal plan were mostly guided through top-down processes (e.g. by specifying them in BML) and by directly aligning the timing of synchronization points to (maybe predicted) interlocutor events.

In sum, the two different realizers provide useful concepts for enabling the kind of flexibility needed to simulate natural coordination and co-articulation in behavior realization. AsapRealizer builds upon both approaches, combines them into a coherent architectural framework and adds new concepts. AsapRealizer's design generalizes ACE's incremental scheduling mechanism to one that supports behavior specification in BML blocks. It further provides the ability to do bottom-up adaptation of ongoing behavior. For example, AsapRealizer implements the bottom-up gesture modification that is employed in ACE, and combines it with Elckerlyc's flexible plan representation and its approach to enable the specification and implementation of interactional coordination.

3 Design Considerations

AsapRealizer should enable both inter-personal coordination and inter-behavioral synchrony through incremental behavior construction and flexible scheduling. As a realizer component within the SAIBA framework, it should be easy to use in many virtual human applications and experiments. To achieve this, AsapRealizer's design satisfies the following requirements:

1. Generate multimodal behavior specified in BML.
2. Generate behaviors incrementally and link increments fluently, with natural co-articulation between the increments.
3. Process underspecified BML behavior specifications that constrain only those features the author or Behavior Planner is really interested in achieving; that

is, keeps all valid realization possibilities open for as long as possible. Figure out unspecified timing or shape (e.g. trajectory, hand shape, amplitude) of a motor behavior in a biologically plausible way.

4. Allow last minute changes in shape and timing of behavior, even when the behavior is currently ongoing; check for and maintain validity of the constraints specified in BML.
5. Enable top-down (through BML) and bottom-up processes (e.g. changing predictions, co-articulation with new behavior) to adapt a behavior.

We make use of several design elements and implementations from both Elckerlyc and ACE to satisfy these requirements. Req. 1 is satisfied by building AsapRealizer on top of the Elckerlyc BML Realizer. To satisfy Req. 2, we have designed novel algorithms for both the specification (in BMLa) and the implementation of incremental construction of behavior using BML blocks. In Section 5.1 we explain these algorithms in detail. Req. 3 is behavior-specific. We adopt mechanisms from ACE to automatically construct preparation and retraction phases of gestures that provide biologically plausible timing. The timing of these gesture retractions and preparations is updated on the fly. Req. 4 and 5 are satisfied by combining Elckerlyc's flexible behavior plan representation with ACE's behavior and chunk state management. The latter allows flexible bottom-up changes of the ongoing behavior plan, while the first assures that all adaptations are subject to the constraints specified in BML. The implementation of this functionality is discussed in Section 4. AsapRealizer also supports all of Elckerlyc's top-down behavior adaptations to achieve interpersonal coordination.

Fig. 2 illustrates how we have incorporated design features from ACE and Elckerlyc into AsapRealizer. We make use of BML to specify behavior, enhanced

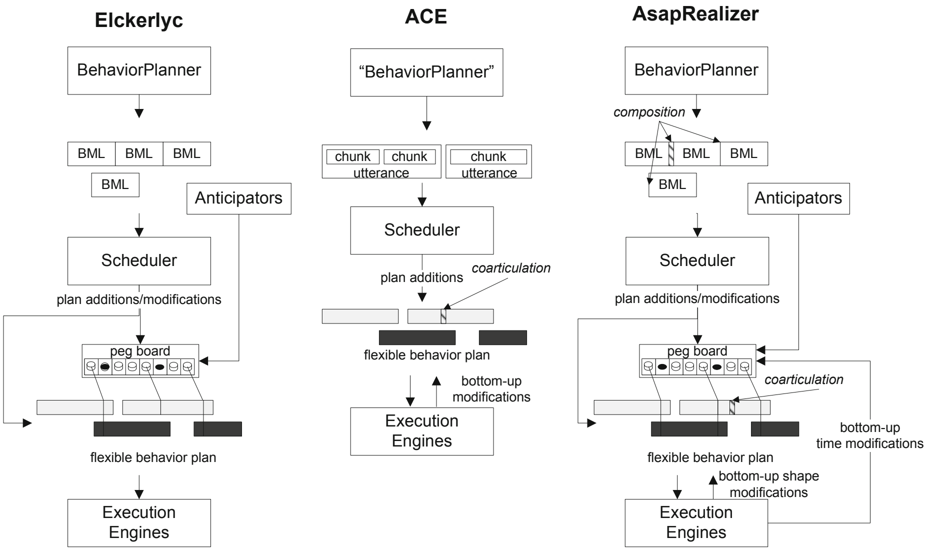


Fig. 2. The design of Elckerlyc, ACE and AsapRealizer

to allow the specification of whether or not behavior co-articulation may occur between BML blocks. The scheduling of specified behaviors results, as in Elckerlyc, in a flexible behavior plan representation –the PegBoard– that allows one to do timing modifications to the behavior plan in such a way that the BML constraints remain satisfied and no expensive re-scheduling is needed. An ExecutionEngine executes the constructed plan and, like ACE, continuously makes modifications to this ongoing plan. These modifications are split into shape modifications that modify the form of behaviors and time modifications that directly act upon the PegBoard. These bottom-up time modifications thus do not invalidate the time constraints as specified in BML. In addition, like Elckerlyc, AsapRealizer can align (and continuously update) the timing of behavior to anticipated events and exert top-down plan or behavior modifications (e.g. interruption and parameter changes in ongoing behavior). Finally, we have implemented a novel gesture co-articulation strategy that extends the strategy used in ACE.

4 State-Based Behavior Scheduling

AsapRealizer realizes a stream of BML blocks, each of which specifies the timing (e.g. sync points X of behavior A and Y of behavior B should occur at the same time) and shape (e.g. behavior A should be performed with the left hand) of the desired behaviors. Generally, BML blocks are under specified and leave realizers freedom in their actual realization. Realizers can make use of this to achieve natural looking motor behavior, e.g. by setting a biologically plausible duration of a gesture preparation. However, most realizers [10,14,11,12] exploit this freedom only for behavior plan construction/scheduling. After scheduling, no more changes can be made and the plan is executed ballistically. AsapRealizer employs Elckerlyc's plan representation [5] that allows changes even when being executed, while retaining the specified BML constraints.

The generation of a BML block and its individual behaviors are managed by two state machines as shown in Fig. 3. These state machines are adopted from Elckerlyc but extended according to the phases of ACE's incremental production model: The behavior state machine (Fig. 3, left) 1) adds a SUBSIDING state, and 2) supports continuous bottom-up adaptation of ongoing behavior using the `updateTiming` function. The BML block state machine (right) 1) provides a SUBSIDING state and 2) has a mechanism to delay the start of a BML block until all its chunk targets are either retracted or finished. These two machines work as follows: Behaviors start out in the IN_PREP state. Once all behaviors of a BML block are scheduled, they move into the PENDING state. This transition is triggered by the central scheduler. At the same time, the block machine moves into the LURKING state. BML blocks can contain ordering constraints that require them to start (fluently) after other blocks. Once these constraints are satisfied, the block moves into the IN_EXEC state. The scheduled plan might already not be the most suitable anymore since the behavior context (e.g. hand position, resting posture) might have changed during scheduling or while waiting

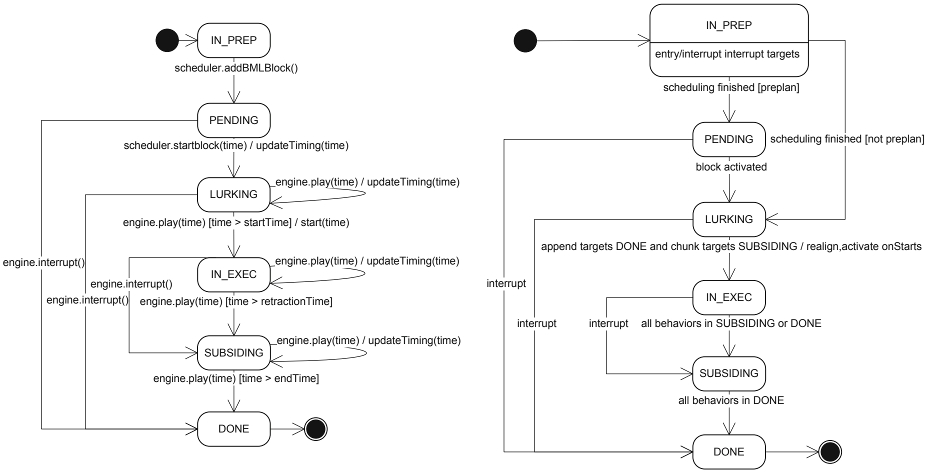

IN_PREP
scheduler.addBMLBlock()
PENDING
scheduler.startblock(time) / updateTiming(time)
LURKING
engine.play(time) / updateTiming(time)
engine.play(time) [time > startTime] / start(time)
engine.interrupt()
IN_EXEC
engine.play(time) / updateTiming(time)
engine.interrupt()
engine.interrupt()
engine.play(time) [time > retractionTime]
SUBSIDING
engine.play(time) / updateTiming(time)
engine.play(time) [time > endTime]
DONE
IN_PREP
entry/interrupt interrupt targets
scheduling finished [preplan]
PENDING
scheduling finished [not preplan]
block activated
LURKING
append targets DONE and chunk targets SUBSIDING / realign,activate onStarts
interrupt
IN_EXEC
interrupt
interrupt
all behaviors in SUBSIDING or DONE
SUBSIDING
all behaviors in DONE
DONE

Fig. 3. The behavior (left) and BML block (right) state machines in AsapRealizer

for other blocks to finish. Therefore this state transition triggers a light-weight behavior realignment step, in which the timing of each behavior is re-evaluated.

Once the new timing is set, all behaviors in the block move to the LURKING state. Behavior state updates are then managed in a bottom-up fashion, through the playback loop of a specific execution engine (e.g. SpeechEngine, AnimationEngine, FaceEngine). Within an animation loop, each playback step is generally executed on an Engine by calling its play function with the current time. This first invokes a timing and shape update on the behavior, using the `updateTiming` function. If the start time of the behavior is greater than the current time, the behavior will be started, moving it to IN_EXEC state. Some behaviors (e.g. gestures) contain a SUBSIDING state. The SUBSIDING state is held while moving the behavior back to a resting position after some meaningful motor behavior was executed within its IN_EXEC state. During the IN_EXEC and SUBSIDING stages, the behavior is executed on the virtual character. While the behavior is being executed, its shape and the timing of its sync points are continuously updated using its `updateTiming` function. Once its end time is reached, the behavior moves to the DONE state. The IN_EXEC, SUBSIDING and DONE phases of a BML block represent the cumulative state of all behaviors in the block. A BML block enters the SUBSIDING state when all of its behaviors are either SUBSIDING or DONE and moves to the DONE state when all its behaviors are DONE. Behaviors and BML blocks may (gracefully) be interrupted at any time after they are scheduled.[1] Interruption can be triggered both top-down (e.g. by specifying in new BML blocks that certain behaviors must be interrupted) or bottom-up (e.g. when the execution of a behavior fails). When

[1] AsapRealizer also provides functionality to forcefully stop a behavior or BML block at any time. This functionality is mainly used to exit or reset the realizer. For clarity reasons it is not show in Fig. 3.

behaviors are interrupted they move into their SUBSIDING phase and are gracefully retracted. The exact implementation of this retraction is behavior-specific; Section 5.2 discusses the implementation used for gesture.

5 Results

AsapRealizer adheres to the new BML 1.0 standard.[2] Compliance to this standard is tested using the RealizerTester framework [15]. We have supported this compliance testing effort by providing several new BML 1.0 test cases.

Through the combination of key features from ACE and Elckerlyc, AsapRealizer provides two main capabilities that go beyond other realizers. Firstly, AsapRealizer improves upon ACE by providing more generic co-articulation mechanisms. Section 5.1 explains how this co-articulation can be specified using BML, as well as how the mechanisms are implemented. Secondly, AsapRealizer provides novel, highly flexible capabilities for specifying and executing graceful interruption of ongoing behaviors, as discussed in Section 5.2.

5.1 Simulating Gesture Co-articulation

AsapRealizer's state-based scheduling and planning allow for simulating interactions between successive behaviors. This, in addition to the interpersonal coordination capabilities of Elckerlyc, makes it suitable for the fluent incremental generation of behavior in which natural co-articulation effects emerge. In this section we demonstrate how the implementation and specification of gesture co-articulation (as shown in Fig. 4) is achieved within our architecture.

Specifying Gesture Co-articulation in BML. The occurrence of gesture co-articulation (or the lack thereof) can well have a communicative function (e.g. marking information boundaries) [1] and is not a matter of simply 'gluing together gestures' in a realizer. Therefore, we allow the Behavior Planner to have control over whether or not gesture co-articulation should occur. This means that some manner of expressing gesture co-articulation has to be provided in BML. We achieve this using the BML `composition` attribute, which allows for merging a BML block into the ongoing behavior plan (i.e. starting it instantly) and for appending a BML block to the behavior plan (i.e. starting it after *all* ongoing behavior). BMLa adds two new composition attribute values that allow us to specify the relation of our block with the ongoing behavior plan in more detail: `append-after(X)` and `chunk-after(X)`. `append-after(X)` specifies the BML block to start after all behavior in the set of BML blocks X is finished. `chunk-after(X)` specifies the BML block to start as soon as all behavior in the set of BML blocks X are either finished or in retraction (i.e. the blocks are SUBSIDING). BML Example 1 illustrates the use of this attribute.

[2] `http://www.mindmakers.org/projects/bml-1-0/wiki`

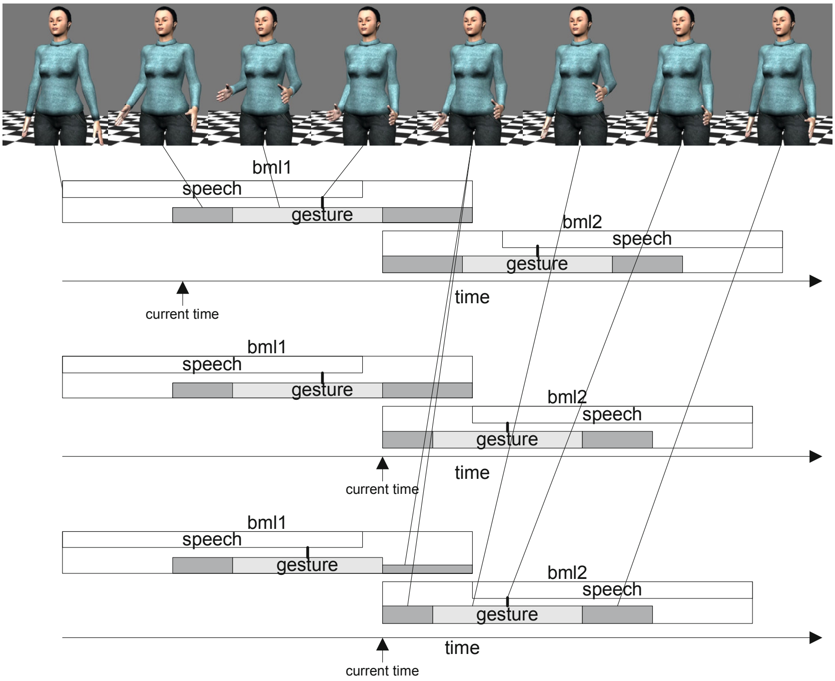


Fig. 4. An example of gesture co-articulation: First BML block `bml1` is being executed and a preliminary plan for `bml2` is being created (top plan graph). As `bml1` is subsiding, `bml2` is re-aligned to fit the current behavior state (middle). This involves shortening the gesture preparation since the hand is still in gesture space. As the gesture of `bml1` is being retracted, it has a lower priority than the preparation of the gesture of `bml2` and is overridden by it (bottom plan graph). Since `bml2`'s gesture acts only on the left hand, a cleanup motion is generated for the right hand part of `bml1`'s gesture.

BML Example 1. Expressing gesture co-articulation in BML.

```
<bml id="bml2" composition="chunk-after(bml1)">
  <speech id="speech1">
    <text>At <sync id="s1"/>6 pm you have another appointment</text>
  </speech>
  <gesture id="gesture1" lexeme="BEAT" stroke="speech1:s1"/>
</bml>
```

Implementation of Gesture Co-articulation. To allow gesture co-articulation between BML blocks, a realizer needs information on when the new BML block can be started and requires an animation system that allows a new gesture to fluently overtake a retracting old gesture. To fulfill the first requirement, AsapRealizer keeps track of each BML block's state using the block state machine (Fig. 3). The state transition from LURKING to IN_EXEC is triggered for a new BML block only if all chunk targets of the block are either SUBSIDING or DONE.

AsapRealizer's AnimationEngine builds upon Elckerlyc's mixed dynamics capabilities [16], allowing a mix of the physical realism provided by physical simulation and the control (in timing and limb placement) provided by procedural animation or motion capture. These capabilities are integrated with the bottom-up re-planning and adaptation provided by ACE. In addition, AsapRealizer provides a novel animation conflict resolution solver that employs a dynamic resting state rather than specifying the resting state only implicitly using the now deprecated BML behavior persistence (as in Elckerlyc) or as a preset joint configuration (as in ACE). Here we highlight how this functionality is employed for gesture co-articulation.

In AsapRealizer's AnimationEngine each behavior is executed using a TimedMotionUnit (TMU, henceforth), which specifies (among other things) the state of its behavior, a priority and the set of skeletal joints it controls. Whenever a TMU is played back, it can set joint rotations, apply a physical controller to the physical part of the body model, or set a new RestingTMU. The latter is a special TMU that creates motions leading into and managing a dynamic 'resting state' of the virtual human. Implementations of a RestingTMU could, e.g., model balanced lower body movement using a physical balance controller or slightly move the body using Perlin noise. There is only one RestingTMU and it is always executed with the lowest priority, i.e., the other TMUs take precedence over it. Such a precedence may be partial (e.g. only on limbs steered by other TMUs).

Gesture co-articulation is achieved using a conflict resolution mechanism within the AnimationEngine. TMUs that execute gesture behaviors automatically reduce their priority when the behavior enters the retraction phase. The AnimationEngine executes TMUs in the order of their priority. Whenever a TMU needs to control joints that are already controlled by higher priority TMUs, this specific TMU is interrupted by the AnimationEngine. Such a TMU might however have previously exerted control upon joints that are not taken over by the higher priority TMUs. These joints need to be gracefully moved back to their resting state. This is automatically taken care of by the AnimationEngine: a new, low priority 'cleanup' TMU is created and inserted in the animation plan.

Bottom-up adaptive timing is crucial in gesture preparation and retraction, since the start position of the preparation, the hand position at the start/end of the stroke and the posture state at the end of the gesture are all subject to change during gesture execution. The hand position may vary by previously executed motion and/or posture changes, the hand position at the start/end of the stroke may vary by parameter adjustments in the gesture, and the rest posture state may change during execution. We have implemented an adaptive timing process for the preparation and retraction of gestures. It makes use of Fitts' law to dynamically determine a biologically plausible duration of the hand movement trajectory from the current position to the hand position at the start of the stroke phase (for the preparation), or from the end of the stroke to the current rest position (for the retraction).

5.2 Graceful Interruption

AsapRealizer allows one to specify graceful interruption of ongoing behavior, including, when desired, replacement behavior that is fluently connected to the interrupted behavior. The mechanisms used for this are based on the handling of graceful interruption in Elckerlyc [17]. In Elckerlyc, this was handled completely in a top-down fashion (see also BML Example 2). Assuming we want to interrupt a BML block `bml1` containing some speech and a gesture `gesture1`, a Behavior Planner using Elckerlyc had to adapt the following interruption strategy:

1. If `gesture1` did not start yet, interrupt everything in `bml1` (BML Example 2a).
2. If `gesture1` is already finished or retracting, there is no need to interrupt it; only interrupt all other behaviors in `bml1` (BML Example 2b)
3. If `gesture1` is currently being executed, replace it by a movement that moves it back to the rest pose (BML Example 2c)

BML Example 2. The specification of graceful interruption in Elckerlyc.

(a) Interrupt all behavior in `bml1`

```
<bml id="yieldturn">
  <bmlt:interrupt id="i1" target="bml1"/>
</bml>
```

(b) Interrupt all behavior in `bml1` excluding `gesture1`.

```
<bml id="yieldturn">
<bmlt:interrupt id="i1" target="bml1" exclude="gesture1"/>
</bml>
```

(c) Interrupt all behavior in bml1. Insert a behavior (relaxArm) that gracefully moves the gesturing arm back to its rest position.

```
<bml id="yieldturn">
  <bmlt:interrupt id="i1" target="bml1"/>
  <bmlt:controller id="relaxArm" class="CompoundController"
    name="leftarmhang"/>
</bml>
```

This verbose interruption strategy requires the Behavior Planner to have knowledge on the state the gesture is in, on the current resting state of the virtual human and on how a graceful interruption behavior can be selected that moves the virtual human towards this state. In AsapRealizer, all this knowledge and functionality is available in the AnimationEngine. Therefore, the Behavior Planner can achieve graceful interruption using BML Example 2a, regardless of the state the gesture is in. This interrupt request is then handled by the AnimationEngine, which generates an automatic retraction motion to the rest pose (using functionality provided by the RestingTimedMotionUnit), if needed. When full control over the exact retraction motion is required, retraction motions can still be specified manually. Such a specified retraction will have a higher priority than the automatically generated one, and as such simply overrule it. The interrupted gesture (that is now in its retraction phase) can also be overwritten by a new gesture, using the AsapRealizer's gesture co-articulation mechanisms.

In summary, the combination of ACE's co-articulation and bottom-up adaptation capabilities, Elckerlyc's top-down interruption specification and any-time adaptability, and the new dynamic pose specification, gives AsapRealizer a novel, highly flexible mechanism for specifying and executing graceful interruption of ongoing behavior, and (when desired) insertion of new co-articulated gestures.

6 Discussion

We have introduced AsapRealizer, a new, BML 1.0 compliant –as tested using the RealizerTester framework [15]– realizer. AsapRealizer's unique capability to continuously and automatically adapt ongoing behavior while retaining its original specification constraints makes it eminently suitable for virtual human applications that require interactional coordination and incremental, fluent behavior generation. Its flexibility is achieved by implementing a fusion of the state of the art multimodal behavior generation features of ACE and Elckerlyc. In this paper, we illustrated how AsapRealizer goes beyond other realizers by discussing its more generic co-articulation mechanisms and its novel, highly flexible capabilities for specifying and executing graceful interruption of ongoing behaviors. AsapRealizer allows us to realize interaction scenarios that go beyond the capabilities of each the individual realizers.

The co-articulation mechanism generalizes ACE's incremental generation of chunks into a mechanism that uses BML blocks as increments instead of chunks –which can also describe many other synchronization possibilities. The interruption scenario illustrates a capability for highly responsive interaction that cannot be realized through either of the contributing systems alone.

Another scenario of responsive interaction enabled by the *combination* of ACE and Elckerlyc is that of interactional coordination. Elckerlyc, and therefore AsapRealizer as well, provides BML specification mechanisms to synchronize the behavior of the virtual human to (anticipated) time events in interlocutor behavior. This functionality has been used to make micro-adjustment to the timing of behavior, for example to align the movement of a virtual fitness trainer to that of the user she exercises with [18], or to delay the start of an utterance after receiving user feedback in an attentive speaker [17]. There are other interaction coordination scenarios that cannot be satisfied by Elckerlyc's time adjustment mechanism alone, but also require (bottom-up) shape adjustments and/or the insertion of new behavior segments. For example, the virtual human could point at an object, wait for the interlocutor to gaze at this object –achieving joint attention– and then retract the pointing gesture and continue speaking. This requires the insertion of a hold motion if the user is not yet gazing at the object as the gesture finishes its stroke, and the automatic, fluent, continuation after a hold motion once joint attention is achieved.[3] Such larger plan adaptations cannot easily be realized with Elckerlyc, since Elckerlyc requires the content and timing of any adaptation to be fully specified in a top-down fashion by the Behavior Planner. AsapRealizer can use its bottom-up adaptation mechanisms to automatically insert fillers and adjust motion shape on the basis of changes in the prediction of time events of interlocutor behavior it synchronizes to. Thus, the combination of Elckerlyc's synchronization to predicted interlocutor events with ACE's bottom-up last minute shape adaptation thus allows us to address an even wider range of interactional coordination scenarios.

References

1. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. In: Key, M.R. (ed.) The Relation of Verbal and Nonverbal Communication, pp. 207–227. Mouton (1980)
2. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press (1995)
3. Bernieri, F.J., Rosenthal, R.: Interpersonal coordination: Behavior matching and interactional synchrony. In: Feldman, R.S., Rimé, B. (eds.) Fundamentals of Nonverbal Behavior. Studies in Emotional and Social Interaction. Cambridge University Press (1991)
4. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15(1), 39–52 (2004)
5. Reidsma, D., van Welbergen, H., Zwiers, J.: Multimodal Plan Representation for Adaptable BML Scheduling. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 296–308. Springer, Heidelberg (2011)
6. Salem, M., Kopp, S., Wachsmuth, I., Joublin, F.: Towards an integrated model of speech and gesture production for multi-modal robot behavior. In: Symposium on Robot and Human Interactive Communication, pp. 649–654 (2010)
7. Schlangen, D., Skantze, G.: A general, abstract model of incremental dialogue processing. Dialogue & Discourse 2(1), 83–111 (2011)
8. Nijholt, A., Reidsma, D., van Welbergen, H., op den Akker, R., Ruttkay, Z.: Mutually Coordinated Anticipatory Multimodal Interaction. In: Esposito, A., Bourbakis, N.G., Avouris, N., Hatzilygeroudis, I. (eds.) HH and HM Interaction. LNCS (LNAI), vol. 5042, pp. 70–89. Springer, Heidelberg (2008)
9. 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)
10. Thiebaux, M., Marshall, A.N., Marsella, S.C., Kallmann, M.: Smartbody: Behavior realization for embodied conversational agents. In: International Foundation for Autonomous Agents and Multiagent Systems, pp. 151–158 (2008)

[3] Goodwin describes another example of such adjustments in conversation: when a listener utters an *assessment* feedback, the speaker, upon recognizing this, will slightly delay subsequent speech (e.g. by an inhalation or production of a filler) until the listener has completed his assessment [19].

11. Heloir, A., Kipp, M.: Real-time animation of interactive agents: Specification and realization. Applied Artificial Intelligence 24(6), 510–529 (2010)
12. Čereković, A., Pandžić, I.S.: Multimodal behavior realization for embodied conversational agents. Multimedia Tools and Applications, 1–22 (2010)
13. 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)
14. Mancini, M., Niewiadomski, R., Bevacqua, E., Pelachaud, C.: Greta: a SAIBA compliant ECA system. In: Troisiéme Workshop sur les Agents Conversationnels Animés (2008)
15. 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)
16. van Welbergen, H., Zwiers, J., Ruttkay, Z.M.: Real-time animation using a mix of physical simulation and kinematics. Journal of Graphics, GPU, and Game Tools 14(4), 1–21 (2009)
17. Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K.P., van Welbergen, H.: Continuous interaction with a virtual human. Journal on Multimodal User Interfaces 4(2), 97–118 (2011)
18. Reidsma, D., Dehling, E., van Welbergen, H., Zwiers, J., Nijholt, A.: Leading and following with a virtual trainer. In: Workshop on Whole Body Interaction. University of Liverpool (2011)
19. Goodwin, C.: Between and within: Alternative sequential treatments of continuers and assessments. Human Studies 9(2-3), 205–217 (1986)

Thalamus: Closing the Mind-Body Loop in Interactive Embodied Characters

Tiago Ribeiro, Marco Vala, and Ana Paiva

INESC-ID and Instituto Superior Técnico, Technical University of Lisbon
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal
tiago.ribeiro@gaips.inesc-id.pt,
{marco.vala,ana.paiva}@inesc-id.pt

Abstract. We present the Thalamus framework, which is based on SAIBA and extends it by adding a perceptual loop. This perceptual loop enables embodied characters to perform continuous interaction. The framework was tested in a case study involving a NAO and an EMYS robots. After showing that our extension works, we point out some issues that were encountered during the development of the case study. We also suggest that the definition of a formal Perception Modelling Language (PML) based on the SAIBA framework can enable SAIBA-compliant embodied characters to perform continuous interaction, while still performing synchronized multimodal behavior based on BML.

Keywords: SAIBA, BML, Continuous Interaction, PML.

1 Introduction

We all imagine a future with robots living around us, behaving and interacting with humans and between themselves. But although fun to imagine, scientists have actually been struggling to create these interactive characters that can operate autonomously. Nevertheless, there have been great efforts in the community, and current research on interactive embodied characters (IEC) has been taking steps towards a unified form of behavior expression. However, for continuous and autonomous interaction, we need the character to be able to react to its environment, and trigger behaviors on that environment.

In this paper we present the Thalamus framework, which abstractly closes the loop between mind and body of an IEC. We close the loop by adding the ability to receive perceptions from the character's embodiment and send them up to the mind, in order to allow for continuous interaction, while maintaining an expressive system based on BML [9, 6].

Therefore, the two main contributions of this framework are: a) support for any kind of embodiment, virtual or robotic; 2) acting as an abstract interface to the character's sensors.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 189–195, 2012.

2 Related Work

Our work builds on the SAIBA framework [3, 9], shown in Figure 1. SAIBA is a representational framework for unified multimodal behavior generation. One of the cores of SAIBA is the Behavior Modelling Language (BML) [6].

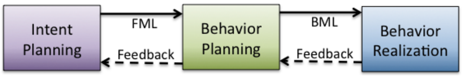


Fig. 1. The three stages of behavior generation in the SAIBA framework and the two mediating languages FML and BML. [9, 6]

Several BML realizers have been developed throughout the community. Greta [4] is a virtual ECA that follows the three-level architecture of SAIBA along with BML. [5] has extended Greta's architecture by adding the ability to communicate with a NAO robot instead of the virtual character. They separate the Behavior Realizer in two sub-layers: Keyframe Generator which is common for both agents, and Animation Generator, which is specific to the embodiment.

Kipp et al. have also proposed that the Realization phase of the SAIBA framework should be separated into Realization Planning, and Presentation [2]. In their architecture, BML serves as input to the Realization Planning, just like on Greta's Keyframe Generator layer. However, the Realization Planner produces EMBRScript [1], which is an executable animation script that is sent into the Presentation module, which controls a 3D character

Smartbody [8] can also be used to control any virtual humanoid character. BML is given as input to a Behavior & Schedule Manager, which produces and runs the plan.

Elckerlyc [10] was developed as a more flexible BML realizer. A first stage parses the BML blocks and schedules them. The scheduler builds a plan that acts like a peg board. Whenever a behavior is scheduled, each sync point is solved and placed in a slot of the peg board so that sync points that should be executed at the same time are placed together. This allows for continuous interaction, because it is possible to change the plan after it has been scheduled, by modifying the placement of the syncpoints in the pegs.

We find that the current state of the art leads to being able to control characters independently of their embodiment, and to be able to continuously interact with them.

3 Thalamus Framework

The Thalamus Framework is a cross-media body interface for multiple simultaneous artificial embodied characters. It supports and is largely based on the architecture of BML. Having a framework that can act as an abstract interface to the character's sensors is especially important when dealing with robotic characters, and follows on the proposal by [12] of having a tight feedback loop between the embodiment and the behavior planner.

We also follow the trend of dividing SAIBA's Behavior Realization level in two sub-layers, as can be seen in Figure 2. The first one is the Behavior Scheduling, which actually keeps in line with [2, 5]. The second sub-layer of our Realization level is the Body Execution.

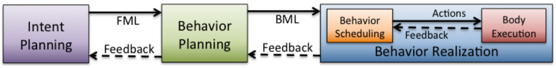


Fig. 2. Our subdivision of SAIBA's Behavior Realization layer into the Behavior Scheduling and Body Execution sub-layers

3.1 Structure

A Character in Thalamus is composed of a mind interface and a body interface, as can be seen in Figure 3. BML blocks are sent via the mind into the character, and the character sends them for scheduling to the plan. The scheduling process solves the sync points of each behavior in order to create *ActionEvents* for it in the *Eventline*. When an *ActionEvent* is activated, it launches the execution of the corresponding behavior. This behavior will call the corresponding actions in the *BodyInterface*, with the correct parameters.

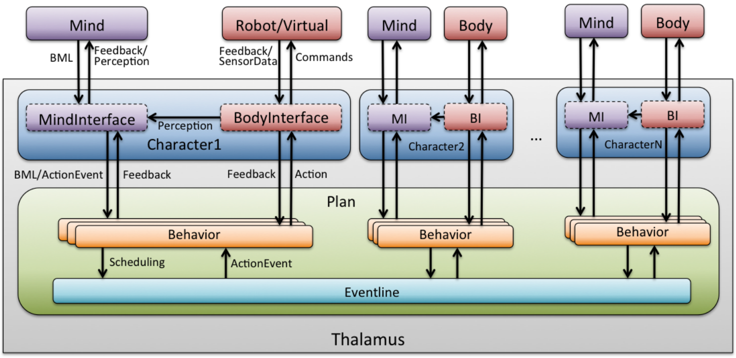


Fig. 3. The Thalamus Framework architecture

3.2 Behavior Scheduling

The scheduling process is inspired by Elckerlyc's peg-board mechanism [10], which we call *Eventline* in our architecture. The *Eventline* contains slots that relate *Events* to *ActionEvents*. An *Event* can be an absolute time instance, like *time=1*, a *SyncPoint* from BML, or any external event that is sent to the plan and that does not originate in BML.

It is important to emphasize the fact that an event is not sent when the plan executes the behavior, but only when the character acknowledges that it has actually started. This is very important with robots, as they usually have some delay between the request for executing an action, and actually starting to execute it.

Our *Eventline* makes it possible for continuous interaction, as the behaviors are not hard-constrained on a timeline, just like in Elckerlyk [10]. Conflicts and overlaps are managed by the usual BML mechanisms.

3.3 Body Interface

There may be several different *Characters*, each with its own *BodyInterfaces*, specific to different embodiments. All *BodyInterfaces* follow the same interface, so they can implement the same set of routines, thus allowing the behaviors to call them regardless of the embodiment they represent.

However, if a behavior tries to call a routine that has not been implemented in a specific *BodyInterface*, it will report back to the plan and mind that it failed.

Body Events. Besides implementing the necessary set of routines for executing BML behaviors, the BodyInterface also supports receiving events. These events can be BML events, non BML events, or sensory events.

The BML events are used for the *BodyInterface* to notify the plan about the executed behaviors. These events are, for example, *SpeechStart*, *SpeechEnd*, *FaceLexemeStart*, *FaceLexemeEnd*, etc.. There is also a *SyncPoint* event for notifying about a specific syncpoint, which is useful for the <*Sync..*> tags in BML *Speech* nodes. This *SyncPoint* event can also be used to send specific non BML events to the *EventLine*.

The *BodyInterface* also supports perception events, originated by the embodiment's sensors. These are sent by the *BodyInterface* to the *Character* in the form of a *Perception* structure which is represented in Figure 4. The perception has an *Id*, which is a unique identifier for each perception, and a *Type*.

The currently supported perceptions were defined for our scenario. They can be of type *SoundLocated*, *SensorTouched*, or *VisionObjectDetected*. Each parameter is composed of a *Name* and a *Value*. The *Name* is a *String*, while the value can currently be an *Integer*, a *Float*, a *Boolean* or a *String*. This list can furtherly be extended.

The list of supported or required parameters depends on the type of the perception. Taking a *SoundLocated* perception as example, it can have parameters

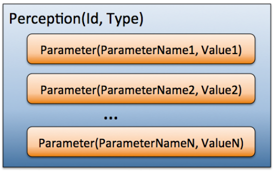

Perception(Id, Type)
Parameter(ParameterName1, Value1)
Parameter(ParameterName2, Value2)
...
Parameter(ParameterNameN, ValueN)

Fig. 4. The Perception structure. Each perception contains a Name, a unique Id, and a set of Parameters.

"Angle":float, and *"Intensity":float*, so that the character may know where the sound came from, and how loud it was. A *SensorTouch* perception can have parameters *"SensorName":string* and *"State":bool*, so that it may know which sensor was touched, and if it was actually touched or released (touched would mean a *True "State"*, while released would mean a *False "State"*).

Mind Events. The perceptions that are generated by the environment are sent through the *Character* into the mind. This way the mind can take appropriate action.

The mind may be a deliberative mind with intent and behavior planning, or even a simple reactive mind which can just immediately react to the events it receives from the body [11]. The only requirement on the mind is that it must also implement an interface we call *MindInterface*, which is capable of maintaining bidirectional communication with the *Character*.

Currently, the *MindInterface* supports a) receiving *Perceptions* and behavior execution *Feedback*, b) requesting the execution of pre-loaded behaviors (stored in .bml files) by their Id, c) sending BML code to the *Character*, and d) request the execution of an *ActionEvent* by the *Character*. This last feature is very interesting to support continuous interaction and interruption of behavior execution, as it makes it possible for the mind to interfere on the scheduled plan.

4 Case study: The Path of NAO

To test our framework, we created a scenario[1] in which two completely different robots interact with each other by running a set of BML scripts, while also interacting with the environment through their sensors. The robots used are a NAO robot[2] and an EMYS robot [7].

The EMYS robot is a robotic head, that can speak, gaze and perform facial animations. It currently has only a sound location sensor that is accomplished by a Microsoft Xbox Kinect™. It will perform *Speech* and *Face-Lexeme* behaviors. The perception its mind will react to is *SoundLocated*.

The NAO robot is a humanoid robot that can walk and perform body animations. It also has lots of sensors that can be used to interact with the environment. It will perform *Locomotion*, *Speech* and *Posture-Pose-Lexeme* behaviors. The perceptions its mind will react to are *VisionObjectDetected* and *SensorTouched*.

Beause we have not implemented a Behavior Planner that could generate BML for this scenario, we have previously written it as BML blocks which are preloaded and scheduled into the plan. Each of these characters has a very simple reactive mind, which, on reception of each perception, send an *ActionEvent* back through its *Character* and into the plan. This ActionEvent can interrupt the current behaviors, and eventually trigger new ones.

[1] The full scenario is shown in the video that accompanies this paper.

[2] `www.aldebaran-robotics.com`

4.1 Discussion

We found our framework to be capable of executing the aforementioned scenario. However, there were some issues that we noted and are worth mentioning.

On the NAO robot's side, we found some complications in having accurate and responsive control both over the robot's actions and sensors. Sometimes the robot raises an internal event stating that the animation has started when in fact it hasn't. That triggers a *start SyncPoint* in the plan, which in turn, triggers a speech that should start synchronized with the animation, but that actually starts before it.

It is clear that robotic animation systems may have these kind of flaws. We thus suggest robotic control system developers to work more closely with users and high-level developers, by looking at this kind of needs.

Taking another example, EMYS' control system was developed by us, and therefore, matches our needs in a higher level. The result is accurate control over it's behavior: when we play or stop and animation, it responds immediately, and sends accurate events back to the character's body interface.

As to using sensors to trigger BML behaviors, we sometimes encountered false positives, both due to noise in the sensor's circuitry and also due to misintepretation of perceptual data. On sound location sensors, for example, echoes might introduce noise in the readings. NAO's head touch sensor however, actually suffers from noise, and frequently triggers events without having been touched.

We therefore had to filter the perceptions in the mind level, before reacting to them.

However, in this kind of framework, it would be useful to be able to filter out false perceptions before they reach our character. After having solved the sensor-imperfection related problems, the scenario ran correctly, except for ocasional delays on NAO's motion response.

5 Conclusions and Future Work

We have developed the Thalamus framework, which is based on the SAIBA framework, but extends it in order to support a perceptional loop for virtual or robotic embodied characters. The perceptional loop can interact with the BML behavior loop, in order to provide continuous interaction based on the SAIBA framework. In order to process the perceptual data in our framework, we have abstracted the data from the sensors into a generic perceptual structure. This perceptual structure was shown to be adequate for the perceptual loop, however it currently lacks formal specification.

We conclude by suggesting the definition of a formal Perception Modelling Language (PML) which meets our extended SAIBA architecture and can interact with BML. Our results show that a widespread specification of PML may provide current embodied characters with generalized continuous interaction capabilities, thus closing the interactive loop.

Acknowledgements. This work was supported by the EU FP7 ICT-215554 project LIREC (LIving with Robots and intEractive Companions). This research was supported by EU 7th Framework Program (FP7/2007-2013) under grant agreement no. 215554, FCT (INESC-ID multiannual funding) through the PIDDAC Program funds.

References

[1] Heloir, A., Kipp, M.: EMBR: A realtime animation engine for interactive embodied agents. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, vol. 1, pp. 1–2 (September 2009), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5349524

[2] Kipp, M., Heloir, A., Schr, M.: Realizing Multimodal Behavior: Closing the gap between behavior planning and embodied agent presentation. Framework (2010)

[3] Kopp, S., Krenn, B., Marsella, S., Marshall, A.N.: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. Information Sciences (2006)

[4] Mancini, M., Niewiadomski, R., Bevacqua, E., Pelachaud, C.: Greta: a SAIBA compliant ECA system. Language (2008)

[5] Niewiadomski, R., Obaid, M., Bevacqua, E., Looser, J., Le, Q.A., Pelachaud, C.: Cross-media agent platform 1(212), 11–20 (2011)

[6] Reidsma, D., Welbergen, H.V.: BML 1.0 Standard, http://www.mindmakers.org/projects/bml-1-0/wiki/Wiki

[7] Ribeiro, T., Paiva, A.: The Illusion of Robotic Life Principles and Practices of Animation for Robots. In: HRI 2012, vol. 1937 (2012)

[8] Thiebaux, M., Rey, M., Marshall, A.N., Marsella, S., Kallmann, M.: SmartBody: Behavior Realization for Embodied Conversational Agents. Information Sciences (Aamas), 12–16 (2008)

[9] Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: 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)

[10] Welbergen, H., Reidsma, D., Ruttkay, Z.M., Zwiers, J.: Elckerlyc. Journal on Multimodal User Interfaces 3(4), 271–284 (2010), http://www.springerlink.com/index/10.1007/s12193-010-0051-3

[11] Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley and Sons (2002)

[12] Zwiers, J., van Welbergen, H., Reidsma, D.: Continuous Interaction within the SAIBA Framework. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 324–330. Springer, Heidelberg (2011)

Lip-Reading: Furhat Audio Visual Intelligibility of a Back Projected Animated Face

Samer Al Moubayed, Gabriel Skantze, and Jonas Beskow

KTH Royal Institute of Technology,
Department of Speech, Music and Hearing. Stockholm, Sweden
{sameram,skantze,beskow}@speech.kth.se

Abstract. Back projecting a computer animated face, onto a three dimensional static physical model of a face, is a promising technology that is gaining ground as a solution to building situated, flexible and human-like robot heads. In this paper, we first briefly describe *Furhat*, a back projected robot head built for the purpose of multimodal multiparty human-machine interaction, and its benefits over virtual characters and robotic heads; and then motivate the need to investigating the contribution to speech intelligibility *Furhat*'s face offers. We present an audio-visual speech intelligibility experiment, in which 10 subjects listened to short sentences with degraded speech signal. The experiment compares the gain in intelligibility between lip reading a face visualized on a 2D screen compared to a 3D back-projected face and from different viewing angles. The results show that the audio-visual speech intelligibility holds when the avatar is projected onto a static face model (in the case of *Furhat*), and even, rather surprisingly, exceeds it. This means that despite the movement limitations back projected animated face models bring about; their audio visual speech intelligibility is equal, or even higher, compared to the same models shown on flat displays. At the end of the paper we discuss several hypotheses on how to interpret the results, and motivate future investigations to better explore the characteristics of visual speech perception 3D projected faces.

Keywords: Furhat, Talking Head, Robot Heads, Lip reading, Visual Speech.

1 Introduction

During the last two decades, there has been on-going research and advancement in facial animation. Researchers have been developing human-like talking heads that can engage in human-like interactions with humans (Beskow et al. 2010) realize realistic facial expressions (Gratch et al. 2006; Ruttkay & Pelachaud, 2004), express emotions (Pelachaud, 2009; De Melo & Gratch, 2009) and communicate behaviours (Granström & House, 2007; Gustafson et al. 2005; Kopp et al. 2005). In addition to the human-like nonverbal behaviour implemented in these heads, research has also taken advantage of the strong relation between lip movements and the speech signal, building talking heads that can enhance speech comprehension when used in noisy environments or as a hearing aid (Massaro, 1998; Salvi et al. 2009).

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 196–203, 2012.

Several talking heads are made to represent personas embodied in 3D facial designs (referred to as ECAs, Embodied Conversational Agents (Cassel et al. 2000)) simulating human behaviour and establishing interaction and conversation with a human interlocutor. Although these characters have been embodied in human-like 3D animated models, this embodiment has almost always been displayed using two dimensional display (e.g. flat screens, wall projections, etc.) having no shared access to the three dimensional environment where the interaction is taking place. 2D displays come with several illusions and effects, such as the *Mona Lisa gaze effect*. For a review on these effects, refer to (Todorovi, 2006).

In robotics on the other hand, the accurate and highly subtle and complicated control of digital computer models (such as eyes, eye-lids, wrinkles, lips, etc.) does not easily map onto mechanically controlled heads. Such computer models require very delicate, smooth, and fast control of the motors, appearance and texture of a mechanical head. In addition to that, mechatronic robotic heads, in general, are significantly heavier, noisier and demand more energy and maintenance compared to their digital counterpart, while they are more expensive and exclusive.

To bring the talking head out of the 2D display, and into the physical situated space, we have built *Furhat* (Al Moubayed et al, 2012a). *Furhat* is a hybrid solution between animated faces and robotic heads. This is achieved by projecting the animated face, using a micro projector, on a three dimensional plastic mask of a face. This approach has been shown to deliver accurate situated gaze that can be used in multiparty dialogue (Al Moubayed et al, 2012b; Edlund et al. 2011). It has also been shown to accurately regulate and speed up turn-taking patterns in multiparty dialogue (Al Moubayed & Skantze, 2011). *Furhat* relies on a state-of-the-art facial animation architecture that has been used in a large array of studies on human verbal and non-verbal communication (e.g. Siciliano et al. 2003; Salvi et al. 2010; Beskow et al. 2010). Figure 1 shows several snapshots of the *Furhat* head.

The question we address in this work is whether this solution comes with negative effects on the readability of the lips. Since the mask is static, jaw and lip movements are merely optical and might not be perceived as accurately as in a flat display due to that the physical surface they are projected onto (the jaw and lips) is not moving according to their movements. The other question is whether the contribution of the lip movements to speech intelligibility is affected by the viewing angle of the face. In 2D displays, the visibility of the lips is not dependent on the viewing angle of the screen (the location of the looker in relation to the screen), due to the Mona Lisa effect, and hence, if there is an optimal orientation of the face, it can be maintained throughout the interaction with humans, something that cannot be established with 3D physically situated heads.

2 Lip-Reading and Speech Perception – Evaluation of Furhat

One of the first steps in building a believable, realistic talking head is to animate its lips in synchrony with the speech signal it's supposed to be producing. This is done not only to enhance the illusion that the talking head itself is the source of the sound

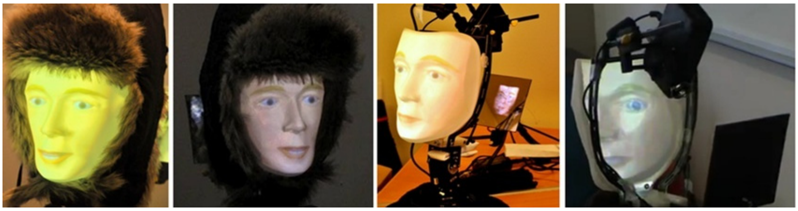


Fig. 1. photos of how Furhat looks like from the inside and outside

signal the system is communicating (rather than a separate process), but also for the crucial role lip movements play in speech perception and comprehension.

The visible parts of the human vocal tract carry direct information about the sounds the vocal tract is producing. The information the lips carry can be perceived by humans and hence help communicate the information in the speech signal itself (Summerfield, 1992; McGurk & McDonald, 1976). These important advantages of lip movements have been taken into account since the early developments on talking heads, and different models and techniques have been proposed and successfully applied to animate and synchronize the lips with the speech signal itself as input (e.g. Beskow, 1995; Massaro et al. 1999; Ezzat & Poggio, 2000).

However, when it comes to Furhat: Furhat's plastic mask itself is static, although the projected image on Furhat is animated, the fact that the mask itself is static might introduce inconsistency and non-alignment between the projected image and the projection surface, and so the fact that the animated lips do contribute to speech perception does not need to naturally hold with Furhat. The following study presents an experiment comparing audiovisual speech intelligibility of Furhat against the same animated face that is used in Furhat but visualized on a traditional flat display.

3 Lip-Reading Experiment

The setup used in this experiment introduces subjects to acoustically degraded sentences, where the content of the acoustic sentence is partially intelligible when listening only to the audio. The sentences are then enriched by a lip-synchronized talking head to increase their intelligibility. The sentences are presented in different experimental conditions (6 in total) and the perceived intelligibility of the sentence (the number of correct words recognized) is compared across conditions.

In the experiment, the audiovisual stimuli consisted of a collection of short and simple Swedish sentences, which vary in length between three to six words, with a basic everyday content. e.g., *"Den gamla raven var slug"* (The old fox was cunning).

The audio-visual intelligibility of each sentence was calculated as *the number of words correctly recognized, divided by the number of content words in the sentence*.

The speech files were force-aligned using an HMM aligner (Sjolander, 2003) to guide the talking head lip movement using the phonetic labelling of the audio file.

The audio signal was processed using a 2-channel noise excited vocoder (Shannon et al. 1995) to reduce intelligibility. This vocoder applies band-pass filtering and replaces the spectral details in the specified frequency ranges with white noise.

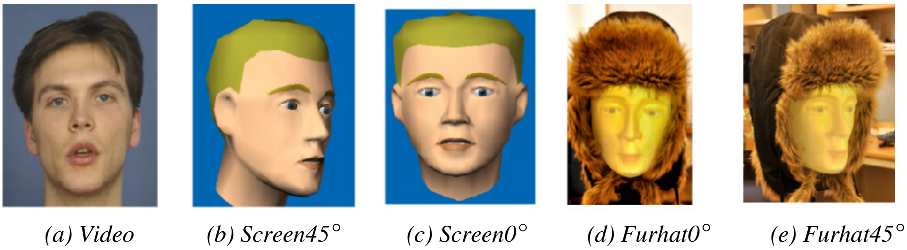

(a) Video (b) Screen45° (c) Screen0° (d) Furhat0° (e) Furhat45°

Fig. 2. Snapshots of the different conditions of the visual stimuli

The number of channels was decided after a pilot test to ensure an intelligibility rate between 25% and 75%, as to avoid any floor or ceiling recognition rate effects.

The stimuli were grouped into a set of 15 sentences per group and every set was only used for one condition. The groups were randomly matched to the conditions for each speaker in order to avoid interaction effects between the sentence difficulty and the condition. As a result, each subject was introduced to all the conditions, but was never introduced to the same stimulus more than once. At the beginning of the experiment, one set was always used as training and only in audio mode, as to avoid any training effects during the experiment. During training, subjects were allowed to listen to the degraded audio file as many times as they wished, and feedback was given to them with the correct content of the audio sentence.

3.1 Conditions

Figure 2 shows snapshots of the stimuli associated with the conditions.

1. ***Audio Only:*** In the audio-only condition, subjects were listened to the acoustically degraded sentences without any visual stimuli.
2. ***Screen0°: Talking head on a flat screen viewed at 0° angle:*** In this condition, the animated face was presented to the subjects along with the acoustic signal. The subject is seated in front of the screen, looking straight at the talking head. The talking head in the screen is oriented to look frontal (0 degrees rotation inside the screen), and hence the name *Screen0°*.
3. ***Furhat0°: Furhat viewed at 0° angle:*** In this condition the sentences were presented to the subject with the animated model and back projected on *Furhat*. The subjects were seated frontal to *Furhat*.
4. ***Screen45°: Talking head on a flat screen viewed at 45° angle:*** This condition is identical to the *Screen0°* condition, except that the head is rotated 45° inside the screen. This condition is designed to compare the audio-visual intelligibility of the sentences with the condition *Screen0°* and *Furhat45°* (see further, condition 5).
5. ***Furhat45°: Furhat viewed at 45° angle:*** This condition is identical to *Furhat0°*, except that subjects were seated at a 45° from *Furhat*. The viewing angle is hence identical to the one in condition *Screen45°* (condition 3). This condition is meant to compare to *Screen45°* condition, except for the projection surface.

6. ***Video:*** In this condition, subjects were presented with the original video recordings of the sentences, viewed on the same flat display used to show the agent, and the size of the face was scaled to match the size of the animated face.

The conditions were systematically permutated among 10 normal hearing subjects, with normal or corrected to normal vision. All subjects were native speakers of Swedish. During the experiments, the subjects were introduced to all conditions but never to the same sentence twice. Since every condition contained 15 sentences, this resulted with 900 stimuli in total for the experiment (6 condition * 15 sentences * 10 subjects), with every condition receiving 150 sentence stimuli. Subjects were given a cinema ticket for their participation, and the experiment took ~25 minutes/ subject.

4 Analysis and Results

An ANOVA analysis was carried out on the sentence recognition rate (accuracy rate) as a dependent variable and the condition as an independent variable. The test shows a significant main effect [F(5)=21.890, p<.0001]. The mean accuracy for each condition is shows in Figure 3, along with the standard error bars.

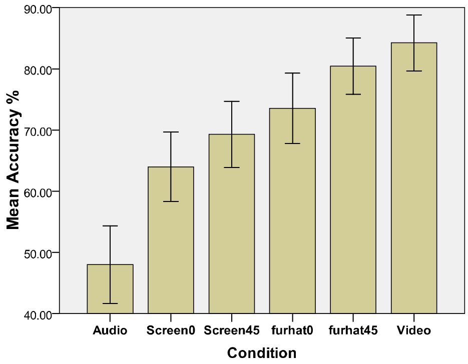


Fig. 3. The average percentage accuracy rates for the different experimental conditions

A post-hoc LSD analysis was carried out to measure the significance values between the accuracy rates of each of the conditions. The p values for the conditions according to the means are shown in Table 1. All other combinations not included in the table are significantly different from each other.

The results firstly show that the Screen0 condition (and all other conditions), provide an audio visual intelligibility that is significantly higher than the audio condition alone. The results also show that there is no significant difference in the audio-visual intelligibility of the face being looked at either frontal or at a 45° (no significant difference between Screen0 and Screen45, or between Furhat0 and Furhat45). The results show that the Mona Lisa effect would not benefit the audio-visual intelligibility of the face using a flat over a spatially situated head, at least not between 0 and 45° rotation angles.

More importantly, the results show that there is no loss in the audio-visual intelligibility when using the Furhat‘s physically-static mask as a projection surface compared to using a flat screen, for either 0 or 45° viewing angles of the face. This shows that the Furhat robot head is a valid alternative to the screen in terms of lip readability, and would be a possible interface to aid human speech perception and comprehension. The more surprising finding is that Furhat, not only does not hinder the audio-visual intelligibility of the animated mask, but rather enhances it significantly over the flat display, and for both viewing angles (the rate is significantly higher for Furhat0 over Screen0, and for Furhat45 over Screen45).

Table 1. p-values from the significance test for all combinations of the different conditions.

Condition1		**Condition2**	***p*-value**
Screen0	*	Screen45	.167
Screen45	*	Furhat0	.266
Furhat0	*	Furhat45	.079
Furhat45	*	Video	.335
All other	**combinations**		**< *.01***

5 Discussion and Conclusions

In the design of the Furhat mask, the details of the lips were removed and substituted by a smooth protruded curvature in order to not enforce a static shape of the lips. Because of this the size of the lips, when projected, is perceived slightly larger than the lips visualized on the screen. This enlargement in size might be the reason behind the increased intelligibility. Another possibility is that looking at Furhat is cognitively easier than looking at a flat display since it is spatially situated and more human-like than a virtual agent presented on a 2D screen.

A main difference between interacting with a face shown on a 2D or 3D surface is that the 2D surface comes with the Mona Lisa effect. For our experiment, this means that the visibility of the face and lips to a subject standing straight in front of the screen or at an angle is the same, and hence if there is an optimal lip reading angle of a face, the face on a 2D screen can maintain that angle and guarantee optimal intelligibility. This is not the same with a 3D head (a physical object). Obviously, the visibility of the lips depends on where the onlooker is standing in relation to the face, and so if looking at the face with an angle is worse than looking at it straight frontal, this would introduce a variable intelligibility depending on the viewing angle. This is found to be the case when reading human lips. In (Erber, 1974) it was found that the lip-reading contribution drops down when looking beyond 45 degrees, but is not significantly different between 0 and 45 degree.

In conclusion, this study aimed at investigating the differences in audio-visual intelligibility (lip readability) of *Furhat* compared to its in-screen counterpart. The results are promising, and validate the suitability of the head as an alternative to animated avatars displayed on flat surfaces. The results also show that people benefit from *Furhat* in terms of lip reading significantly more than showing the same model on a flat display. This is indeed interesting, and motivates future work to investigate the sources of these differences.

Acknowledgments. This work has been partly funded by the EU project IURO (Interactive Urban Robot) No. 248314.

References

1. Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.: Furhat: A Back-projected Human-like Robot Head for Multiparty Human-Machine Interaction. In: Esposito, et al. (eds.) Cognitive Behavioural Systems. LNCS. Springer (2012)
2. Al Moubayed, S., Beskow, J.: Effects of Visual Prominence Cues on Speech Intelligibility. In: Proceedings of Auditory-Visual Speech Processing AVSP 2009, Norwich, England (2009)
3. Al Moubayed, S., Edlund, J., Beskow, J.: Taming Mona Lisa: Communicating gaze faithfully in 2D and 3D facial projections. ACM Trans. Interact. Intell. Syst. 1(2), Article 11, 25 pages (2012)
4. Al Moubayed, S., Skantze, G.: Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. In: Proceedings of the international conference on Auditory-Visual Speech Processing AVSP, Florence, Italy (2011)
5. Beskow, J.: Rule-based visual speech synthesis. In: Proc. of the Fourth European Conference on Speech Communication and Technology (1995)
6. Beskow, J., Edlund, J., Granström, B., Gustafson, J., House, D.: Face-to-Face Interaction and the KTH Cooking Show. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) Second COST 2102. LNCS, vol. 5967, pp. 157–168. Springer, Heidelberg (2010)
7. Cassel, J., Sullivan, J., Prevost, S., Churchill, E.E.: Embodied Conversational Agents. MIT Press (2000)
8. de Melo, C.M., Gratch, J.: Expression of Emotions Using Wrinkles, Blushing, Sweating and Tears. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 188–200. Springer, Heidelberg (2009)
9. Edlund, J., Al Moubayed, S., Beskow, J.: The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 439–440. Springer, Heidelberg (2011)
10. Ezzat, T., Poggio, T.: Visual Speech Synthesis by Morphing Visemes. Visual speech synthesis by morphing visemes. International Journal of Computer Vision 38, 45–57 (2000)
11. Erber, N.P.: Effects of angle, distance and illumination on visual reception of speech by profoundly deaf children. J. of Speech and Hearing Research 17, 99–112 (1974)
12. Granström, B., House, D.: Modeling and evaluating verbal and non-verbal communication in talking animated interface agents. In: Dybkjaer, l., Hemsen, H., Minker, W. (eds.) Evaluation of Text and Speech Systems, pp. 65–98. Springer-Verlag Ltd. (2007)
13. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S.C., 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)
14. Gustafson, J., Boye, J., Fredriksson, M., Johanneson, L., Königsmann, J.: Providing Computer Game Characters with Conversational Abilities. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 37–51. Springer, Heidelberg (2005)

15. Kopp, S., Gesellensetter, L., Krämer, N., 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)
16. 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)
17. Massaro, D.: Perceiving talking faces: from speech perception to a behavioral principle. A Bradford Book. MIT Press, Cambridge (1997) ISBN: 978-0262133371
18. Massaro, D., Beskow, J., Cohen, M., Fry, C., Rodgriguez, T.: Picture my voice: audio to visual speech synthesis using artificial neural networks. In: Proceedings of the International Conference on Auditory-Visual Speech Processing AVSP 1999, Santa Cruz, USA (1999)
19. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature 264, 746 (1976)
20. Pelachaud, C.: Modeling Multimodal Expression of Emotion in a Virtual Agent. Philosophical Transactions of Royal Society B Biological Science, B 364, 3539–3548 (2009)
21. Raskar, R., Welch, G., Low, K.-L., Bandyopadhyay, D.: Shader lamps: animating real objects with image-based illumination. In: Proc. of the 12th Eurographics Workshop on Rendering Techniques, pp. 89–102 (2001)
22. Ruttkay, Z., Pelachaud, C. (eds.): From Brows till Trust: EvaluatingEmbodied Conversational Agents. Kluwer (2004)
23. Salvi, G., Beskow, J., Al Moubayed, S., Granström, B.: SynFace—Speech-Driven Facial Animation for Virtual Speech-Reading Support. EURASIP Journal on Audio, Speech, and Music Processing (2009)
24. Shannon, R., Zeng, F., Kamath, V., Wygonski, J., Ekelid, M.: Speech Recognition with primarily temporal cues. Science 270(5234), 303 (1995)
25. Siciliano, C., Williams, G., Beskow, J., Faulkner, A.: Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. In: Proceedings of the International Congress of Phonetic Sciences, pp. 131–134 (2003)
26. Summerfield, Q.: Lipreading and audio-visual speech perception. Philosophical Transactions: Biological Sciences 335(1273), 71–78 (1992)
27. Sjolander, K.: An HMM-based system for automatic segmentation and alignment of speech. In: Proceedings of Fonetik, pp. 93–96 (2003)
28. Todorovi, D.: Geometrical basis of perception of gaze direction. Vision Research 45(21), 3549–3562 (2006)

Subjective Optimization

Chung-Cheng Chiu and Stacy Marsella

University of Southern California,
Institute for Creative Technologies,
12015 Waterfront Drive,
Playa Vista, CA 90094
{chiu,marsella}@ict.usc.edu

Abstract. An effective way to build a gesture generator is to apply machine learning algorithms to derive a model. In building such a gesture generator, a common approach involves collecting a set of human conversation data and training the model to fit the data. However, after training the gesture generator, what we are looking for is whether the generated gestures are natural instead of whether the generated gestures actually fit the training data. Thus, there is a gap between the training objective and the actual goal of the gesture generator. In this work we propose an approach that use human judgment of naturalness to optimize gesture generators. We take an important step towards our goal by performing a numerical experiment to assess the optimality of the proposed framework, and the experimental results show that the framework can effectively improve the generated gestures based on the simulated naturalness criterion.

1 Introduction

One of the main challenges in building a virtual human is to create its non-verbal behaviors such as gestures. Instead of spending time defining these animations manually, an alternative approach is to build a gesture generator to generate animations for dialogs automatically. One common design for building gesture generators is to apply machine learning algorithms to model the relationship between dialogs and gestures from human conversation data, and the derived model can then generate animations for similar dialogs automatically. The most common design for existing machine learning approaches trains models to fit the training data, but this conventional design is orthogonal to the actual goal of gesture generators, which is to produce natural motions that match well the corresponding dialogue. Although human conversation data collected for training are samples of natural gestures, unless the machine learning model can achieve 100% accuracy on fitting the data while also generalize to novel dialogue, there is no guarantee that the resulting model can generate natural motion. Conversely, since samples in the training data represent only a small portion of possible human gestures, there are many natural gestures that are quite different from

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 204–211, 2012.

the training samples that the gesture generator can learn to generate. Thus, there is a gap between the actual goal and existing approaches, such that the criterion of existing approaches makes the problem more challenging. One way to reduce this gap is to develop a naturalness criterion and use this criterion to optimize gesture generators. The naturalness criterion is a subjective criterion and a reasonable way to acquire this criterion is to collect feedback from humans.

The goal of this work is to propose an algorithm to refine gesture generators using the naturalness criterion. We adopt the idea of pairwise comparisons to help address the issue of the noisiness of absolute subjective judgments. To use these judgments to help learn a model of gesture generation, we frame the problem as a *dueling bandits problem* which optimizes the model from the results of the pairwise comparisons. At each step, the process generates a gradient vector to modify the parameters of the gesture generator and then generates gestures with the new parameters. If the generated animations are evaluated to be more natural than the original one, then the process generates a new set of parameters based on the gradient direction and use the parameters for the next optimization loop. The procedure performs several rounds to refine gesture generators and is applicable for gesture generators with numeric parameters. In this work we incorporate our framework into the gesture generator based on the Hierarchical Factored Conditional Restricted Boltzmann Machines (HFCRBMs) which generates motions based on given prosody.

Because HFCRBMs have very high dimensional parameters and optimizing the model with human subject evaluations is costly, in this paper we perform a numerical experiment to assess the optimality of the proposed framework. We define a metric to simulate human judgment and provide numerical evaluations for the optimization results. The numerical experiment shows that the algorithm can improve the gesture generator significantly.

2 Optimization Framework

In our optimization framework, the process first calculates a gradient vector for the parameters of the specified gesture generator and modifies the parameter with the gradient to get a new model. The framework then uses the two models to generate gesture animations for the same dialogue, pairs animations for the same dialogue from both models together into videos, and evaluates them in a pairwise comparison. The evaluated results are applied to generate the new gradient vector. The optimization process is shown in Figure 1.

One issue raised by the process is the evaluation of the naturalness criterion. Naturalness is a subjective opinion and it requires subject evaluations. Thus, the optimization process requires a mechanism to get evaluations for the quality of the generated gestures based on individual judgments. A conventional approach is to ask individuals to give absolute scores of the gestures (e.g. rate gestures from 1 to 10). However, individuals in general are poor at making absolute judgments as compared to discriminating judgements [6]. In fact, individuals make

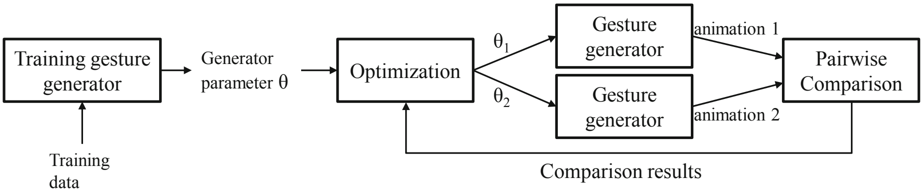


Fig. 1. The flow of the optimization framework

relative judgments for subjective evaluation and require reference points for making absolute judgments [7]. Since different individuals have different standards in mind, the evaluation results of absolute judgments can be inconsistent between individuals. Moreover, reference points of an individual change from one trial to the next [5] which suggests that the results of absolute judgments from the same individual may be inconsistent in itself. An alternative approach is to ask individuals to compare two animations and ask them to judge which one is better. This relative judgment approach not only resolves these biases but also makes the evaluation more reliable as individuals are more consistent in making qualitative judgments than estimating scores. Thus, our framework shows two animations to subjects and asks them to do pairwise comparisons to get naturalness evaluations.

Subject evaluations are very expensive, especially when each evaluation task takes several minutes. A common approach is to do crowdsourcing to collect a large number of evaluations with low cost [4], but this approach can still become expensive in our case when the model is high-dimensional as it may require hundreds of optimization iterations to get a reasonable improvement. We use HFCRBMs [1] as the model for gesture generators in the framework, and there are approximately one million parameters which requires a very large number of updates to optimize. Thus, in this work we define a metric as the naturalness criterion to bypass the expensive cost of human evaluation. We do not propose to replace human evaluations with the metric but rather use it here as an efficient approach to assess the framework. We currently use the metric in this work and will incorporate this with human evaluations as our next step. Thus, the framework is still designed for incorporating human evaluations.

The crucial step of the optimization process is to determine gradient vectors for model parameters. With the design of the pairwise comparisons, the step can naturally be formulated as a *dueling bandits problem* [12]. The dueling bandits problem refers to finding an optimal decision that minimizes regret based on pairwise comparisons. Its discrete version is the K-armed dueling bandits problem [11] in which there are a collection of K bandits and the process needs to determine which bandit leads to the best result based on pairwise comparisons. Dueling bandits problems, in contrast, do not feature K distinct choices of bandits but optimize over an entire finite continuous space of them. Each iteration comprises a comparison between two selected bandits A and A', and the optimization process uses the comparison result to determine the two bandits for

Algorithm 1. Optimization Algorithm

```
Input: $\gamma, \delta, \theta, X$
  $G \leftarrow generate(\theta, X)$
  for $t = 1 \rightarrow T$ do
    $u \leftarrow$ Sample unit vector uniformly with the same dimension as $\theta$
    $\theta' \leftarrow \theta + \delta u$
    $G' \leftarrow generate(\theta', X)$
    if Comparisons show $G'$ is better than $G$ then
      $\theta \leftarrow \theta + \gamma u$
      $G \leftarrow generate(\theta, X)$
    end if
  end for
```

the next comparison. The evaluation results derived from pairwise comparisons are assumed to be noisy where there is a chance that A is evaluated to be better than A' despite A' actually being better than A. The optimality of the selected bandits is quantified as a regret, defined as:

$$R_T = \sum_{t=1}^{T} \epsilon(A_t^*, A_t) + \epsilon(A_t^*, A_t'),$$

where A_t and A_t' are two bandits selected at time t, and A_t^* is the best bandit at time t.

Dueling bandits problems can naturally be related to optimization problems where bandits correspond to gradient vectors for model parameters, and the decision process is equiavlent to choosing an effective gradient for optimizing the model. An algorithm called *Dueling Bandit Gradient Descent* (DBGD) has been proposed to perform online optimization with pairwise comparisons, and its regret bound has been shown in previous work [12].

We apply DBGD to perform online optimization for gesture generators. The algorithm is described in Algorithm 1. In Algorithm 1 δ is the step size for exploring effective gradient direction, and γ is the step size for exploiting this gradient direction in the current model, θ denotes parameters of the gesture generator, X denotes required data for generating gesture animation, and function *generate* represents the specified gesture generator which generates animations Gs based on θ and X. Previous work [12] has shown that when G' is evaluated as better than G the corresponding u indicates a gradient direction that can improve the model.

The algorithm takes θ as input and uses it as an initial point of the optimization process. The optimization process depends on the naturalness evaluation, and without initializing the parameter with the original training algorithm, the generated animations will in general be too poor to make such a comparison. Thus, our framework trains gesture generators with their original training algorithm before starting the optimization process. This algorithm is applicable for gesture generators with numeric parameters.

2.1 Gesture Generator

We follow the same design of the previous work [1] for our HFCRBM-based gesture generator. The HFCRBM-based gesture generator generates gesture motion based on prosody information and past motion. It learns the model from motion capture data of human conversations. The HFCRBM decomposes the gesture learning problem into two parts: it first learns the hidden factors of generating human motion and then learns the correlation between prosody and these hidden factors. This design is based on the idea that human motion is driven by a set of motor signals that in combination produce a sequence of movements. In contrast, the motion capture data is represented as real valued vectors that in essence obscure the factors and constraints that were involved in generating the data. Thus, the model first infers the underlying causes of motion and then learns the relation between speech and the hidden factors. The HFCRBM is comprised of two components, a reduced conditional restricted Boltzmann machine (RCRBM) [2] for inferring hidden factors of motion and a factored conditional restricted Boltzmann machine (FCRBM) [8] for modeling the relationship between prosody and hidden factors. To learn gesture generators from human conversation data, the model first trains RCRBMs with motion capture data. After this training step, the HFCRBM maps motion data onto hidden factors with RCRBMs and trains the top-layer FCRBMs conditioned on audio features. For gesture generation with HFCRBMs, the generation process works by taking previous motion frames and audio features as input to generate the next motion frame. The generation of a sequence of motion is done in a recurrent way in that the generated motion frame becomes part of the input of the next generation step. The process can generate a motion sequence with the same length as the audio.

The HFCRBM for the gesture generator has many parameters, and performing our optimization algorithm with all the parameters is inefficient. It is more reasonable to pick a subset of parameters for the optimization process. We can use the fact that the HFCRBM comprises two separate modules and focus on optimizing only one of them. The reason for proposing this optimization framework is to improve the generalization of gesture generators for novel prosody, and therefore it is more reasonable to focus on improving the modeling among prosody and hidden factors with the naturalness criterion. Thus, we only update the parameters of FCRBMs. Because we only optimize the top model, in order to give it better control we choose to use RCRBMs as the bottom model instead of CRBMs [9], which results in a HFCRBM with the same architecture as [2]. In our HFCRBM model, the hidden layers of the RCRBMs and FCRBMs each have 300 nodes. The prosodic features for gesture generators at each time frame have a window of ±1/3 seconds.

2.2 Simulating Pairwise Comparisons

The crucial step of the optimization algorithm is the pairwise comparison. We define a metric to simulate human evaluation and simulate the pairwise comparison by comparing the metric values of two animations. The metric function for an animation set G is defined as:

$$metric(G) = \sum_{i=1}^{n} ||cor(g_i^*) - cor(g_i)||_2^2 \, /2n.$$

In this function $G^* = g_1^*, ..., g_n^*$ represents a set of human motion for the corresponding prosody, and $G = g_1, ..., g_n$. The function $cor(g)$ is defined as:

```
**Input:** $g$
  $p \leftarrow pitch(g)$
  $i \leftarrow intensity(g)$
  $v \leftarrow velocity(g)$
  $a \leftarrow acceleration(g)$
  **return** $correlation(p, a)$,$correlation(i, v)$
```

where $correlation(x, y)$ calculates the linear correlation of the two input sequences x and y. We proposed this metric under an assumption that the pitch is more likely to be correlated with the movement acceleration and the intensity is more likely to be correlated with the movement velocity. One possible alternative is to directly compare the generated motion and the motion capture data, but since the motion capture data only shows one instance of possible gestures this explicit comparison can limit the variety of generated gestures. Thus, we choose a more implicit comparison metric to increase the flexibility for the gesture generation.

3 Experiments

We assessed our framework by conducting an experiment to analyze its performance empirically. To train the gesture generator, we used a dataset originally created for a different study that examined how audio and body motion affected the perception of virtual conversations [3]. We followed the approach suggested in [1] to extract the data. Unlike their configuration, we only use prosody features as contextual information and exclude correlation parameters, as they did not seem to improve the quality of generated gestures. There are a total of 1140 frames (38 seconds) of training data.

We trained the gesture generator with the HFCRBM training process, and then used our optimization framework with the simulated pairwise comparison described in subsection 2.2. We used the training data applied in the HFCRBM training process for the simulated comparison process, in which the gesture generator is requested to generate gestures with the prosody in the training data and the generated gestures are compared with the corresponding motion capture data based on our defined metric. We ran the optimization loop for 1000 iterations, and the output of the simulation metric (can be understood as error values) of each successive training iteration are shown in Figure. 2. After 1000 iterations, the value drops to less than half of the original value.

The experiment shows that the algorithm can effectively improve the generation result of the gesture generator. One limitation of this framework for gesture

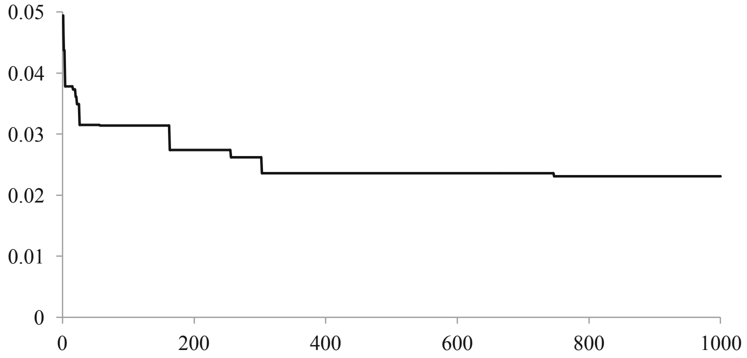


Fig. 2. Error values of the optimization process

generator is that when the model has many parameters, it requires a lot of iterations for the optimization process to improve the results. For example, [10] uses 10^5 iterations to train their model for search ranking. Also, there is no guarantee that the function we are optimizing is convex, which makes it necessary to try several initializations to avoid local minimum. These two properties suggest that we need a large number of pairwise comparisons to get promising results which leads to expensive demands on human effort. A possible solution is to improve the HFCRBM so that it can maintain similar quality in gesture generation by using far fewer parameters. Another method is to introduce some heuristic function based on domain knowledge to narrow the search space. Further work needs to be done before we can investigate the incorporation of real human judgment into this framework.

4 Conclusions

This work seeks to address the common problem of existing machine learning based gesture generators in which the training objective does not match the naturalness criterion people expect for gesture generators. We have proposed a framework to improve gesture generators with a naturalness criterion. The framework lays a foundation for training gesture generators using a naturalness criterion. Specifically, we applied our framework to improve a HFCRBM-based gesture generator. The framework identifies a gradient that can improve the model and updates the parameters iteratively. The optimization algorithm uses the information from comparing two generated gesture animations, and the pairwise comparisons are simulated with a metric based on prosody and motion. The efficacy of the framework is demonstrated in experiments that show significant improvement of the HFCRBM-based gesture generator. The major limitation of the framework is that the cost of the optimization process can become impractical when applied to gesture generators with too many parameters. Future work needs to address this parameter problem before we can proceed with moving from a simulated human judgments to actual human judgments.

Acknowledgements. This research was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) Simulation Training and Technology Center (STTC). The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. 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)
2. Chiu, C.-C., Marsella, S.: A style controller for generating virtual human behaviors. In: Proceedings of the 10th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011, vol. 1 (2011)
3. Ennis, C., McDonnell, R., O'Sullivan, C.: Seeing is believing: body motion dominates in multisensory conversations. In: ACM SIGGRAPH 2010 Papers, SIGGRAPH 2010, pp. 91:1–91:9. ACM, New York (2010)
4. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 453–456 (2008)
5. Mozer, M., Pashler, H., Wilder, M., Lindsey, R., Jones, M., Jones, M.: Improving human judgments by decontaminating sequential dependencies. In: Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 1705–1713 (2010)
6. Stewart, N., Brown, G.D.A., Chater, N.: Absolute identification by relative judgment. Psychological Review 112(4), 881–911 (2005)
7. Stewart, N., Chater, N., Brown, G.D.: Decision by sampling. Cognitive Psychology 53(1), 1–26 (2006)
8. Taylor, G., Hinton, G.: Factored conditional restricted Boltzmann machines for modeling motion style. In: Bottou, L., Littman, M. (eds.) Proceedings of the 26th International Conference on Machine Learning, pp. 1025–1032. Omnipress, Montreal (2009)
9. Taylor, G.W., Hinton, G.E., Roweis, S.T.: Modeling human motion using binary latent variables. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19, pp. 1345–1352. MIT Press, Cambridge (2007)
10. Yue, Y.: New learning frameworks for information retrieval. Ph.D. thesis, Cornell University (2011)
11. Yue, Y., Broder, J., Kleinberg, R., Joachims, T.: The k-armed dueling bandits problem. In: Conference on Learning Theory (2009)
12. Yue, Y., Joachims, T.: Interactively optimizing information retrieval systems as a dueling bandits problem. In: ICML (2009)

Understanding the Nonverbal Behavior of Socially Anxious People during Intimate Self-disclosure

Sin-Hwa Kang, Albert (Skip) Rizzo, and Jonathan Gratch

Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Playa Vista, CA 90094 USA
{kang,rizzo,gratch}@ict.usc.edu

Abstract. This study explores the types of nonverbal behavior exhibited by socially anxious users over the course of an interview with virtual agent counselors that talked about themselves. The counselors provided self-disclosure using human back stories or computer back stories. The video data was collected from a previous study. We defined nine types of nonverbal behavior to investigate the associations between the types of nonverbal behavior and users' anxiety levels. The results of preliminary data analysis show that five features out of the nine features are positively correlated with different levels of users' anxiety in the "computer back stories" condition. These five types of nonverbal behavior are gaze aversion, moving arms and hands, constant rocking, shaking a head, and fidgeting arms and hands. There are no significant relationships between the kinds of nonverbal behavior and users' anxiety levels in the "human back stories" condition.

Keywords: nonverbal behavior, embodied virtual agents, social anxiety, affective behavior, self-disclosure, virtual humans, evaluation.

1 Introduction

This study is part of a larger program to demonstrate the suitability of virtual agents that sense client nonverbal cues in therapeutic interactions where human clients need to be encouraged to disclose sensitive information. In clinical interactions, nonverbal behavior is widely considered a crucial indicator of a client's mental state [14]. Likewise, nonverbal behavior can help build intimacy between a client and a clinician [1,4] as nonverbal cues serve to communicate inner feelings and intentions [2,14,17]. In fact, some evidence suggests that nonverbal cues may serve as more credible indicators of clinical states than the verbal information communicated by clients [11]. For example, research in clinical psychology has found that the nonverbal behavior of human clients unintentionally revealed intimate information that is not disclosed in their verbal behavior [5,17]. More broadly, nonverbal behavior plays a vital role in the creation and maintenance of a therapeutic relationship by constructing rapport between counselors and clients in psychotherapeutic interactions [17].

Human clinicians invest considerable time and effort into carefully observing the nonverbal behavior of clients and adjusting their own nonverbal behavior to respond

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 212–217, 2012.

appropriately and build intimacy. If virtual agents were capable of detecting and interpreting both verbal and nonverbal signals from human clients, such agents could obtain a better understanding of the clients' intention and, thus, respond in a more appropriate manner. For instance, a virtual counselor must be able to understand the psychological states of human clients to approach the level of rapport and diagnostic efficacy of human clinicians in psychotherapeutic interactions. Thus, giving virtual agents the ability to recognize and understand these indicators would greatly enhance their relevance in clinical settings. This may lead to the creation of a virtual clinical setting run more similarly to the way that human counselors interact with their clients in the real world.

Social anxiety is the most common clinical condition faced by clinicians, occurring in 18% of the general population [15]. Therefore, recognizing nonverbal indicators of social anxiety is a priority for our research in clinical virtual agents. In this study, we aim to identify the kinds of nonverbal behavior displayed by human clients with varying anxiety levels during their interaction with a virtual counselor. We introduce the preliminary results of our study in this paper.

We investigated the types of nonverbal behavior displayed by socially anxious users over the course of an interview with a virtual counselor that self-disclosed personal information. The video data analyzed in this paper was recorded during a previous study [10] described in detail in the following section.

2 Experimental Design

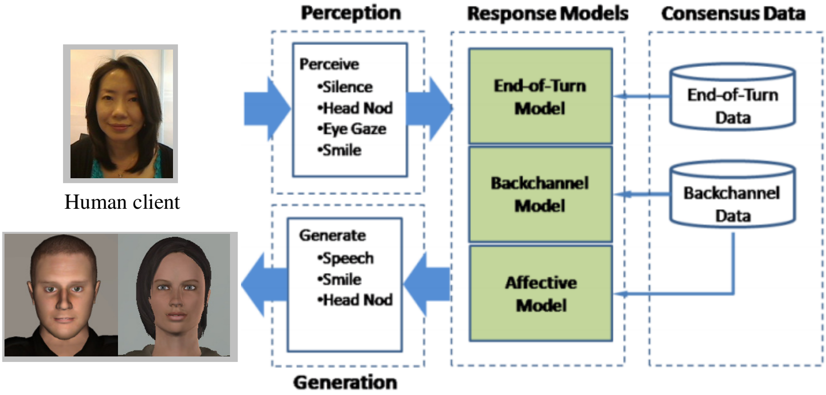
In a previous study [10], we investigated whether the different types of virtual agent counselors' self-disclosure affected real human clients' social responses in psychotherapeutic interactions. We designed a between-subjects experiment involving two different kinds of self-disclosure from virtual counselors in an interview setting: i) human back stories, e.g. "I was born and raised in LA"; ii) computer back stories, e.g. "I was designed and built in LA." Each experimental condition was presented to same gender combinations of dyadic partners.

2.1 Participants and Procedure

Forty people (50% women, 50% men; average 31 years old) from the general Los Angeles area were recruited using Craigslist.com and compensated for seventy five minutes of their participation. The participants were randomly assigned to one of the two experimental conditions. Participants were given instruction describing the counseling interview interaction. The interview questions were modified from ones used in a previous study [9]. The virtual counselors preceded each interview question with some information about themselves before asking each counseling question to participants. Participants in all conditions viewed the virtual agents on a 30-inch screen display that approximated the size of a real human sitting 4 feet away. They wore a lightweight close-talking microphone and spoke into a microphone headset. The monitor was fitted with a camcorder and a webcam. To control for gender effects, two types of gender dyads were used in equal numbers in each experimental condition: male-male and female-female. The typical interaction was allowed to last about thirty minutes.

2.2 Stimulus Materials

The Rapport Agents [8] were used as virtual counselors (see the image (a) in Figure 1) that presented timely positive feedback, such as smile and head nods, by recognizing and responding to the audiovisual features of a participant (human client) (see the image (b) in Figure 1).



(a) Virtual counselors

(b)

Fig. 1. (a) Rapport Agents (male & female); (b) System architecture of the Rapport Agent

To generate the virtual counselor's behaviors, the Rapport Agent first collected and analyzed the attributes from the voice, smile, head nods, eye-gaze, and upper-body movements of a human client. To detect the client's behaviors, a webcam was placed in front of the client. An audio cue detector extracted data such as the intensity of the client's voice from the raw signal using the signal processing package, Praat. A visual cue detector tracked the direction of eye-gaze, head nods, smile levels, and body movements. The backchannel, end-of-turn and affective models of the Rapport Agent were unique in their ability to make real-time decisions and generate the most appropriate responses to client statements using perceived audiovisual features. For example, the virtual counselor may provide back-channeling in the form of a smile if the human client smiles. To generate speaking behaviors of the virtual counselor to provide self-disclosure, an experimenter controlled the buttons that retrieved pre-recorded voice messages. The same male and female virtual agents were used in all conditions (see the images in Figure 1 (a)).

2.3 Measurements

Social anxiety. The pre-questionnaire packet included questions about one's social anxiety as a dispositional personality trait. We utilized the modified Cheek & Buss shyness scale [3] to measure users' anxiety levels. Scales ranged from 1 (disagree strongly) to 5 (agree strongly). Sample items include: 'I feel tense when I'm with people I don't know well' and 'I feel inhibited in social situations.'

Nonverbal behavior. We defined nine nonverbal features to explore the types of nonverbal behavior exhibited by socially anxious users: gaze aversion, frowning eyebrows, leaning, moving arms and hands, constant rocking, touching on body, shaking a head, fidgeting arms and hands, and fidgeting feet and legs. These nonverbal behaviors were extracted from an extensive literature review and previously observed features of social anxiety in previous work [13]. A coder annotated the frequency of nonverbal cues present by tallying the occurrences of behavior displayed by participants.

3 Preliminary Findings

We ran a Pearson Correlation for users' anxiety levels (M = 2; SD = .65) and frequency of nonverbal behaviors present in each experimental condition. The results show that five features are positively correlated with users' anxiety levels in the "computer back stories" condition (see Figure 2). These five features are gaze aversion (r = .5), moving arms and hands (r = .5), constant rocking (r = .55), shaking a head (r = .53), and fidgeting arms and hands (r = .61). There is a general trend of positive associations between the users' anxiety levels and the rest of the behaviors in the condition. It is also worth noting that there is no significant difference in the length of conversations as a function of anxiety level. Thus, these correlations indicate that these nonverbal behaviors are derived from the quality rather than the quantity of their speech. It seems that the users' existing views toward programmed characters' own stories might have, in turn, prompted a more awkward interaction that contributed to showing nonverbal behaviors associated to higher anxiety when speaking to the counselor using computer back stories. There are no significant associations between these two variables in the "human back stories" condition.

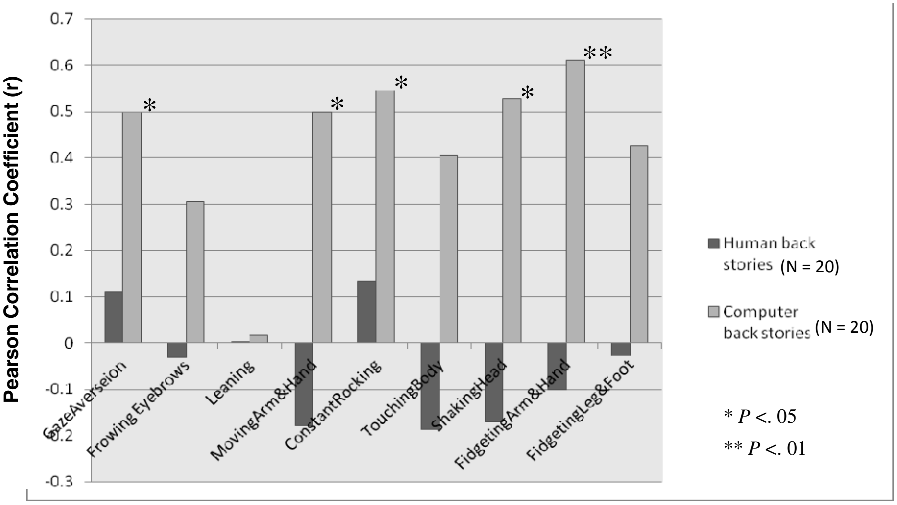


Fig. 2. Correlations between users' Social Anxiety Levels and frequency of Nonverbal Behaviors in two experimental conditions

4 Discussion and Future Work

Previous studies indicate that the nonverbal signs of social anxiety include gaze aversion, facial expressions, and body movements which relay discomfort, such as extremity movements [13]. The output of our study is in line with and reinforces these prior findings. Our examination of the nonverbal behaviors associated with social anxiety levels in users interacting with virtual counselors supports a movement in the direction of developing more effective virtual agents in the future.

Detecting the nonverbal signals of users could complement comprehension of their verbal content and result in a virtual agent that employs this crucial information to assess the user's emotional state. The virtual agent could then utilize this information to create a higher fidelity model of the user's state that would enhance the quality of the agent's feedback. Potentially, such nonverbally-aware agents would result in more robust communication between the agent and humans and facilitate establishment of rapport [6,7].

In our future work, we plan to extend this analysis to other clinically-relevant client states, such as depression and post-traumatic stress disorder, building on a larger recently-collected dataset of clinical interviews with clients with such conditions. By validating the presence of nonverbal cues that correlate with clinical states, this research informs the design of automatic techniques that aspire to recognize such cues in real-time, within the context of clinical interactions [16]. Although considerable technical and ethical hurdles must be overcome as this research proceeds, ultimately this research is advancing the potential of virtual human agents that can assist in clinical contexts.

Acknowledgements. This work was sponsored in part by the DARPA Detection and Computational Analysis of Psychological Signals (DCAPS) program 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. Argyle, M., Dean, J.: Eye-contact, distance, and affiliation. Sociometry 28, 289–304 (1965)
2. Cacioppo, J.T., Petty, R.E., Losch, M.E., Kim, H.S.: Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. Journal of Personality and Social Psychology 50, 260–268 (1986)
3. Cheek, J.M.: The Revised Cheek and Buss Shyness Scale (RCBS). Wellesley College, Wellesley (1983)
4. Edinger, J., Patterson, M.: Nonverbal Involvement and Social Control. Psychological Bulletin 93(1), 30–56 (1983)
5. Farber, B.: Self-Disclosure in Psychotherapy. Guilford, New York (2006)

6. 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)
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. Huang, L., Morency, L.-P., Gratch, J.: Virtual Rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
9. Kang, S., Gratch, J.: People like virtual counselors that highly-disclose about themselves. The Annual Review of CyberTherapy and Telemedicine 167, 143–148 (2011)
10. Kang, S., Gratch, J.: Socially Anxious People Reveal More Personal Information with Virtual Counselors That Talk about Themselves using Intimate Human Back Stories. In: Proceedings of the 17th Annual CyberPsychology & CyberTherapy Conference (2012)
11. Knapp, M., Hall, J.: Nonverbal Communication in Human Interaction. Wadsworth | Cengage Learning, Boston (2010)
12. Morency, L.-P., Sidner, C., Lee, C., Darrell, T.: Contextual Recognition of Head Gestures. In: Proceedings of the 7th International Conference on Multimodal Interactions, Torento, Italy (2005)
13. Perez, J., Riggio, R.: Nonverbal social skills and psychopathology. In: Philippot, P., Feldman, R., Coats, E. (eds.) Nonverbal Behavior in Clinical Settings. Oxford University Press, New York (2003)
14. Philippot, P., Feldman, R., Coats, E.: The Role of Nonverbal Behavior in Clinical Settings. In: Philippot, P., Feldman, R., Coats, E. (eds.) Nonverbal Behavior in Clinical Settings. Oxford University Press, New York (2003)
15. Raj, B., Sheehan, D.: Social Anxiety Disorder. Medical Clinics of North America 85(3), 711–733 (2001)
16. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Rizzo, A., Morency, L.-P.: Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. In: Nakano, Y., et al. (eds.) IVA 2012. LNCS (LNAI), vol. 7502, pp. 455–463. Springer, Heidelberg (2012)
17. Tickle-Degnen, L., Gavett, E.: Changes in nonverbal behavior during the development of therapeutic relationships. In: Philippot, P., Feldman, R., Coats, E. (eds.) Nonverbal Behavior in Clinical Settings. Oxford University Press, New York (2003)

Virtual Reality Negotiation Training Increases Negotiation Knowledge and Skill

Joost Broekens[1], Maaike Harbers[1], Willem-Paul Brinkman[1],
Catholijn M. Jonker[1], Karel Van den Bosch[3], and John-Jules Meyer[2]

[1] Delft University of Technology
joost.broekens@gmail.com, {m.harbers,w.p.brinkman,c.m.jonker}@tudelft.nl
[2] Utrecht University
jj@cs.uu.nl
[3] TNO Human Factors Soesterberg
karel.vandenbosch@tno.nl

Abstract. In this paper we test the hypothesis that Virtual Reality (VR) negotiation training positively influences negotiation skill and knowledge. We discuss the design of the VR training. Then, we present the results of a between subject experiment (n=42) with three experimental conditions (control, training once, repeated training) investigating learning effects on subjects' negotiation skill and knowledge. In our case negotiation skill consists of negotiation outcome (final bid utility) and conversation skill (exploratory conversational choices in VR scenario), and negotiation knowledge is the subjects' quality of reflection upon filmed behavior of two negotiating actors. Our results confirm the hypothesis. We found significant effects of training on conversation skill and negotiation knowledge. We found a marginally significant effect of training on negotation outcome. As the effect of training on negotiation outcome was marginally significant and only present when controlling for overshadowing effects of the act of reflecting, we postulate that other learning approaches (e.g., instruction) are needed for trainees to use the information gained during the joint exploration phase of a negotiation for the construction of a bid. Our results are particularly important given the sparse availability of experimental studies that show learning effects of VR negotiation training, and gives additional support to those studies that do report possitive effects such as with the BiLAT system.

1 Introduction

Virtual Reality systems are effective tools to change human behavior in a wide variety of domains including training medical skils, education of children, military procedures, flying, but also the treatment of phobias through VR exposure therapy (see [22,29,19,26,28,2]). A key characteristic of these systems is that they are effective at inducing cognitive and behavioral changes for a relatively constrained and well-defined setting. Systems that have shown to be effective include treating particular anxieties through exposure of the subject such as fear of heights [8], training particular skills such as teaching children to safely cross

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 218–230, 2012.

a street [32], a particular procedure such as emergency situtation triage [1], or a particular sensory-motor skill such as a specific type of surgery [12]. More recently VR training has been proposed for ill-defined training tasks such as cultural understanding, persuasion, social skills and negotiation, usually in the form of a serious game [6,25,14,30]. However, for these more complex, and often ill-defined tasks, it is difficult to develop the right simulation content, storyline, interactions, and outcome measures [14]. As a result of these difficulties and the novelty of the field, there is only sparse evidence of such VR systems showing measureable learning effects, a point explicitly made in [30].

We focus on negotiation support systems for novice negotiators and within that context aimed to develop a VR negotiation training. Only several accounts exist of experimentally verified learning effects of VR negotiation training [20,7], and with the same system (i.e., BiLAT, [18]). It is therefore important to investigate learning effects targeted at the same phenomenon (i.e., negotiation knowledge and skills) with a different system, because positive results could easily be tied to the specific choices of a system with respect to domain, implementation, and content. In this paper we present an experiment with a virtual training system for negotiation that has been carefully constructed, involving a virtual agent that is able to express emotions and explain its behavior. VR negotiation training is in essence a role play between a human and a virtual human, as often used in traditional negotiation training. Therefore, the use of intelligent virtual agents equipped with human-like capabilities such as emotion and explanation is a logical choice. This paper addresses two topics. First it describes in detail the design of the system, so that choices and assumptions are made explicit. Second, we present results of an experiment investigating learning effects of the training on negotiation skill and knowledge. In our case negotiation skill consists of an outcome measure and a process measure; i.e., negotiation outcome and conversation skill. We define negotiation outcome as the utility of the final bid proposed by the subject. We define conversation skill as the number of times a subject selects responses that open the conversation towards finding underlying concerns minus the number of times a subject selects responses directing the conversation towards a premature ending. In our scenario, opening responses are responses that are polite, show interest in the other and ask for underlying interests instead of prematurely fix issues. We define negotiation knowledge as the subjects' quality of reflection upon the filmed behavior of two negotiating actors. Negotiation skill in our experiment thus measures in-game non-transferred skills as displayed in the actual negotiation behavior of subjects while playing the VR scenario. Negotiation knowledge measures implicit knowledge transferred from the VR training to the analysis of negotiation behavior of others.

In section 2 we provide background on negotiation and distill the requirements for our negotiation training. In section 3 we discuss the design of VR negotiation training in detail. In section 4 we present the experimental setup and results. Section 5 presents a more general discussion.

2 VR Training Requirements

The naive view on a negotiation is that it is a single task aimed at claiming the highest outcome value by bargaining the best price for a particular good. This naive view on negotiation has several important shortcomings resulting in a difficulty to reach a win-win outcome [10,27,31]. A win-win outcome is an agreed-upon bid that is optimal in terms of overall outcome value for both sides of the negotiation. First, the naive view focusses on a single issue, i.e., money, while any meaningful negotiation involves multiple issues, relationships, and emotions. Second, it focusses mainly on the bidding process and approaches bidding as bargaining (e.g., about price). This hinders getting a good overview of all issues that play a role in the negotiation and thus limits the possibility to place interesting bids that are good for both sides. Third, and related to the previous, it does not emphasize the different phases in a negotiation process. Any negotiation can be separated into at least four phases: preparation, joint exploration, bidding and closing (see, e.g., [15]). The preparation takes place before the negotiation partners meet, and involves the collection of information about one's own and the partner's desires. In the exploration phase, the negotiation partners start to explore each others' wishes. Subsequently, in the bidding phase the negotiation partners exchange actual bids, and in the closing phase the partners leave each other with or without an agreement, make plans for further negotiation, renegotiation, and make sure the relationship is well-managed.

A more realistic view on negotiation is thus that it is a four step process involving the exploration of issue preferences of and by the different parties in the negotiation in order to be able to get closure on a deal that has value for all parties and will be respected afterwards. Although such a process seems overkill for simple day-to-day negotiations it is not [32]. Even the distribution of household tasks among couples is a multi-issue negotiation including issues such as doing the dishes, putting the kids to bed, cooking, and doing finances and taxes. Partners have preferences for or against doing these tasks and usually figure out a win-win bid that honors these preferences. These bids are renegotiable each day, and often *are* being renegotiated. The bids are complete bids (I don't feel like doing the dishes, but I don't mind putting the kids to bed, etc.) and not based on single issue bargaining. When getting home from work one usually has preferences about the different tasks and in fact privately prepares the negotiation. Then in a short exploration phase the different issues and preferences are explored (I don't feel like doing X today, I don't mind Y, you don't mind X?, etc.). Several bids are exchanged, a deal is made and should be honored (no-one will get away in the long run with not honouring the fact that you said you would do the dishes but then simply decide not to). In fact these simple negotiations are perfect examples of negotiations in separate phases, and show the shortcomings of the naive view on negotiation: you rarely bargain about a single household issue and then think it is fair to claim as much value (as little work) as possible.

The example also highlights the importance of ensuring a good relationship. Most negotiations involve a relation between the two negotiation partners. Even after buying a car a relationship follows, albeit a very limited one (service agreement). This brings us to an important element in negotiation: emotion. Emotions play a role before, during and after a negotiation. People have preferences about issues that are in essence affective attitudes. People have an opinion about negotiation in general and about having to enter one in particular. People experience emotions during negotiations, and use emotions strategically. As such, it is critical to address and be aware of your own and the other side's emotions in a negotiation [9], and the importance of emotion in negotiation has been experimentally shown in a large number of psychological studies (for review [4]).

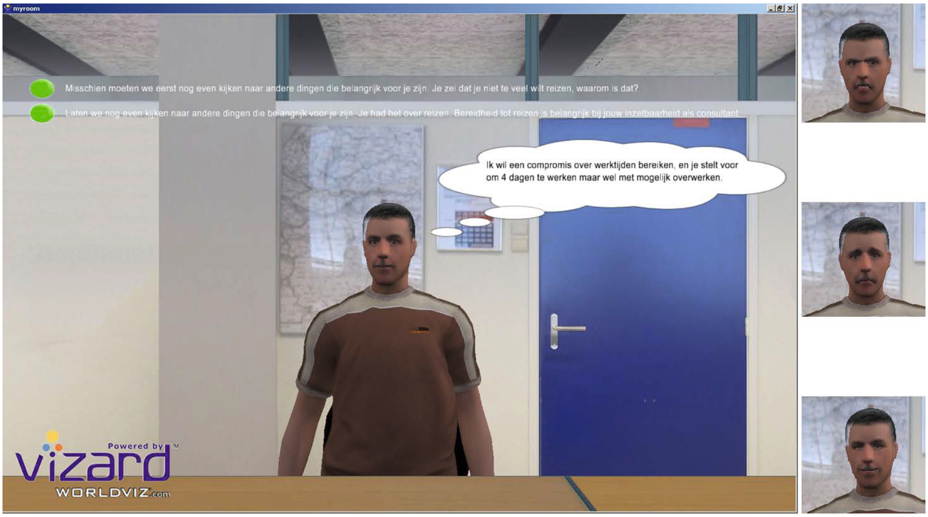
An often-made mistake by novice negotiators in the joint exploration phase is to only explore each others' preferences on issues, e.g. the height of a salary, and forget to ask about the other's interests, e.g. the need of enough money to pay the mortgage. It is important to learn that by exploring someone's interests, alternative solutions can be found that are profitable for both partners, e.g. a lower monthly salary but with a yearly bonus. This mistake was confirmed by a diary study we performed as a preparation for the development of the virtual reality scenario. The study involved 8 subjects who were asked to keep track of their negotiation for a new job or a new house. Subjects often reported about issues, but rarely reported how these issues were derived from one's own underlying interests, let alone the interests of the other party.

These case studies and theoretical analysis have been the basis for the requirements of our negotiation training. First, trainees must follow a phase-based negotiation, with a clear separation between exploration and bidding. Second, emotions play an important role during the negotiation training. Third, the training should focus on investigating underlying concerns, rather than issues.

3 VR Training Design

The main training goal is to make people realize the importance of, and get skilled at, investigating issues and interests (underlying concerns). The training involves a negotiation about terms of employment and involves a human player in the role of an employer and a virtual agent playing the future employee. It has two negotiation phases: the joint exploration phase and the bidding phase. The trainee can interact with the agent by selecting a conversational response from a multiple choice selection (Figure 1). Choices influence the course of the scenario as explained below. The scenario is represented as a conversation tree with branches that can be conditionally activated or deactivated based on previous choices. Total playtime averages around 10 minutes, and the tree consists of about 150 sentence nodes. The virtual agent communicates in natural speech, pre-recorded by a professional voice actor. Beforehand, the virtual training and scenario were reviewed and approved by a professional negotiator.

In more detail, the training scenario focuses on the exploration phase in which the trainee and the character explore each others' standpoints concerning topics such as monetary gain and commute time. A total of four topics are explored in a fixed sequence. Throughout the scenario the trainee can make subtle conversational choices approaching the topic either from an underlying interest point



Misschien moeten we eerst nog even kijken naar andere dingen die belangrijk voor je zijn. Je zei dat je niet te veel wilt reizen, waarom is dat?

Laten we nog even kijken naar andere dingen die belangrijk voor je zijn. Je had het over reizen. Bereidheid tot reizen is belangrijk bij jouw inzetbaarheid als consultant

Ik wil een compromis over werktijden bereiken, en je stelt voor om 4 dagen te werken maar wel met mogelijk overwerken.

Fig. 1. The negotiation training showing two conversational options, the Virtual Character and the explanation as a thought bubble (left). Emotional expressions (right).

of view or an issue point of view. Conversational choices that approach the topic based on underlying interests will eventually broaden the range of issues that can be used to resolve a conflict. The mechanism is the same for all four topics. Interest-based exploration will trigger the Virtual Character (VC) to introduce a non-distributive issue to resolve a conflict around a distributive issue for a particular topic. Values for a distributive issue are positively related to the utility for one negotiator but negatively to the utility for the other (if one wins, the other looses), while values of a non-distributive issue have the same relation to the utility of both negotiators (both win or loose together). For example, if the trainee keeps asking about why the virtual character (the future employee) needs a particular salary, eventually the VC will tell the trainee that he is planning a world trip in one year (interest) and needs to have a certain amount of money for this, but that it is also possible to get this money as an end-of-year presentation-based bonus instead of a fixed salary. This should be acceptable to the trainee as this limits the financial risk of hiring personnel and gives incentive to the employee to work hard (the trainee is told in the role description that he/she owns a startup and hence risk and motivated personnel is an important thing to manage). The end-of-year bonus is a non-distributive issue that can be used to replace the distributive issue salary. All interests and issues used in the scenario are based on the diary studies.

When all four topics in the exploration phases have been explored, the trainee constructs one complete bid based on the issues that have been found during the exploration phase. This bid typically consists of distributive and non-distrubutive issues as found through conversation with the VC. For each topic the trainee has three options, two are always available, the third has to be 'unlocked' by exploring the agent's interests in the exploration phase as explained above.

The first option is the value for the distributive issue according to the trainee's original standpoint (hardliner). The second option is a compromise value for that issue, in between the trainee's and VC's standpoint. The third option is a win-win value for the non-distributive issue. The utility of the bid is scored as follows. For each non-distributive issue used in the bid the utility gain equals 2. For each compromise on a distributive issue the utility gain equals 1. For each hardliner value, the utility gain equals 0. This means that the utility ranges between 0 and 8 (4 topics in total). A win-win agreement is defined as a utility>6, no agreement is defined as a utility<3. Other values involve compromise agreements. As a result, only subsequent finding and use of the non-distributive issue in constructing the bid can lead to a win-win solution with a high utility, reflecting the fact that the bid must be good for both parties.

To enhance the realism of the virtual training, and to emphasize the importance of emotions during the negotiation process, the VC facially expresses three basic emotions as feedback to the trainee's selected response option: happiness, sadness and anger. These expressions have been evaluated beforehand in a separate study (n=19) and showed to be uniquely identifiable [5]. These three emotions have been chosen because of their meaning for giving feedback to a conversation partner. Happiness signals a - for the virtual character - potentially positive outcome of a chosen option (i.e., happyness is the VC's reaction to a trainees selection of a response that opens up the conversation towards underlying concerns), sadness signals a potentially bad outcome (i.e., expressed when the trainee selects a response that steers away from underlying concerns), while anger signals an actual bad outcome (i.e., a reaction to the trainee selecting a response that eliminates the possibility to use a non-distributive issue in the final bid). This meaning is compatible with a goal-based intrerpretation of emotions as in cognitive appraisal theory [24], as well as operant conditioning where positive (social) feedback is given to reinforce behavior and negative (social) feedback is given to discourage behavior.

Because understanding of the other side's preferences is an important shortcoming of novice negotiators (see above), we support trainees in their learning by making the negotiation agent (the VC) able to explain its own behavior. Explanations about agent behavior aim to help trainees to better understand played training sessions and learn from them, see e.g. [16,33,11]. The explanations in this system are based on our previous work on the development of explainable BDI agents for virtual training [13]. The approach is based on folk psychology, i.e. the way people think that they think. Namely, humans explain and understand their own and others' behavior in terms of its underlying desires, goals, beliefs, intentions and the like [17,21]. In earlier work, we explored which explanation types people prefer in which situation (e.g. belief or goal-based explanations) [3], and proposed guidelines for the explanation of agent behavior [13]. These guidelines have been used to develop the explanations for the training. The explanations are offered in the form of thought clouds (see Figure 1) to offer explanations at the time they are most relevant, but without disturbing the flow of the scenario.

4 Experiment

To evaluate training effects of the negotiation training, we have conducted an experiment. Our main hypothesis was that VR negotiation training improves negotiation knowledge and skills. We now detail the experimental design, protocol, subject sample and materials.

4.1 Method

We performed a standard between-subject experiment with three conditions. In the *control* condition subjects did not perform a VR training session prior to collecting effect measures. In the *single* session condition subjects performed one VR training session. In the *repeated* condition subjects performed 5 sessions. As preliminary studies with smaller number of subjects and only a single training session did not show learning effects, the repeated condition was added to make sure subjects had enough training. One training session took approximately 10 minutes. Table 2 shows the experimental protocol schematically.

First, all subjects rated their daily life self-reported negotiation skill, negotiation liking, negotiation frequency, and negotiation perseverance when negotiating. Ratings were on a 5-point Likert scale. As Cronbach's alpha for the four items was acceptable (alpha=0.71), the four items were integrated into one construct measuring self-reported *negotiation tendency.* This tendency to negotiate thus consists of self-reported doing, liking, skill and perseverance in negotiation.

Second, the subjects in the single session and repeated conditions performed the VR negotiation training with explanations and emotions in an office setting behind a standard desktop pc wearing headphones. Subjects in the single session condition were asked to play as well as possible. Subjects in the repeated condition were asked to explore the training in the first four sessions but to play as well as possible in the 5th session. We recorded final negotation outcome and conversation skill for both conditions. Subjects in the control condition did nothing but continued immediately with the next step in the protocol.

Third, we presented each subject with 5 pre-recorded scenes showing a similar job negotiation acted by two actors. Beforehand, the scenes had been judged plausible by a professional negotiator. We asked subjects to take the role of advisor for the employer and write a reflection upon these scenes in an open response format. Subjects were asked to answer two questions per scene: What just happened in the scene? And, what is your advice for the employer? After the experiment two independent raters rated the quality of the reflection for each subject based on the following coding scheme that identifies knowledge and understanding of negotiation:

- The advisor proposes to ask for the underlying reason for the employee's preference for part-time work (scene 1).
- In case of an impasse, the advisor proposes to broaden the negotiation by explicitly mentioning new issues or interests (scene 2, 3).
- The advisor proposes a clear closure of the negotiation (scene 4).

– The advisor assumes negotiation partners are equal: there are no signs of hierarchy, single-sided dependency, dominance or a ’battle for points’.
– The advisor stresses the importance of a good atmosphere.

For each item 1 point could be gained, effectively creating a 6 point scale (1-6). As the inter-rater reliability between the two raters was excellent (Cronbach’s alpha=0.91) we combined the independent ratings, resulting in one rating per subject. This rating of reflection quality is our measure of *negotiation knowledge*.

Finally, all subjects performed a test in which they played the VR negotiation scenario again. We recorded negotiation outcome and conversation skill. Subjects were again asked to play as well as possible.[1]

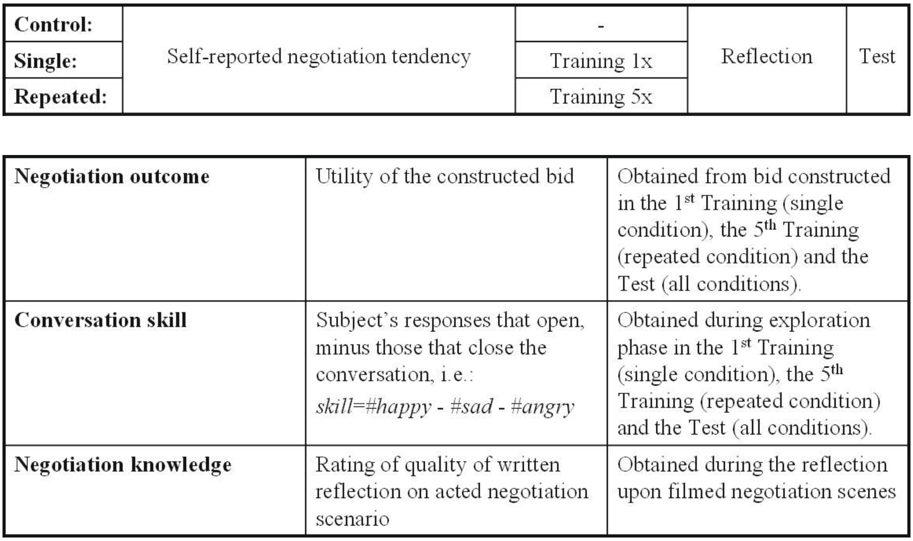
Subjects (mean age=24.7, std=4.5) were recruited of two different universities, and were gender balanced accross conditions, resulting in 7 males and 7 females in each condition totalling 42 subjects, 14 in each condition. Assigment to the conditions was random. The conditions did not differ significantly in age (ANOVA F(2, 39)=2.241, p=0.120), nor in self-reported negotiation tendency (ANOVA F(2, 39)=1.585, p=0.218).

4.2 Main Results

To investigate if VR training increases negotiation skill and knowledge, we performed a multivariate ANOVA with training condition as independent variable and test conversation skill, test negotiation outcome and negotiation knowledge as outcome measures.[2]. We found a significant effect of training (Wilks’ Lambda F(6, 74)=2.668, p=0.021). In detail we found a significant effect on negotiation knowledge (F(2, 39)=4.315, p=0.020) and conversation skill (F(2, 39)=3.668, p=0.035), but not on negotiation outcome (F(2, 39)=0.593, p=ns). Contrasts (LSD method) showed that lack of training results in a significantly lower rating for negotiation knowledge (mean=2.39, std=1.11) than a single session (mean=3.5, std=1.14, p=0.011) or repeated sessions (mean=3.39, std=1.04, p=0.021), and that lack of training results in a significantly lower (mean=-0.93, std=4.46) conversation skill than repeated training (mean=4.71, std=5.74, p=0.01). None of the contrasts between the three conditions were significant or even approached significance for negotiation outcome (all $p>0.32$). Conversation skill after a single training did not significantly differ from either no training or repeated training (mean=2.43, std=6.27, p=0.12 and p=0.28 respectively). This confirms our hypothesis. Training has a positive influence on negotiation knowledge and conversation skill. Apparently more training is needed for gaining skill,

[1] Emotional facial expression and explanations were omitted, as we will use this test as baseline performance in future experiments aimed at testing the influence of emotion and explanation as separate factors. A second reason to omit these is that they are informative means of feedback aimed at learning, while we wanted to use this as a test.

[2] Gender effects were non-significant in a MANOVA with condition and gender as independent variables (Wilks’s Lambda F(6, 68)=2.438, p=0.081), and no interaction effect between gender and training was found.



Control:	Self-reported negotiation tendency	-	Reflection	Test
Single:		Training 1x		
Repeated:		Training 5x		

Negotiation outcome	Utility of the constructed bid	Obtained from bid constructed in the 1st Training (single condition), the 5th Training (repeated condition) and the Test (all conditions).
Conversation skill	Subject's responses that open, minus those that close the conversation, i.e.: *skill=#happy - #sad - #angry*	Obtained during exploration phase in the 1st Training (single condition), the 5th Training (repeated condition) and the Test (all conditions).
Negotiation knowledge	Rating of quality of written reflection on acted negotiation scenario	Obtained during the reflection upon filmed negotiation scenes

Fig. 2. Experimental conditions and protocol (top); dependent variables (our outcome measures; bottom).

but a single session of about 10 minutes is enough for gaining implicit knowledge as measured by the quality of reflection on filmed scenes.

As reflecting upon scenes could overshadow the effect of training on negotiation outcome, we performed a simple ANOVA with single versus repeated training as independent variable and negotiation outcome taken from the single training session and the 5th repeated session *before the reflections* as outcome measure. The effect of single versus repeated training approached significance (F(1, 26)=4.002, p=0.056) with higher negotiation outcome for repeated training (mean=4.21, std=1.31) compared to a single session (mean=3.29, std=1.13). To analyse the main effect of reflection, we performed a within-subject multivariate repeated measures ANOVA with reflection as independent variable and conversation skill and negotiation outcome as dependents. We found a marginally significant effect of reflection (F(2, 26)=2.807, p=0.079), that was significant only for conversation skill (F(1, 27)=5.480, p=0.027) with pre-reflection conversation skill being lower (mean=0.32, std=7.09) than post-reflection negotiaton skill (mean=3.57, std=6.02).

4.3 Addional Analyses

In this section we highlight several trends and findings that are not directly related to our main hypothesis, but are relevant for negotiation training.

Gender Effects. Gender effects approached significance for negotiation knowledge (F(1, 36)=3.55, p=0.068) and conversation skill (F(1, 36)=4.03, p=0.052). Female participants had lower negotiation knowledge ratings (mean=2.79, std=1.20) than males (mean=3.40, std=1.11) and they had lower post-reflection conversation skill ratings (mean=0.81, std=4.61) than males (mean=3.76, std=6.63). We found a significant effect of gender on self-reported negotiation tendency (F(1, 40)=11.380, p<0.01), with female participants reporting a lower tendency (mean=2.21, std=0.57) than males (mean=2.90, std=0.74). We did not find significant difference between male and female participants when it comes to negotiation outcome, neither pre-or post reflection.

Conversation-Outcome Relation We found pre- and post-reflection conversation skill to correlate with pre- and post-reflection negotiation outcome (r=0.832, p<0.01; r=0.688, p<0.01; respectively). This indicates that subjects with a high conversation skill are also good at reaching a high outcome, which is interesting for two reasons: (a) it shows that the training is coherent (better exploration = better bidding), and (b) increasing conversation skill through VR training is a useful goal, as the two are linked.

We found a significant correlation between self-reported negotiation tendency and pre- and post-reflection conversation skill (r=0.412, p=0.029; r=0.329, p=0.033; respectively), and marginally significant correlations between self-reported negotiation tendency and pre- and post negotiation outcome (r=0.348, p=0.069; r=0.280, p=0.073; respectively). We interpret these findings as indicating that the tendency to negotiate is an indicator of actual negotiation skill.

5 Discussion and Conclusion

Our main results confirm our hypothesis that VR training has a positive effect on negotiation conversation skill and negotiation knowledge. Further, reflecting upon filmed scenes has a positive effect on conversation skill in the VR training. Our results support recent findings of others showing positive learning effects of VR training [20,7]. To assess if developing a VR training is worth the effort, future work should investigate differences between VR training and traditional training methods such as paper-based materials and role playing. Our results also show that there is (a) transfer from the training to the quality of reflections on a negotiation of others, and (b) transfer from reflection to conversation skill in the VR training. This highlights the importance of controlling for outcome measurement effects as measurement itself can overshadow or interact with the actual manipulation, a point stressed by our finding that *before* reflection the effect of single vs repeated training approached significance but not *after* reflection.

Our results show that *although* VR training increases conversation skills and knowledge, it did not automatically result in a better negotiation outcome, *even though* there is a strong correlation between individual ratings of conversation skill and outcome. We interpret this as follows: good negotiators already understand the link between exploration and bidding, while those that do not understand this link can get knowledge and conversation skills out of VR training but need additional forms of teaching (e.g., explicit instruction or negotiation rules) in order to understand the link. This lack of a learning effect is consistent with work described in the negotiation training literature [23] concluding that pure experience-based learning is largely ineffective. Our results nuance this conclusion slightly by supporting the following view: although experience-based learning does not positively influence the joint outcome, experience-based learning does positively influence negotiation knowledge and conversation skills. We hypothesize that the reason for the lack of a positive effect on the actual joint outcome of the negotiation is due to participants' inability to bridge the gap between exploring the negotiation space and translating the results of the exploration into a concrete bid. Future work should investigate if different negotiation phases need different learning approaches.

We did not observe a clear effect of gender when it comes to negotiation outcome, conversation skills and knowledge related to VR training and reflection. However, lower self-reported negotiation tendency for women does indicate that further study towards gender differences in negotiation training should be done, especially since our results seem to indicate that this tendency to negotiate is related to negotiation skill in the VR training.

Acknowledgments. This research is supported by the GATE project (NWO and ICT regie) as well as the Pocket Negotiator VICI-project 08075 (STW). Special thanks to Barnier Geerling from iamstudios for the voice acting, Mark van Gurp from Praction for evaluating the negotiation scenario and the filmed scenes, and Wietske Visser and Jeroen Meijer for rating reflection quality.

References

1. Andreatta, P.B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., Stern, S., Frankel, J.: Virtual reality triage training provides a viable solution for disaster-preparedness. Academic Emergency Medicine 17(8), 870–876 (2010)
2. Brinkman, W., Hartanton, D., Kang, N., de Vliegher, D., Kampmann, I., Morina, N., Emmelkamp, P.M.G., Neerincx, M.: A Virtual Reality Dialogue System for the Treatment of Social Phobia (page in press, 2012)
3. Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., Meyer, J.-J.: Do You Get It? User-Evaluated Explainable BDI Agents. In: Dix, J., Witteveen, C. (eds.) MATES 2010. LNCS, vol. 6251, pp. 28–39. Springer, Heidelberg (2010)
4. Broekens, J., Jonker, C., Meyer, J.-J.: Affective negotiation support systems. Journal of Ambient Intelligence and Smart Environments 2, 121–144 (2010)
5. Broekens, J., Qu, C., Brinkman, W.-P.: Factors influencing user perception of affective facial expressions in virtual characters (submitted)
6. Core, M., Traum, T., Lane, H., Swartout, W., Gratch, J., Van Lent, M.: Teaching negotiation skills through practice and reflection with virtual humans. Simulation 82(11), 685–701 (2006)
7. Durlach, P.: Cultural awareness and negotiation skills training: Evaluation of a prototype semi-immersive system. Technical report, DTIC Document (2008)
8. Emmelkamp, P., Bruynzeel, M., Drost, L., van der Mast, C.: Virtual reality treatment in acrophobia: a comparison with exposure in vivo. CyberPsychology and Behavior 4(3), 335–339 (2001)
9. Fisher, R., Shapiro, D.: Beyond reason: using emotions as you negotiate. Random House Business Books (2005)

10. Fisher, R., Ury, W., Patton, B.: Getting to yes: negotiating agreement without giving in. Houghton Mifflin Harcourt (1991)
11. Gomboc, D., Solomon, S., Core, M.G., Lane, H.C., van Lent, M.: Design recommendations to support automated explanation and tutoring. In: Proc. of BRIMS 2005, Universal City, CA (2005)
12. Grantcharov, T.P., Kristiansen, V.B., Bendix, J., Bardram, L., Rosenberg, J., Funch-Jensen, P.: Randomized clinical trial of virtual reality simulation for laparoscopic skills training. British Journal of Surgery 91(2), 146–150 (2004)
13. Harbers, M., Broekens, J., van den Bosch, K., Meyer, J.-J.: Guidelines for developing explainable cognitive models. In: Proceedings of ICCM 2010, pp. 85–90 (2010)
14. Hays, M.J., Ogan, A., Lane, H.C.: The Evolution of Assessment: Learning about Culture from a Serious Game, pp. 37–44 (2010)
15. Hindriks, K., Jonker, C.: Creating human-machine synergy in negotiation support systems: Towards the pocket negotiator. In: Brinkman, W.-P. (ed.) Proc. of the 1st Int. Working Conference on Human Factors and Computational Models in Negotiation, HuCom 2008, Delft, pp. 47–54 (2008)
16. Johnson, L.: Agents that learn to explain themselves. In: Proceedings of the Conference on AI, pp. 1257–1263 (1994)
17. Keil, F.: Explanation and understanding. Annual Reviews Psychology 57, 227–254 (2006)
18. Kim, J.M., Hill, J.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)
19. Krijn, M., Emmelkamp, P.M.G., Olafsson, R.P., Biemond, R.: Virtual reality exposure therapy of anxiety disorders: A review. Clinical Psychology Review 24(3), 259–281 (2004)
20. Chad Lane, H., Hays, M.J., Auerbach, D., Core, M.G.: Investigating the Relationship between Presence and Learning in a Serious Game. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 274–284. Springer, Heidelberg (2010)
21. Malle, B.: How people explain behavior: A new theoretical framework. Personality and Social Psychology Review 3(1), 23–48 (1999)
22. Mantovani, F.: Vr learning: Potential and challenges for the use of 3d environments in education and training. In: Riva, G., Galimberti, C. (eds.) Towards Cyberpsychology: Mind, Cognition, and Society in the Internet Age, pp. 207–225 (2001)
23. Nadler, J., Thompson, L., Boven, L.V.: Learning negotiation skills: Four models of knowledge creation and transfer. Management Science 49(4), 529–540 (2003)
24. Ortony, A., Clore, G.L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press (1988)
25. Parsons, S., Mitchell, P.: The potential of virtual reality in social skills training for people with autistic spectrum disorders. Journal of Intellectual Disability Research 46(5), 430–443 (2002)
26. Powers, M.B., Emmelkamp, P.M.G.: Virtual reality exposure therapy for anxiety disorders: A meta-analysis. Journal of Anxiety Disorders 22(3), 561–569 (2008)
27. Raiffa, H.: The Art and Science of Negotiation. Harvard University Press (1982)
28. Reznek, M., Harter, P., Krummel, T.: Virtual reality and simulation: Training the future emergency physician. Academic Emergency Medicine 9(1), 78–87 (2002)

29. Rizzo, A.S., Kim, G.J.: A swot analysis of the field of virtual reality rehabilitation and therapy. Presence: Teleoperators and Virtual Environments 14(2), 119–146 (2005)
30. van den Spek, E.: Experiments in Serious Game Design. SIKS Dissertation series. University of Utrecht, Utrecht (2011)
31. Thompson, L.L.: The Heart and Mind of the Negotiator. Pearson Prentice Hall, Upper Saddle River (2005)
32. Thomson, J., Tolmie, A., Foot, H., Whelan, K., Sarvary, P., Morrison, S.: Influence of virtual reality training on the roadside crossing judgments of child pedestrians. Journal of Experimental Psychology: Applied 11(3), 175 (2005)
33. Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: Proc. of IAAA 2004. AAAI Press, Menlo Park (2004)

Towards Multimodal Expression of Laughter

Radosław Niewiadomski[1] and Catherine Pelachaud[2]

[1] Telecom ParisTech, Paris, France
radoslaw.niewiadomski@telecom-paristech.fr
http://perso.telecom-paristech.fr/~niewiado
[2] CNRS-LTCI, Telecom ParisTech, Paris, France
catherine.pelachaud@telecom-paristech.fr
http://perso.telecom-paristech.fr/~pelachau

Abstract. Laughter is a strong social signal in human-human and human-machine communication. However, very few attempts to model it exist. In this paper we discuss several challenges regarding the generation of laughs. We focus, more particularly, on two aspects a) modeling laughter with different intensities and b) modeling respiration behavior during laughter. Both of these models combine a data-driven approach with high-level animation control. Careful analysis and implementation of the synchronization mechanisms linking visual and respiratory cues has been undertaken. It allows us to reproduce the highly correlated multimodal signals of laughter on a 3D virtual agent.

Keywords: multimodal animation, expression synthesis, laughter, intensity.

1 Introduction

Laughter is one of the most frequently used communicative signals. It is mostly associated with positive reactions to humorous stimuli, but it can also be a social signal such as an indicator of social position or relations [1] or even be a conversation regulator (e.g. such as a punctuator) [2]. Recent studies [3,4] enumerate at least 23 different types of laughter such as angry, anxious, desperate, hysterical or contemptuous or sulky laughter. Laughter plays an important part in an interaction; it is a very contagious behavior [2]. It may also have a positive impact on health [5].

Humans are very sensitive to laughter. The social and communicative quality of laughter is crucial in human-human interaction. It is not surprising there is a new interest in laughter modeling for human-machine interaction, in particular when the machine takes the appearance of a virtual agent. Virtual agents might use laughter to communicate their intentions and social attitudes or to improve the relations with their human interlocutor. However, so far, there exist only few interactive systems ([6,7], see Section 2) that make use of laughter and of its contagious effect. There are even fewer attempts of laughter synthesis.

Laughter is a highly multimodal expression in which different modalities are highly synchronized. In laughter, the body movements and the tight

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 231–244, 2012.

synchronization between audio and visual signals of the expression is crucial. Laughter is composed of very quick rhythmic shoulders and torso movements, visible inhalation, several facial expressions which are often accompanied with some rhythmic as well as communicative gestures [8]. Recent studies on laughter suggest that the various types of laughter may have different expressive patterns [4]. Consequently, even a small incongruence in laughter synthesis may influence its perception. This makes its synthesis particularly challenging. Careful attention needs to be put on the synchronization between modalities which seems to be a key factor to successful laughter synthesis. Thus laughter synthesis requires, first of all, a good understanding of the underlying physiological processes, the relations between the modalities as well as the communicative meanings of laughter cues.

Our long term goal is to build a virtual agent able to laugh believably and multimodally. In this paper we focus more particularly on the modeling of laughter visual intensity and on the respiration animation. In our approach we combine several existing animation techniques such as data-driven and procedural animation. In the remaining of this paper we present our model and the data we used.

This paper is structured as follows. The next section is dedicated to the description of the existing works on laughter synthesis. Then in Section 3 we discuss the challenges of the laughter synthesis. Next, we focus on two aspects of laughter: in Section 4 we present an approach for modulating perceived intensity of laughter while Section 5 is dedicated to the study and the animation of respiration during laughter. Finally we conclude the paper in Section 6.

2 State of Art

Only few visual laughter synthesis models were proposed so far. In existing approaches the generation of laughter animation is often driven by some audio parameters. Cosker and Edge [9] propose a data-driven model for non-speech related articulations such as laugh, cries etc. The model based on HMM is trained from motion capture data and audio segments. First, the data are captured with the motion capture system Qualysis, with 30 markers placed on the face, and normalized to one chosen identity (i.e. one facial model). During the training phase, the model learns the correlations between the recoded audio and the visual data. For this purpose, the number of facial parameters is reduced using PCA, while MFCC is considered for the audio input.

DiLorenzo et al. [10] propose a physically-based parametric model of human chest that can be automatically driven from prerecorded audio laugh samples. The model uses an anatomically inspired and physics-based model of a human torso that is a combination of rigid-body and deformable components. It is based on the assumption that there exists a relation between lung pressure and the laughter phase that can be derived from the amplitude of the audio signal. It takes into account the volume and lungs pressure, the air flow volume rate and the chest wall cavity. The model is restricted to the respiration during the laughter act; it does not involve other body moments. The animation is not generated in real-time.

While not directly related to laughter synthesis, some interesting works on respiration synthesis were proposed recently. De Melo et al. [11] study the role of respiration in expressing emotional states. Their respiration model is based on target morphing technique. They use 2 morph targets and morphing is applied to interpolate between these two targets. The model provides a set of parameters to control the respiration rate and depth. These parameters are manipulated manually to define the respiration profiles for 14 emotions. The evaluation shows that adding respiration improves significantly the perception of some emotions such as excitement, boredom or relief. Kider et al. [12] propose an anatomically inspired and data-driven model of human fatigue that includes, among others, respiration. The model is driven by data from different physiological sensors that control the visual appearance of a character. Regarding the respiration animation it uses an underlying anatomical model of the lungs. This approach is based on respiration sensor and on data on the vital capacity of lungs, expiratory reserve volume and tidal volume. Finally, the anatomical model of lungs is used to simulate realistic chest movement.

Recently laughter starts to play a more significant role in human-machine interaction. Urbain et al. [6] proposed the AudioVisualLaughterCycle machine, a system able to detect and respond to human laughs in real time. With the aim of creating an interaction loop between a human and an agent, the authors built a system capable of analyzing the user's laugh and of responding to it with a similar laugh produced by the virtual agent. This similar laugh is automatically chosen from an audio-visual laughter database. Its selection is done by measuring the acoustic similarities between the input laughter and the outputted one. The visual data corresponds to motion capture data of facial expressions. While the audio content of the similar laugh is directly replayed, the corresponding motion capture data is retargeted to a virtual model. Recently Fukushima et al. [7] built a system capable of increasing users' laughter reactions. It is composed of a set of toy robots that shake their heads and play prerecorded laughter sounds when the system detects user's laugh. An evaluation study shows that the system enhances users' laughing activity (i.e., it favors laugh contagion).

3 Laughter Synthesis

Laughter synthesis is a challenging task. A large quantity of movements occurs across modalities. Laughter is characterized by highly multimodal expressive patterns composed of different facial actions (see Section 4.1), head movements such as tilts, visible respiration (see Section 5.1), shaking of the shoulders, straighten or vibration of the trunk that are often completed with body inclinations, swinging, but also some gestures such as clapping hands or thighs [8]. The synchronization of all these modalities is critical. The preliminary audio-driven models of laughter described in the previous section are a first step towards the construction of a laughing virtual agent. However they often focus only on a single modality. Moreover their animation is not modifiable; e.g. it cannot be altered through an intensity parameter. These models still lack complete and time-efficient solution for body animation. In particular, compulsive chest and torso movements, which strongly contribute to laughter production, need to be captured and synthesized. At the same time capturing the data over the various modalities (face and body) at the same time is still considered a technical challenge. Hence, we believe it is important to develop an approach in which different modalities can be synthesized separately and then synchronized. It includes synchronization between audio and visual cues as well as between single communicative modalities such as face, head, torso and gestures. New approach should also permit to control synthesis process with the high-level easily interpretable parameters such as communicative intentions or intensity.

With the long aim of multimodal laughter synthesis in this paper we focus two aspects, namely, the modulation of the laughter visual intensity and the respiration animation. In particular, not much is known on which visual cues participate to the perception of laughter intensity. Interestingly the so-called silent laughter (i.e. when no sound is perceivable) can be perceived as highly intense [13]. Controlling the intensity of synthesized laughter is an important aspect of laughter synthesis. The intensity is a very intuitive high-level variable that can be used to control laughter synthesis and to generate laughs that are appropriate to the situation context and the communicative intentions of the laughing agent.

Respiration has also a particular role in laughter production. It is one of the most significant cues and is strongly visible. Moreover, as a physiological process, it drives the audio and visual expressive patterns and probably decides on the synchronization of different modalities.

3.1 Database and Annotation

For the purpose of this work we used the AudioVisualLaughterCycle (AVLC) corpus [14] that contains about 1000 spontaneous audio-visual laughter episodes with no overlapping speech. The episodes were recorded with the participation of 24 subjects. Each subject was recorded watching a 10-minutes comedy video. Thus, it is expected that the corpus contains mainly only one type of laughter namely hilarious one. Each episode was captured with one motion capture system (either Optitrack or Zigntrack) and synchronized with the corresponding audiovisual sample. The material was manually segmented into episodes containing just one laugh. The number of laughter episodes for a subject ranges from 4 to 82.

Next, through perceptive study, human coders annotated the perceived intensity of the AVLC laughter episodes using a Likert scale from 1 (low intensity) to 5 (high intensity) (for details see [13]). Each episode was manually annotated by minimum 6 and maximum 9 coders. In the rest of the paper we use a part of this dataset: 1528 intensity annotations for 249 laugh episodes of AVLC corpus (corresponding to 6 subjects). The intensity coder's agreement, Krippendorff κ coefficient, for this part of the dataset is 0.65.

Finally in the respiration study we used the manual annotation of the respiration phases proposed by Urbain and Dutoit [15]. Each laugh episode in AVLC corpus [14] was manually annotated to indicate the airflow direction (inhalation or exhalation).

4 Laughter Intensity Modulation

In this section we propose a solution to model laughter with different intensity values. Our rational is to modify laughter production so that its intensity is perceived differently. We use a data driven approach for facial animation and we propose a method to modulate automatically the intensity of the laughter expression.

At first, we carried out a study to discover which facial cues are related to the perception of intensity level. We rely on the annotation of perceived intensity of laughter (see Section 3.1). We link the perceived intensity with facial cues. Facial cues are characterized by a set of distances between facial markers (recorded with motion capture system). They were chosen as they correspond to specific muscular activities (related to action units [16]). For each episode of the AVLC corpus we extract such distances between the facial markers. We are particularly interested in the cues that can be associated with intense laughs. Having the values of these distances, we check their correlation with perceived intensity. Then we use such results to elaborate a computational model that modulates any animation of laughter depending on its intensity level.

4.1 Facial Expression of Laughter

To define the cues that are significant in laughter expression we look at the facial actions that occur in laughter expressions. Laughter contains usually the activation of the zygomatic major muscle that corresponds to the action unit AU12. Additionally the cheek raise is often present; it corresponds to the activation of the orbicularis oculi muscle (action AU6). Its presence is associated with the Duchenne or hilarious laughter. There are other facial actions that may occur in the expression of laughter. For example mouth opening (AU25) and jaw drop (AU26) were observed quite often in the study of Beermann et al. [17]. The same study reports, albeit more rarely, the occurrence of the lid tightener action AU7 and of the lip corner depressor action AU15. In the acted expressions of hilarious laughter, the actions AU25, AU26 are frequently observed while AU7, AU27 (mouth stretch), AU4 (frown) occur less and AU1 (inner raise eyebrow), AU2 (outer raise eyebrow), AU9 (nose wrinkling) and AU20 (lip stretcher) occur occasionally [18]. On the other hand in naturally occurring laughter, actions AU5 (upper lid raiser), AU6, AU7, and AU9 are observed [8]. Apparently some of these actions may be particularly related to laughter intensity. Indeed, Darwin [19] claimed that intense laughter will include lowering of the eyebrows (AU4). Other actions such as AU7 and AU9 are also sometimes considered to be an indicator of strong laughter.

4.2 Data Analysis

We analyzed the motion capture data of the AVLC corpus [6] which was annotated with intensity values. We introduced 12 measures (D1 - D12) that correspond roughly to the action units observed in the laughter expressions: AU6,

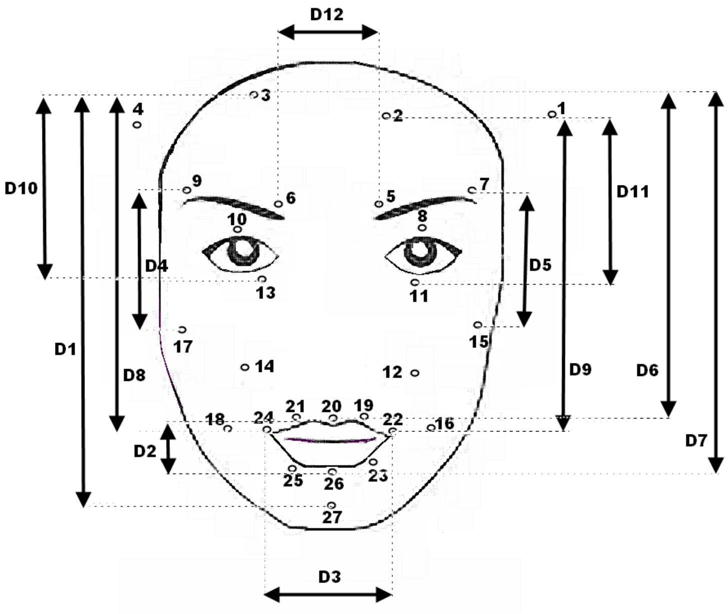


Fig. 1. Position of the markers and distances D1-D12

AU4, AU7, AU12, AU25, AU26 (see Figure 1). The computed distances are normalized to the neutral expression in the motion capture data. The distances used in our study are presented in Table 1.

The measurements D4-D5 and D8-D9 correspond roughly to action units considered to be specific for the facial expression of hilarious laughter, namely cheek raising - AU6 and smile (lip corner up) - AU12. The remaining measurements correspond to the action units which may occur in certain laughs (see Section 4.1) i.e. AU4 (D12), AU25 and AU26 (D1, D2, D3, D6, D7), AU7 (D10-D11). All these distances are computed at 25 FPS.

Table 1. Distances D1-D12

id	name	value	direction
D1	jaw movement	3-27	vertical
D2	lip height	20-26	vertical
D3	lip width	24-22	horizontal
D4	cheek raising (right)	17-9	vertical
D5	cheek raising (left)	15-7	vertical
D6	upper lip protrusion	3-20	depth
D7	lower lip protrusion	3-26	depth
D8	lip corner movement (right)	24-3	vertical
D9	lip corner movement (left)	22-2	vertical
D10	lower eyelids movement (right)	3-13	vertical
D11	lower eyelids movement (left)	2-11	vertical
D12	frown	5-6	horizontal

For each laugh episode, the values D1-D12 of single frames are mapped to a fixed-length feature vector with the help of the following functions: minimum, maximum, range and mean. Since we have 12 facial distances and 4 functions, we obtain a feature vector of 48 features per episode. We calculate the correlations between distances D1-D12 and the intensity annotations of the corresponding laugh episode (see Section 3.1). The detailed data are presented in Table 2.

Table 2. Correlation between laughter median intensity and the distance features

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
min	0.40	0.34	0.14	0.19	0.11	0.17	0.27	0.20	0.17	-0.23	-0.22	0.10
max	0.67	0.63	0.46	0.27	0.29	0.20	0.58	0.33	0.44	-0.09	-0.06	0.38
range	0.49	0.44	0.42	0.39	0.41	0.40	0.51	0.34	0.28	0.45	0.49	0.21
mean	0.63	0.61	0.42	0.25	0.25	0.19	0.52	0.31	0.42	-0.18	-0.16	0.31

The strongest correlation is observed for the maximum jaw and lip openings, i.e. the distances D1 and D2, using the "max" function computed over the whole episode ($\rho = .67$ and .63, respectively). Strong correlation is also observed for maximal lower lip protrusion (D7) ($\rho = .58$). All these three measures received comparable strong correlations when computed as a mean over the whole episodes. On the other hand, these three distances correspond to the activation of the action units AU 25 and AU 26. This might suggest that the perceived degree of intensity is correlated with the mean and maximal activation of AU 25/26 and, in other words, with the mouth opening. Relations between the perceived intensity and the other distances were less strong. In our test the correlation between the perceived intensity and the measures D4 and D5 (cheek raise) was weak ($\rho = .27$ and .29). It does not mean that this activity was not observed in the dataset. Indeed cheek raise is present in the considered episodes but its presence is not so strongly correlated with the perceived intensity. We obtained similar results for AU12. The correlation between perceived intensity and the measurements D3, D8, and D9 is only slightly higher (0.33-0.46 for maximum function, and 0.31 - 0.42 for mean function) than for the distances corresponding to AU6. Distances D10-D11 corresponding to the lower eyelids movement (AU7) are even less correlated with the perceived intensity. This result, however, might be influenced by the eyes blinking. Finally also D12 that corresponds to AU4 (frowning) is slightly correlated with the perceived intensity ($\rho = .38$).

Interestingly, the overall duration of the laugh is not strongly correlated ($\rho =$ .54, see Figure 2) with the perceived intensity. In other words, an intense laugh does not necessarily last long, and vice-versa.

These results show that only some visual features are strongly related to the perceived intensity of laughs. In other words, rather than a simple linear increase of the activation of all facial cues, intense laughter is expressed by activating some additional actions units, or by the gradual changes of single action units. This important conclusion will be used, in the next section, to synthesize laughs with the desired intensity.

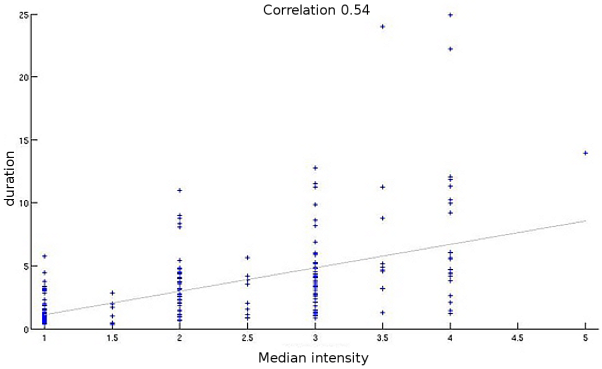


Fig. 2. Correlation between median intensity and laughter duration

4.3 Synthesis

The intensity of the facial expression is often modeled in virtual agents by using a simple linear function (such as proposed in [20]). In such approach all facial parameters of the face are multiplied by one intensity value. Thus the values of all facial parameters are proportional to the intensity value. The results of our study in Section 4.2 shows that such approach in the case of laughter expression would not be appropriate. For this reason we propose a novel approach for the modulation of the intensity of laughter animation. In our approach only facial parameters corresponding to certain actions units (AUs) are modulated while others are not. Moreover facial parameters corresponding to different facial actions are modified independently. Some of them are activated only for high intense expressions. The intensity modulation is done at the level of action units and thus it is independent from different facial animation parameterizations used to animate virtual agents. It can be applied for both procedural animation and motion capture data.

For the synthesis we use the virtual agent system that is MPEG-4 compliant. We modulate the intensity of any original expression according to the results of the analysis described in the previous section. Our intensity modulation module works as a filter that modulates any laughter animation. Figure 3 presents the pipeline. In this application we use the motion capture data from AVLC dataset. The original motion capture data is converted into Facial Animation Parameters FAPS (of MPEG-4 [21]) according to the procedure described in [6]. Next, MPEG-4 compliant animation is modulated by the intensity module described below and the final animation is displayed by the virtual agent. Additionally the original (prerecorded) audio file is synchronously played with the modified animation.

Our intensity module modulates the values of the facial animation parameters FAPS corresponding to AU25, AU26, AU4 and AU7 (see also Section 4.1). Values of FAPs corresponding to the actions AU25 and AU26 are modified proportionally to $\Delta(intensity)$ where $\Delta(intensity)$ is the difference between the perceived intensity of the original data and the requested intensity. The values of FAPS corresponding to AU4 and AU7 are activated only if the values of the

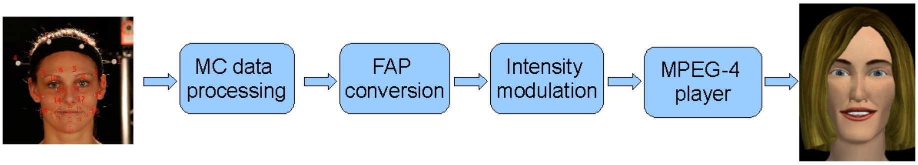


Fig. 3. Laughter intensity modulation

former exceed certain activation values. The FAPS corresponding to the AU12 and AU6 are not changed. MPEG-4 does not allow for the animation of AU9.

It is important to notice that once we have the animation described in facial animation parameters the intensity modulation can be done in real-time (data retargeting from the original motion capture data to facial animation parameters is not yet real-time).

Figure 4 presents several frames of an animation generated with our approach. In the first line one can see the original animation reproduced from the motion capture data. It serves as basis of comparison with the animation modified by the intensity modulation module. The video corresponding to this animation was perceived as medium intensity laugh (median value is 3 in 5 points scale from 1 to 5). In this example we increase this laugh intensity. One can see that the facial expressions on the frames of the modified animation (second row) are characterized by a stronger mouth opening. Additionally in the column Figure 4c) the action units AU4 and AU7 are visible. Thus the final animation is not a simple linear filter applied over the whole face but facial parts are differently modulated according to the empirical result on the perception of the intensity.

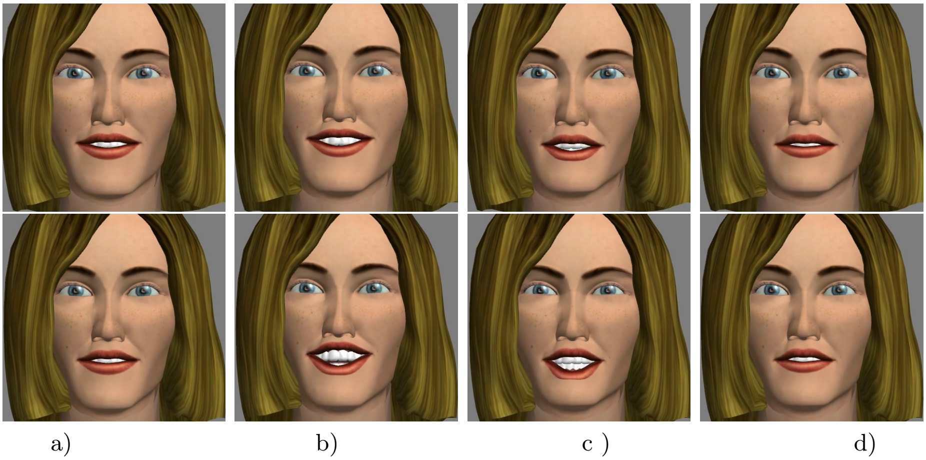


Fig. 4. Example of laughter episode with different intensities

5 Respiration

The visible respiration is an important part of expression of the laughter. A respiration cycle in laughter usually begins with an "initial forced exhalation", that is followed by a "sequence of repeated expirations" of high frequency and low amplitude. In longer laughter episodes the expiratory phases are interlaced with inhalations [8]. Interestingly the synthesis of the respiration is ignored in the virtual agent animation. During respiration, an important cue is the synchronization between modalities such as facial expressions and body movements. Indeed, facial, body and respiration cues are driven by the same physiological processes. It is shown that the two respiration phases correspond to different audiovisual patterns [13]. For example it is argued [8] that the backward tilt of the head facilitates the forced exhalations.

In this section we focus on the respiration during the laughter. First, we study the relation between the respiration phases and the facial cues. We extract some relations that characterize the synchronization mechanism across modalities in laughter. Second, from the physiological respiration data gathered through sensors, we replay the respiration animation with an MPEG-4 compliant agent.

5.1 Data Analysis

We studied the relation between the visual cues and the respiration phases. For this purpose we used again the data from the AVLC corpus and the manual annotation of the respiration phases developed by Urbain and Dutoit [15]. We considered 142 laughs that contain more than one respiration phase. They contain 190 exhalation and 190 inhalation phases. We also used the same 12 measurements, namely D1-D12, and we checked if they have different values in the inhalation and exhalation phases. For each considered respiration phase, the values D1-D12 corresponding to single frames were mapped to a fixed-length feature vector using 4 functions minimum, maximum, range, mean. We used two-sample Kolmogorov-Smirnov test to check whether D1-D12 have different values within the inhalation and exhalation phases. The detailed results are presented in Table 3.

According to the obtained results the distances D1-D12 differ significantly between two respiration phases for 2 functions: range and min. The 'max' values corresponding to the mouth opening (i.e. distances D1-D2) are significantly bigger in the exhalation phase. In other words, mouth opening is more intense in this

Table 3. Mean distance differences between exhalation and inhalation phases (significant values in bold)

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
min	**-0,80**	**-1,20**	**-1,66**	**-0,92**	**-0,91**	**-0,29**	**-1,27**	**-1,06**	**-1,02**	**-0,27**	**-0,23**	**-0,13**
max	**0,46**	**0,45**	0,36	0,26	0,30	0,16	0,61	0,28	**0,34**	**0,18**	**0,23**	0,02
range	**1,26**	**1,65**	**2,03**	**1,18**	**1,21**	**0,44**	**1,88**	**1,35**	**1,36**	**0,45**	**0,47**	**0,16**
mean	-0,08	-0,12	-0,14	-0,06	-0,02	-0,01	-0,11	-0,06	0,01	-0,09	-0,08	-0,04

respiration phase. The maximal distances D4-D5 and D8-D9 that correspond to AU6 and AU12 are also bigger in this phase but only for D9 this difference is significant. From these results, it emerges that in one phase the mouth is more often widely opened. This hypothesis should be however confirmed using more precise respiration data.

5.2 Synthesis

To generate the respiration animation we use the data gathered with the The ProComp Infiniti system[1] which serves for biofeedback data acquisition. This system contains high accuracy respiration sensor that provides the thorax and abdomen expansion with 256 samples per second. On the other side, MPEG-4 standard does not allow us to modify the body shape; it solely permits to animate the body skeleton. To simulate respiration and corresponding chest movement we had to extend the set of body animation parameters. The body of the agent is defined by bones. To simulate the chest movement we added two additional bones at the level of the thorax. Next, the normalized data from the respiration sensor are used to animate the additional bones independently to the rest of the body. Figure 5 presents frames of the body animation corresponding to inhalation and exhalation phases.

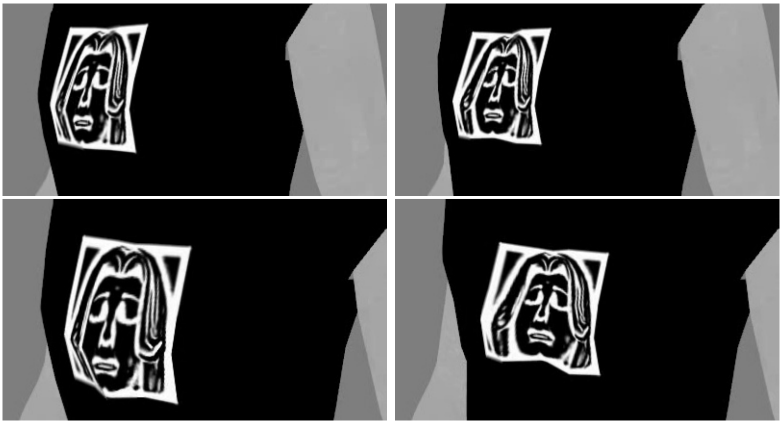


Fig. 5. Animation of two respiration phases: first column - inhalation phase, second column - exhalation phase

6 Conclusion and Future Works

The laughter synthesis is a very complex task that has not been much studied before. In this paper we discussed several issues related to laughter synthesis, in particular: intensity modulation and respiration modeling. We studied the

[1] http://www.thoughttechnology.com/index.html

relation between visual cues of laughter and the perceived laughter intensity, as well as between the visual features and laughter inhalation and exhalation phases. We also proposed a motion capture data-driven laughter intensity model and a physiological data based respiration animation.

Several limitations of this work should be noted. First of all, so far we have worked with only two communicative modalities: face and body motion during respiration. This work is also limited to only one laughter category i.e. hilarious laughter. Concerning data, we use the motion capture dataset corresponding to only 6 human subjects. More subjects have to be considered to allow us to consider individuals as well as intra-subjects differences in expressive behaviors. Indeed there may exist different "laughing styles" that may depend for example on personality factors (e.g. extraversion) or even on physical and physiological characteristics of the laughing person (e.g. weight, lungs volume). In the intensity analysis, so far, we do not consider the dependencies between different distance measurements of facial features. Very likely, some of these measurements are interrelated, e.g. this may be the case for the measurements D8-D9 and D4-D5. Other factors such as the duration of single facial actions in the laughter expression may influence the perceived intensity and, thus, should also be analyzed.

This is an ongoing work. At the moment we are preparing an evaluation of the intensity model. The evaluation will be organized through perceptive tests where the original and intensity modulated animations, as well as videos extracted from the original AVLC dataset will be evaluated by the naive users by using the same intensity scale. Regarding the respiration modeling, we are now working on the synchronization of the respiration animation with the facial animation. For this purpose we use results of the study described in Section 5.1. Last but not least we will extend our model by introducing other modalities. First of all we plan to integrate laughter audio synthesis. The analysis of the relation between certain acoustic parameters and the perceived intensity [13] shows that this relation can be even stronger than in the case of facial cues. Thus intensity model should be able not only to modify facial animation but also corresponding sound of laughter. At the same time we will also work with different corpora to introduce arm gestures to our model.

Acknowledgements. This work was partially supported by the European FP7-ICT-FET project ILHAIRE (grant n270780). We are very grateful to our ILHAIRE partners and in particular Jerome Urbain (University of Mons), Harry Griffin (University College London) and Maurizio Mancini (University of Genova) for their valuable help and for having given access to data.

References

1. Adelsward, V.: Laughter and dialogue: The social significance of laughter in institutional discourse. Nordic Journal of Linguistics 102(12), 107–136 (1989)
2. Provine, R.: Laughter. American Scientist 84(1), 38–47 (1996)

3. Huber, T., Ruch, W.: Laughter as a uniform category? a historic analysis of different types of laughter. In: 10th Congress of the Swiss Society of Psychology. University of Zurich, Switzerland (2007)
4. Huber, T., Drack, P., Ruch, W.: Sulky and angry laughter: The search for distinct facial displays. In: Banninger-Huber, E., Peham, D. (eds.) Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions, pp. 38–44. Universitat Innsbruck (2009)
5. Martin, R.: Is laughter the best medicine? humor, laughter, and physical health. Current Directions in Psychological Science 11(6), 216–220 (2002)
6. Urbain, J., Niewiadomski, R., Bevacqua, E., Dutoit, T., Moinet, A., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: AVLaughterCycle. enabling a virtual agent to join in laughing with a conversational partner using a similarity-driven audiovisual laughter animation. Journal of Multimodal User Interfaces 4(1), 47–58 (2010)
7. Fukushima, S., Hashimoto, Y., Nozawa, T., Kajimoto, H.: Laugh enhancer using laugh track synchronized with the user's laugh motion. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010), pp. 3613–3618 (2010)
8. Ruch, W., Ekman, P.: The expressive pattern of laughter. In: Kaszniak, A. (ed.) Emotion Qualia, and Consciousness, pp. 426–443. Word Scientific Publisher (2001)
9. Cosker, D., Edge, J.: Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations. In: Proceedings of Computer Animation and Social Agents, CASA (2009)
10. DiLorenzo, P., Zordan, V., Sanders, B.: Laughing out loud: Control for modeling anatomically inspired laughter using audio. ACM Transactions on Graphics 27(1) (2008)
11. Melo, C.D., Kenny, P., Gratch, J.: Real-time expression of affect through respiration. Computer Animation and Virtual Worlds 21, 225–234 (2010)
12. Kider, J., Pollock, K., Safonova, A.: A data-driven appearance model for human fatigue. In: Spencer, S. (ed.) Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2011), pp. 119–128. ACM, New York (2011)
13. Niewiadomski, R., Urbain, J., Pelachaud, C., Dutoit, T.: Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases. In: Proceedings of 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals, LREC 2012, Istanbul, Turkey (2012)
14. Urbain, J., Bevacqua, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., Picart, B., Tilmanne, J., Wagner, J.: The AVLaughterCycle Database. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp. 2996–3001. European Language Resources Association (ELRA) (2010)
15. Urbain, J., Dutoit, T.: A Phonetic Analysis of Natural Laughter, for Use in Automatic Laughter Processing Systems. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 397–406. Springer, Heidelberg (2011)
16. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
17. Beermann, U., Gander, F., Hiltebrand, D., Wyss, T., Ruch, W.: Laughing at oneself: Trait or state? In: Banninger-Huber, E., Peham, D. (eds.) Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions, pp. 31–36. Universitat Innsbruck (2009)

18. Drack, P., Huber, T., Ruch, W.: The apex of happy laughter: A FACS-study with actors. In: Banninger-Huber, E., Peham, D. (eds.) Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions, pp. 32–37. Universitat Innsbruck (2009)
19. Darwin, C.: The expression of the emotions in man and animals. John Murray, London (1972)
20. Bartneck, C., Reichenbach, J.: Subtle emotional expressions of synthetic characters. International Journal Human-Computer Studies 62(2), 179–192 (2005)
21. Ostermann, J.: Face animation in MPEG-4. In: Pandzic, I.S., Forchheimer, R. (eds.) MPEG-4 Facial Animation - The Standard Implementation and Applications, pp. 17–55. Wiley, Chichester (2002)

Ada and Grace: Direct Interaction with Museum Visitors

David Traum[1], Priti Aggarwal[1], Ron Artstein[1], Susan Foutz[3], Jillian Gerten[1], Athanasios Katsamanis[2], Anton Leuski[1], Dan Noren, and William Swartout[1]

[1] USC Institute for Creative Technologies, Los Angeles
[2] USC Signal Analysis and Interpretation Laboratory, Los Angeles
[3] Institute for Learning Innovation, Edgewater, Maryland

Abstract. We report on our efforts to prepare Ada and Grace, virtual guides in the Museum of Science, Boston, to interact directly with museum visitors, including children. We outline the challenges in extending the exhibit to support this usage, mostly relating to the processing of speech from a broad population, especially child speech. We also present the summative evaluation, showing success in all the intended impacts of the exhibit: that children ages 7–14 will increase their awareness of, engagement in, interest in, positive attitude about, and knowledge of computer science and technology.

Keywords: virtual human applications, natural language interaction, virtual museum guides, STEM, informal science education.

1 Introduction

Ada and Grace [14], a pair of twin virtual humans that were designed to engage visitors and increase their knowledge and appreciation of science and technology, have been in operation at the Museum of Science in Boston since December 2009. The Twins are physically located in a kiosk inside Cahners ComputerPlace, a room dedicated to exhibits about computers and related technologies (such as robots). They have been designed to answer a variety of questions about science and technology, museum exhibits, and themselves. The characters use speech recognition to recognize the words in a question and then use a statistical classifier to select the most likely answer from a pre-existing set of approximately 150 answers. The answer is then presented using coordinated speech, gestures, eye gaze and body movement.

In the first version, Ada and Grace did not directly interact with visitors. Instead, a museum staff member would wear a head-mounted microphone and pose questions to the Twins. Staff members could either ask their own questions, or they could field questions from the audience and relay them to the Twins. In this paper, we report on our efforts to add two additional interaction use cases with the exhibit: a *direct interaction* condition, in which museum visitors could talk to Ada and Grace without staff member intervention, and a *blended* condition in which both staff and visitors can talk to the Twins. Allowing visitors to interact directly with the characters raised a number of issues, which we address in Section 2. In Section 4, we present a summary of some of the results of an independent summative evaluation that was performed by the Institute for Learning Innovations (ILI).

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 245–251, 2012.

2 Improvements to the Twins to Enable Direct Interaction

Allowing visitors to interact directly with the visitors raised a host of issues related to the wide usability and performance of the system, the most important of which are described below.

Hardware. Initially, museum staff used a standard wired USB microphone to talk to the Twins. This was replaced in June 2010 by a wireless Sennheiser microphone, to allow more mobility and cut down on failure points in the constant use of the museum environment. For visitor use, we needed a fixed microphone that could focus on the voice of the speaker. We experimented with three microphones, a Sennheiser ME66, Shure SM58, and Shure 522. The narrow beam pattern of the Sennheiser ME66 gives superior pickup of quiet signals in a noisy environment, but we found that it was very difficult to get visitors to remain at the optimal distance and orientation. The Shure SM58 alleviates this problem, but visitors still had difficulty in associating the microphone with the separate, wireless press-to-talk switch. The current setup (since February 2011) uses the Shure 522, a desktop microphone with integrated press-to-talk button which is typically used for paging and dispatching applications.

Data Collection. In order to handle questions that visitors frequently ask, as well as be able to recognize them well, data was collected in the museum, which was then transcribed and coded for the speaker type (child, adult male, adult female, or no speech) [1]. A sampling of 17,244 utterances from April and May of 2011 revealed the following composition (identified by listening to the voice): 47% children, 13% adult male, 8% adult female, and 31% no speech. The questions were annotated with the correct answer if possible, or were marked either as questions for which answers could be constructed or questions that would be treated as "off-topic" [9]. Additional details about the corpus collected are given in [1].

Speech models. Speech recognition was performed by the SONIC toolkit [11] until December 2010, and thereafter by OtoSense, a recognition engine that is currently being developed by USC. The transcribed audio recordings from the museum visitors have been used for the adaptation of three separate acoustic models, namely for children's, adult male and adult female speech [12]. Adaptation was performed using Maximum Likelihood Linear Regression while the original children's models were trained on the Colorado University children's speech database [5] and the two adult speech models were trained on the Wall Street Journal corpus [10]. In the deployed system, the three recognizers are used in parallel, each providing an n-best list with confidence scores.

Audio acquisition. Incorporating both the multiple microphones and multiple acoustic models into the Twins system posed a significant engineering challenge. The original speech acquisition subsystem (Acquirespeech) was designed to support a single microphone connected to a single ASR engine. We redesigned Acquirespeech to work with multiple simultaneous inputs and multiple concurrent ASR engines. The current system starts five instances of the OtoSense ASR engine, connects the first two (generic male and female) to the staff microphone and the other three (child, adult male, adult female) to the visitor microphone. Both microphones operate in a push-to-talk mode with separate physical control buttons. Acquirespeech handles both the control button clicks and audio inputs from multiple microphones independently. In theory, both the visitor and the staff member can be talking to Twins at the same time and it's the task of the response selection process to select between two inputs.

When Acquirespeech receives audio input from a microphone, it forwards it to all linked ASR engines concurrently. For example, audio from the visitor microphone goes to all three visitor speech engines simultaneously. Acquirespeech waits for the text response from each engine, selects the one with the highest confidence score, and passes it along to the response selection component (NPCEditor [8]).

Content enhancements. The exhibit was also extended and improved in several ways, based on experiences during the initial period. Captions were added for the Twins output to aid the hearing impaired. We also provided optional indicators on the processing stages (listening, thinking, responding), so that visitors could regulate the timing of asking questions. Many animations were also improved.

An optional "idle" dialogue behavior was added, such that if no one talks to the Twins for a threshold time (usually set to 10 minutes), then the Twins would start talking to each other, to cue visitors that they could ask questions.

A number of changes were made to the language understanding and response selection component. One set of changes was to remove directions to exhibits that had been removed from the space, such as the AI Dome and the Computer Build Bench. However the information about the scientific knowledge related to the exhibits (such as cell phones) was retained. Additional answers about new exhibits, such as Coach Mike's arrival in Robot Park [7] were added. The data collection was also used to identify frequent questions that were not understood or did not have a good answer. We added several answers to handle these classes, such as people speaking to the Twins in other languages, insults and hazing, and asking about dinosaurs.

We also provided a card of about 10 example questions, to help visitors get a sense of the types of questions they could ask. In a sample of over 20,000 utterances asked by visitors, 30% were identical to one of the posted suggestions [1].

3 Performance Evaluation

Automatic speech recognition using the three acoustic models has been systematically evaluated only for a small but representative portion of the Twins corpus comprising 1003 utterances recorded on a single day. The average word error rate was found to be 57%, when automatically selecting the model that has the highest confidence score (the initial configuration in the museum, used at the time of the summative evaluation); this resulted in a response selection accuracy of 42%. The best performing individual model is the child model, with a 53% overall word error rate and response accuracy of 45%; however, using an oracle to choose the best performing model for each utterance lowers the word error rate to 43% and raises the response accuracy to 53%, suggesting that better speaker identification should lead to improved performance overall. The relatively high word error rate is linked to the challenges of the museum setting: (1) Speech is spontaneous, i.e., with frequent hesitations and mispronunciations. (2) Speech is coming mainly from children (76% of the sample). (3) There are no vocabulary constraints on what visitors can say.

4 Summative Evaluation

The redesigned exhibit, together with an accompanying exhibit highlighting the science behind the Twins, was subject to a summative evaluation from an external, independent evaluator, the Institute for Learning Innovation (ILI). The study was designed to address two primary questions: 1) What is the nature of visitors' interactions with the Twins and Science Behind exhibits? and 2) In what ways do interactions with the exhibits impact visitors' knowledge and awareness of, engagement and interest in, and attitudes and perceptions towards computer science and technology? The intended impacts are shown in Table 1. Overall, 15 indicators were identified across the four impact areas. The summative evaluation was designed to determine whether the exhibits achieved these indicators, and therefore, the visitor impacts. The evaluation demonstrated that 14 of these indicators were achieved, as shown in Table 1.

Table 1. Intended impacts of the Twins and Science Behind exhibits

Impacts	Indicators	
Children (ages 7–14) and adults will	Tested	Achieved
– increase their **engagement and interest** in computer science and technology.	5	5
– have a **positive attitude** about computer science and technology.	2	2
– increase their **awareness** about computer science and technology.	5	4
– increase their **knowledge** about computer science and technology.	3	3

Two conditions were tested: direct visitor interaction, and blended staff and visitor interaction. Three methods were used in the study: observation of visitors while they interacted at the exhibits, in-depth interviews with visitors after their interaction, and follow-up online questionnaires 6 weeks after the initial interaction. Observational data included group size and composition, stay time, types of social interaction (between the target visitor and other visitors and between the target visitor and museum staff/volunteers), usability issues encountered while using the exhibit, the number and types of questions that the visitor addressed to the Twins, categorization of the Twins' responses, and visits to the Science Behind exhibit. Interviews were conducted after visitors engaged with either exhibit with the goal of collecting a paired observation and interview with the same participant. Children under 16 years of age were interviewed only after the data collector obtained permission from an adult family member in the visiting group. Interviews included open-ended questions and rating scale questions for use with all visitors designed to elicit visitor interest, attitudes, awareness, and knowledge of themes related to the visitor impacts. Only adult participants were asked to complete retrospective-pre/post-experience ratings in order to measure change in attitude and awareness as a result of the experience.

Observational and interview data were collected at the museum between July 21 and September 11, 2011; online questionnaires were collected between August 20 and October 26, 2011. A total of 225 observations were collected, 180 of which were paired with interviews (for a refusal rate of 20%). A total of 61 follow-up online questionnaires were collected (for a response rate of 42%).

In this paper, we present a selection of the results from the summative evaluation study showing the combined results for both condition and illustrating each of the impact areas. In most cases the trends are the same for the direct and blended condition, however in some cases there are significant differences between the conditions. See [4] for the complete results of the summative evaluation.

Engagement and Interest. Time spent in the exhibit ranged from 19 seconds to just nearly 18 minutes, with a median time of 3 minutes and 7 seconds ($N = 221$). Quantitative rating scale questions were used to determine whether participants had a positive experience at the exhibit. Participants were asked to rate the statements "Interacting with the exhibit" and "Learning more about computers by interacting with the Twins" on a four point scale, where 1 was "boring" and 4 was "exciting." The overall rating for both statements was a median of 3, or "pretty good" on the 4-point scale. Participants were asked this same question six weeks later in the follow-up online questionnaire; researchers compared ratings from the interview and follow-up questionnaire. Ratings remained the same six weeks following the original visit (Wilcoxon Signed Rank Tests).

Attitudes. The same quantitative rating scale question was also used to determine if participants had positive attitudes towards speaking with the Twins, by asking them to rate the statement "Being able to speak with the Twins" . The overall rating for all participants was a median of 3, or "pretty good" on the 4-point scale. As with the engagement questions above, ratings remained the same six weeks following the original visit (Wilcoxon Signed Rank Tests).

An additional quantitative approach to assess visitors' attitudes was used only with adults, to determine if interacting with the exhibit impacted self-reported agreement with the four statements: 1) "I enjoy being able to speak to a computer as a way to interact with it," 2) "Having a computer with a personality is a good thing," 3) "In the future, there will be new and exciting innovations with smarter computers," and 4) "In the future, interacting with computers will be easier." Adults reported a significantly higher rating for all of these measures of attitudes towards computers/virtual humans directly after their interaction with the exhibit.

Awareness. Five indicators were used to indicate awareness. For the statements "I understand what a virtual human is" and "Women have made important contributions in the field of computer science", adults showed significantly higher agreement post than retrospective-pre. In answers to open-ended questions, over 90% of visitors were able to describe the Twins as a computer that acts like a human, and recognize interaction characteristics of the Twins. However, only 39% of participants noted aspects of the connection between the Twins and the main subjects of Cahners (computers, communications, robots) or described the Twins as guides for the space.

Knowledge. To determine whether study participants recognized aspects of computer science needed to create a virtual human, open-ended responses were coded for the presence of five aspects (communications technology, artificial intelligence, natural language, animation/graphics, and nonverbal behavior). 97% of all participants mentioned at least one aspect, while 73% mentioned two or more aspects. The most commonly mentioned aspect was natural language, mentioned by 86% of participants. 64% of onsite participants named at least one technology needed to build a virtual human; this rose to 90% in the follow-up six weeks later. Finally, 84% of participants gained at least one additional understanding about STEM (Science, Technology, Engineering, and Mathematics) domains related to the Twins, while 59% of participants indicated they learned something new about computers or technology from interacting with the exhibit.

5 Related Work and Discussion

There are other efforts, both past and present, to put virtual characters in museums as guides. An early system is Max [6], who was placed at the Heinz Nixdorf Museums-Forum in Paderborn, Germany in 2004. Like the Twins, Max is projected life-size on a screen, and communicates using speech and body animations. Max can engage in both reactive and deliberate conversational behavior, and visitors communicate with him by typing on a keyboard. An analysis of interactions between visitors and Max showed that visitors treat Max conversationally as a person, evidenced by conventional strategies of beginning and ending conversations and general cooperativeness [6].

Another virtual museum guide is Tinker [3], who has been situated since April 2008 in the Museum of Science in Boston, just around the corner from the Twins. Tinker builds relations over time, recognizing a visitor who returns for a second conversation; relations bring about gains in visitors' attitudes toward, engagement with, and learning from Tinker [3]. Users communicate with Tinker through menus on a touch screen.

The main difference between the above two systems and the Twins is the input modality – the Twins understand human speech, allowing unmediated, naturalistic interaction with visitors. Systems similar to the Twins in this regard are Sergeant Blackwell [13], exhibited at the Cooper-Hewitt Museum in New York in 2006–2007 as part of the National Design Triennial exhibition, and Furhat [2], who was shown for four days in 2011 at the Robotville exhibit in the Science Museum in London. These systems do not share the Twins' educational goals, but they do understand speech input and employ a variety of techniques to overcome noisy and difficult-to-recognize speech.

This paper described extensions to the Virtual Human Museum Guides, bringing the system from one that can be demonstrated by an expert to one that can interact directly with visitors (or blending the two, having both interact). This involved hardware, software and content changes. The independent summative evaluation showed that despite some remaining challenges in language understanding performance, the exhibit successfully impacts visitors, as intended, realizing some of the capability for virtual humans to aid in informal science education.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant 0813541.

References

1. Aggarwal, P., Artstein, R., Gerten, J., Katsamanis, A., Narayanan, S., Nazarian, A., Traum, D.: The Twins corpus of museum visitor questions. In: LREC 2012, Istanbul, Turkey (May 2012)

2. Al Moubayed, S., Beskow, J., Granström, B., Gustafson, J., Mirnig, N., Skantze, G., Tscheligi, M.: Furhat goes to Robotville: A large-scale multiparty human-robot interaction data collection in a public space. In: Edlund, J., Heylen, D., Paggio, P. (eds.) LREC Workshop on Multimodal Corpora, Istanbul, Turkey, pp. 22–25 (May 2012)
3. Bickmore, T., Pfeifer, L., Schulman, D.: Relational Agents Improve Engagement and Learning in Science Museum Visitors. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 55–67. Springer, Heidelberg (2011)
4. Foutz, S., Ancelet, J., Hershorin, K., Danter, L.: Responsive virtual human museum guides: Summative evaluation. Tech. rep., Institute for Learning Innovation, Edgewater, Maryland (2012)
5. Hagen, A., Pellom, B., Cole, R.: Children's speech recognition with application to interactive books and tutors. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003), pp. 186–191 (2003)
6. 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)
7. 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)
8. Leuski, A., Traum, D.: NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine 32(2), 42–56 (2011)
9. Patel, R., Leuski, A., Traum, D.: Dealing with Out of Domain Questions in Virtual Characters. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 121–131. Springer, Heidelberg (2006)
10. Paul, D.B., Baker, J.M.: The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the DARPA Speech and Natural Language Workshop, pp. 357–362. Harriman, New York (1992), http://acl.ldc.upenn.edu/H/H92/H92-1073.pdf
11. Pellom, B., Hacıoğlu, K.: SONIC: The University of Colorado continuous speech recognizer. Tech. Rep. TR-CSLR-2001-01, University of Colorado, Boulder (2001/2005), http://www.bltek.com/images/research/virtual-teachers/sonic/pellom-tr-cslr-2001-01.pdf
12. Potamianos, A., Narayanan, S.: Robust recognition of children's speech. IEEE Transactions on Speech and Audio Processing 11(6), 603–616 (2003)
13. Robinson, S., Traum, D., Ittycheriah, M., Henderer, J.: What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco (2008)
14. Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu, S., White, K.: 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)

Spatial Cues in Hamlet

Christine Talbot and G. Michael Youngblood

Game Intelligence Group,
University of North Carolina at Charlotte
{ctalbot1,youngbld}@uncc.edu

Abstract. Many games, films, and virtual environments are very scripted, or use very canned/explicit cut scenes for characters to interact. This requires extensive work for producing new scenes, actions, and other scripts for these environments. It also usually comes with a certain level of expertise in lower-level character control. Current research focuses primarily on the conversational and non-verbal domains of this issue. However, with the growing focus on realistic virtual environments, the spatial domain is becoming a more critical component in creating that realism.

Tools and markup languages, such as Behaviour Markup Language (BML), Functional Markup Language (FML), and BML Realizers are making it possible to abstract the control of virtual characters to a certain extent. Unfortunately, these methods still require a level of expertise and time that can be unreasonable. Therefore, we propose a higher level of abstraction to ease this authorial burden for new scenes and actions in games. To do this, we look to another example of scripted activities that has been used for hundreds of years: play-scripts.

We took the fully annotated play-script for Gielgud's Hamlet in 1964, along with the recording of the same play to create a baseline using BML. We compared this to the use of just the play-script and some very simplistic natural language processing to block the same act of the same play. The results of this comparison show a savings of over four hours for authoring spatial scripts, while maintaining similar spatial blocking.

1 Introduction

Plays use spatial cues and blocking, provided by the director, to control the actors on the stage. In this paper, we take a well-known play and compare the level of effort and accuracy of utilizing a BML interface versus a standardized prompt-script to control the actors on the stage. A Shakespearean play was chosen because of the fact that they represent at least three of the top ten most performed plays in North America, and most top ten lists disqualify Shakespeare plays from their list of the top 10 because the competition would simply be unfair [1]. We then chose Hamlet because of the availability of a rather famous production which is available in both video [2] and as a detailed prompt-script from the director—the 1964 Hamlet played by Richard Burton and directed by John Gielgud [3]. The performance was filmed during three successive stage performances in June/July 1964 by Electronovision, Inc. [3], provides us insight into

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 252–259, 2012.

the actual performance of the annotations within the script, and is available on YouTube in its entirety. This production closed after 138 performances, setting records as the longest-running Hamlet ever to play New York [3].

We utilized a 2D simulation for our experiment (top-down view of the stage), however a BML realizer (like SmartBody) could have been used to provide a 3D view as well. The prompt-script and video were used together to translate the director's annotations into physical actions and movements for our 2D simulator environment. This became our ground-truth for comparing our techniques to spatially control the characters on the stage. We then utilized the Natural Language Toolkit (NLTK) [4], along with some core rule-sets and assumptions (theatre rules, Shakespeare play rules, conversational space rules, and stage directions) to generate BML, which controls the characters.

2 Background

When we give directions to people, we often have a layer of implied meaning built into it. For instance, to give directions in an office, you might tell someone to go down the hall and take the elevator to the fourth floor. Implied in these directions are things like how far is it to the end of the hall; the elevator is within sight when you get there, so you do not mention you have to turn right and go a few feet to the elevator; you do not instruct them how to work the elevator, you assume they know to press the button and wait for it to arrive. We look to provide the capability to continue that level of abstraction with controlling the virtual characters as well.

While reviewing play-scripts, you notice a similar level of abstraction and assumption within the director annotations as we use in every-day language. Shakespeare plays happen to be a genre with very few director annotations in them, unless you can find a director's annotated version. This leads to very different interpretations of his plays, and may contribute to their popularity even after over 400 years [5]. Modern plays tend to have more annotations included in the published versions than the original Shakespearean plays. Richard Burton's Hamlet play is one of the few exceptions to this due to Richard Sterne's book, which includes detailed annotations. The scripts tend to utilize stage directions, such as stage left, center stage, and upstage, along with specific marks and props to guide the actors to appropriate locations. Assumptions are also made that the actors understand some of the hidden rules behind performing these scripts, such as avoid putting your back to the audience, try to keep towards center stage as much as possible, primary characters should be closer to the audience than secondary characters, and general personal space and conversational rules.

We utilized some natural language processing techniques to translate scripted acts into BML. We start with some very basic one-to-one mapping of words, and a simplistic actor-verb-target sentence structure. This removes the authorial burden of creating the BML directly and reduces the time required for authoring a new script. Along with this, we need to incorporate some basic rules to apply to the natural language translations for better quality. For instance, we can say that characters should look at the current speaker, face the audience as much as possible/avoid turning their back to them, be placed closer to the viewer based on their character importance, and so forth. With these tools, we can translate any scripted act/play that follows standard playwright formatting and control the characters in a scene dynamically and without having to hard-code these kinds of spatial movements and acts for everything the author wants to do.

Applications such as SmartBody [6] or Elckerlyc [7] execute behaviours specified by BML on the character in the environment. Per the BML standards, "BML describes the physical realization of behaviours (such as speech and gesture) and the synchronization constraints between these behaviours. BML is not concerned with the communicative intent underlying the requested behaviours." [8] BML is structured like a typical XML message. You can control what is done, when it is done, and what runs concurrently with other commands. However, it is often at such a low-level that this can be extremely time-consuming to build, especially for things like non-verbal behaviours (eye saccade, gesturing while speaking, head nods, and so forth).

The focus of much research has involved virtual characters; however, very little of this work has investigated spatial movement of those characters. The emphasis appears to be more on the speech and emotional interaction with humans or other characters. For instance, Dias proposed changes to the FAtiMA (FearNot Affective Mind Architecture) to include the skill of understanding emotions of others in determining next steps [9]. The FAtiMA architecture was built to create autonomous believable characters that allowed the establishment of empathetic relationships with other characters in the FearNot! system [10].

There are several approaches like the Virtual Storyteller, which enable characters to tell a story with the appropriate gestures, prosody, and so forth [11]. Here, they focus on plot and story creation, mostly in the area of interactive storytelling. For instance, Kriegel proposes a design to help solve the authoring problem for interactive storytelling utilizing the FAtiMA architecture [12]. Thespian expands on these to reduce the programming effort for the speech actions of a story by pre-authoring sections of the speech and utilizing goals to control choices by the characters [13].

Next, let us look at how a typical play-script is formatted so we can use it as a guideline for our formatting. We look at the three types of stage directions that will include our spatial commands we wish to interpret as a part of this effort: scene, stage, and character directions. Examples of this formatting can be seen at Script-Frenzy's website [14].

3 Methodology and Experimentation

Using common sense rules, we translate the speech and annotations of the play to appropriately place the characters on the stage at the right times, doing the right things. Because our goal is to decrease the authorial burden for producing scripted acts that involve spatial movements and actions, we will need to utilize some natural language processing to translate components of the play-script.

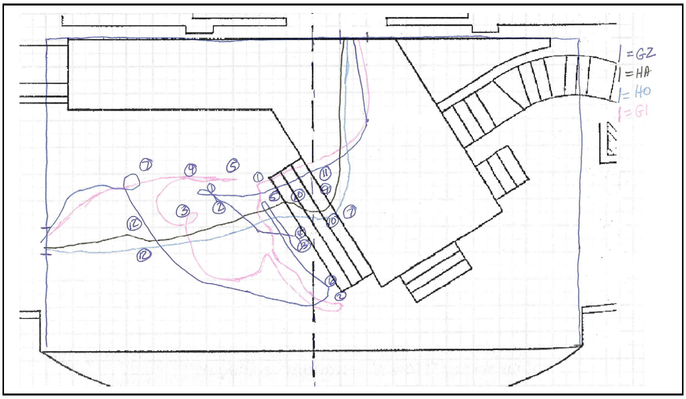


Fig. 1. Hand-mapped blocking of Richard Burton's Hamlet play from 1964

As a first pass, we parse the spatial directions (surrounded by parentheses) to determine the action within those statements and translate them into one of our spatial motions such as walking, pointing, gazing, or picking up an object [4]. This process can also be applied to the speech text within the play to extract things that the character may be doing at that point in time as well. Once this information is extracted, it will be combined with our basic rules to determine what action should be occurring (e.g., characters should face the current speaker).

We manually mapped out about ten minutes of Act III, Scene II from Hamlet, as produced by John Gieguld in 1964 (Figure 1). This is the graveyard scene where Hamlet reminisces about a skull that may have been Yorick, an old friend. The play consists of 280 lines and actions when mapped following the play-script standards for formatting, with the additional annotations provided by Sterne and the video. The position of each of the characters were hand-mapped against the stage layout, utilizing the recording of the 1964 play as a guideline since a direct physical mapping of the character locations were not available. Key aspects captured included walking, pointing, gazing/turning, and picking up/carrying objects at specific moments during the play. These movements were the focus of the spatial aspects of the play, which could be rendered in 2D, and were converted into BML. This resulted in 400 BML and FML commands with similar spacing between character speech and act lines as exist in the play-script. This is a 142.86% increase in commands that were needed to be written to accommodate just the four spatial aspects of moving, pointing, gazing, and picking up objects for a ten minute performance. All of this effort required four hours and twelve minutes, just for writing the BML to support these four actions.

First, a simple application was built to visualize the results of the BML and FML in a 2D environment. This application and BML script became our ground truth for determining how well our method could provide similar spatial controls, while reducing the technical effort and time required to author the script (as seen while creating BML).

Next, we utilized a simplistic natural language processor to identify the actor, what they are doing (of our four identified spatial movements), and to

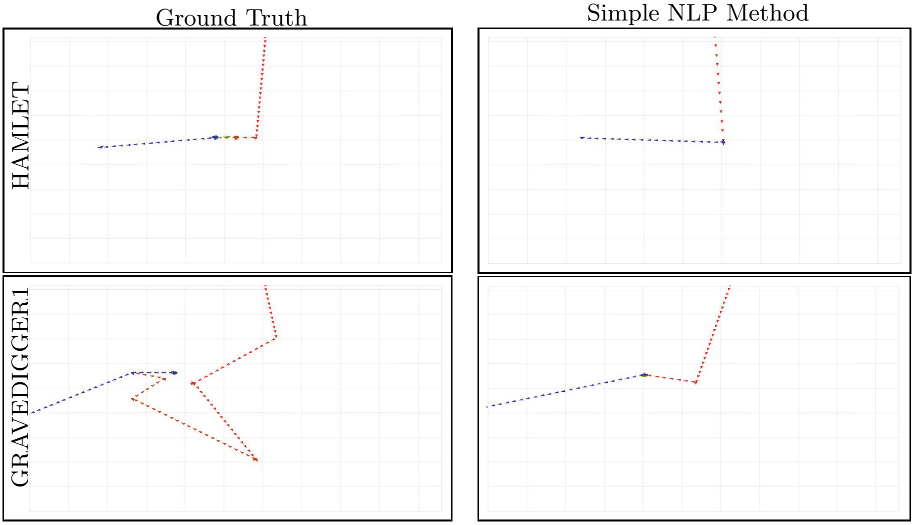


Fig. 2. Comparisons of Character Traces in Hamlet Over Time (Red to Blue)

whom/what they are doing it. Due to the nature of most play-scripts, we decided to focus on the basic noun-verb-noun structure of spatial commands within the script. Sentences are parsed to determine the verbs and nouns. The verbs and their synonyms are each reviewed against a list of synonyms for our key spatial movements (walk, turn, point, and pick-up). Meanwhile, the nouns and their synonyms are each reviewed against our known objects—four characters, Shovel, Lantern, and our nine basic stage positions. Taking the verbs and nouns we identify, we make the assumption that these sentences will take on the basic form of "actor action target." We then generate and send the BML to our simulator to perform the action.

Assumptions were made in this approach due to our understanding of typical play-script contexts, including our simplistic sentence structures. Typically, director's annotations are short and to the point. Often, they are just barely sentences, if not sentence fragments. Therefore our expectation was that the sentence fragments would contain very little information outside of the actor, action, and target. Other assumptions were made about the timing of these spatial events. All sentences, or sentence fragments, within a single set of parentheses were assumed to be independent of each other and required to be acted upon at the same time. These were also to be performed with whatever the next speech action was, unless we were changing the speaking character. The basis for this assumption comes from a basic understanding of how scripts are acted and formatted. Directions are provided before or in the middle of whatever is being said by the characters.

We took the character traces from both our ground truth (hand-coded BML) and our method and compared them. These traces focus on position and gaze direction (direction of arrows) for the characters throughout their time on-stage.

We want our new method to result in character positioning as close to our baseline as possible, however we do not want to penalize for being "close enough." As you can see in Figure 2, we accomplished a similar character trace over time with our new method, although not an exact match. We see that we were able to accomplish a reasonable blocking for this play, thereby saving us more than four hours of work and technical expertise for these ten minutes of script.

One key reason for some of the discrepancies in the character traces is due to the input utilized for the ground-truth versus our method. The ground-truth BML was written to include movements that were not included in the play-script that our method utilized. It included some movements based on what was seen in the video. When comparing one-to-one with what is actually in the play-script, our method recognized the majority of the spatial commands correctly. Other differences between the traces were related to typical natural language processing issues, such as pronoun grounding issues (who "he" refers to), multiple meaning words (steps or hands), compound statement parsing ("He is followed by GRAVEDIGGER2, who carries a T-spade and a pick and whistles"), and actor versus recipient identification. Also, because we are working in 2D, the annotations for the play are 3D in nature, and the annotations include how to say/speak certain items; many statements end up being irrelevant for our 2D model and are discarded. For instance, "(laughing)," or "(The sound of the bell fades out)" have no actions in an environment without sound, however still require processing to determine the sentence is irrelevant for this work. Some other issues also arose from the fact that this script was written in British English and the dictionary utilized (WordNet) was American English (due to availability).

4 Conclusion

The results using the natural language processing techniques indicate promise in retrieving spatial directions from a play-script. We were able to reduce the technical expertise required to write the script by over four hours. We were able to accomplish similar, albeit not exact, movements without requiring any technical BML or FML expertise from the author (see Figure 2). The more robust natural language processing will become necessary as we explore the wider range of motion that comes with moving in a 3D environment.

References

[1] Kabialis, B.D.: When Shakespeare Dependence Hits Point of Diminishing Returns, http://www.berkeleybeacon.com/arts/2012/2/2/when-shakespeare-dependence-hits-point-of-diminishing-returns

[2] Colleran, B., Gielgud, J., Shakespeare, W., Burton, R., Cronyn, H., Drake, A., Herlie, E.: Hamlet. Electronovision Taping (1964)

[3] Sterne, R.L.: John Gielgud Directs Richard Burton in Hamlet. 5th edn. or later edition. Random House (1967)

[4] Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. In: ETMTNLP 2002 Proceedings of the ACL 2002 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, vol. 1, pp. 63–70 (2002)

[5] Mabillard, A.: The Chronology of Shakespeare's Plays (August 2000), http://www.shakespeare-online.com/keydates/playchron.html

[6] Feng, A., Xu, Y., Shapiro, A.: An Example-Based Motion Synthesis Technique for Locomotion and Object Manipulation. In: I3D (2012)

[7] Welbergen, H., Reidsma, D., Ruttkay, Z.M., Zwiers, J.: Elckerlyc. Journal on Multimodal User Interfaces 3(4), 271–284 (2010)

[8] Marshall, A., Vilhjálmsson, H., Kopp, S., Kipp, M., Krieger, M.: Wiß ner, M., Tepper, P., Homer, J., Welbergen, H.V., Hill, A., Bickmore, T., Gruber, J.: Behavior Markup Language (BML) Version 1.0, Proposal (2011), http://www.mindmakers.org/projects/saiba/wiki/Wiki/

[9] Dias, J., Paiva, A.: Agents with Emotional Intelligence for Storytelling. In: D'Mello, S., Graesser, A., Schuller, B., Martin, J.-C. (eds.) ACII 2011, Part I. LNCS, vol. 6974, pp. 77–86. Springer, Heidelberg (2011)

[10] Dias, J., Paiva, A.C.R.: 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)

[11] Theune, M., Faas, S., Heylen, D.K.J.: The Virtual Storyteller: Story Creation by Intelligent Agents. In: Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment TIDSE Conference (2003) (2003)

[12] Kriegel, M.S.O.M., Science, C.: An Authoring Tool for an Emergent Narrative Storytelling System. AAAI Fall Symposium - Technical Report, p 55-62, Intelligent Narrative Technologies - Papers from the AAAI Fall Symposium, Technical Report FS-07-05 (2007)

[13] Si, M., Marsella, S., Pynadath, D.: Thespian: Using Multi-agent Fitting to Craft Interactive Drama. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 21–28. ACM (2005)

[14] Abigail-Nicole: How to Format A Stage Play - Script Frenzy, http://www.scriptfrenzy.org/howtoformatastageplay

[15] Si, M., Marsella, S.C.S., Pynadath, D.V., Rey, M.: Evaluating Directorial Control in a Character-centric Interactive Narrative Framework. In: Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010), pp. 1289–1296 (2010)

[16] Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D.: Building Interactive Virtual Humans for Training Environments. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2007, NTSA (2007)

[17] Thórisson, K., Vilhjalmsson, H.: Functional Description of Multimodal Acts: A Proposal. In: AAMAS 2009 Workshop Towards a Standard Markup Language for Embodied Dialogue Acts, p. 31 (2009)

[18] Games, R.: Rockstar Games Presents: L.A. Noire., http://www.rockstargames.com/lanoire/agegate/ref/?redirect=

[19] Frank, A.: Qualitative Spatial Reasoning: Cardinal Directions as an Example. International Journal of Geographical Information (February 2012) (1996) 37–41

[20] 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)

[21] Leuski, A., Pair, J., Traum, D., McNerney, P.J., Georgiou, P., Patel, R.: How to Talk to a Hologram. In: IUI 2006 Proceedings of the 11th International Conference on Intelligent User Interfaces, pp. 360–362 (2006)

[22] Jan, D., Traum, D.R.: Dynamic Movement and Positioning of Embodied Agents in Multiparty Conversations. In: Proceedings of the Workshop on Embodied Language Processing, EmbodiedNLP 2007, pp. 59–66. Association for Computational Linguistics, Stroudsburg (2007)

[23] Sundstrom, E., Altman, I.: Interpersonal Relationships and Personal Space: Research Review and Theoretical Model. Human Ecology 4(1), 47–67 (1976)

[24] Sommer, R.: The Distance for Comfortable Conversation: A Further Study. Sociometry 25(1), 111–116 (1962)

[25] Vidal Jr, E., Nareyek, A., et al.: A Real-Time Concurrent Planning and Execution Framework for Automated Story Planning for Games. In: Workshops at the Seventh Artificial Intelligence and Interactive Digital Entertainment Conference (September 2011)

[26] 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)

[27] Aggarwal, P., Traum, D.: The BML Sequencer: A Tool for Authoring Multi-character Animations. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 428–430. Springer, Heidelberg (2011)

[28] Mateas, M., Stern, A.: Integrating Plot, Character and Natural Language Processing in the Interactive Drama Façade. Computer 2

[29] Cavazza, M., Charles, F., Mead, S.J.: Agents' Interaction in Virtual Storytelling. In: de Antonio, A., Aylett, R.S., Ballin, D. (eds.) IVA 2001. LNCS (LNAI), vol. 2190, pp. 1–15. Springer, Heidelberg (2001)

[30] Tomaszewski, Z.: On the Use of Reincorporation in Interactive Drama. In: Workshops at the Seventh Artificial Intelligence and Interactive Digital Entertainment Conference (September 2011)

[31] Kenny, P., Hartholt, A., J.G.W.S.D.T.S.M.D.P.: ICT Virtual Human Toolkit, http://vhtoolkit.ict.usc.edu/index.php/Main_Page

[32] Causey, T.: Where is the Best Place to Sit in the Theater? http://theater.about.com/od/faqs/f/faqbestseat.htm

[33] Primrose, J.: Theatrecrafts - Entertainment Technology Resources, http://www.theatrecrafts.com/

[34] Alderson, M.: Theatre Types, http://www.ia470.com/primer/theatres.htm

[35] Talbot, C.: Virtual companions and friends. In: ACMSE Conference, pp. 356–357 (2011)

Interactive Stories and Motivation to Read in the Raft Dyslexia Fluency Tutor

Arthur Ward[1], Margaret McKeown[3], Carol Utay[2],
Olga Medvedeva[1], and Rebecca Crowley[1]

[1] University of Pittsburgh, Department of Biomedical Informatics
[2] Total Learning Centers, Wexford, Pa.
[3] University of Pittsburgh, Learning Research Development Center

Abstract. Improving motivation to read is an important step toward improving reading fluency among children with dyslexia. We describe a test of the RAFT dyslexia fluency tutor which dynamically generates an interactive story by assembling pre-written scene templates in response to user-chosen plot events. Results suggest that interactivity improved voluntary reading, relative to non-interactive versions of the story.

1 Introduction

Dyslexia affects a very large percentage of our children, with recent estimates of its prevalence ranging as high as 15% to 20% [1–3]. Children with dyslexia often have average to above average intelligence, but their specific disability in decoding words leads them to dislike and avoid text [1]. This avoidance can reduce their word decoding practice by as much as a million words per year [4], relative to normal readers. Dislike of text, therefore, contributes to the persistence of poor reading fluency among children with dyslexia, even after their individual word decoding problems have been remediated. We are developing the RAFT (**R**epeated **R**eading **A**daptive **F**luency **T**utor) tutor to remediate poor reading fluency among children with dyslexia. Because of the link between motivation and word exposure, it is important that the RAFT tutor improve motivation to read.

Classroom studies often find that students feel more motivated to read when they can exercise some choice over their reading materials [5,6]. Similarly, a meta-analysis of studies in the self-determination literature [7] found that many varieties of choice can improve motivational and performance outcomes, as well as perceptions of self-efficacy. In addition, the "serious game" literature has found that interactive games can keep students motivated and engaged (e.g. [8,9]). These studies suggest that the RAFT tutor might increase reading motivation by allowing students with dyslexia to make choices in a dynamically generated interactive story. In this work we implement an interactive story generator in RAFT, to test if interactivity leads children with dyslexia to read more.

While there have been a wide variety of approaches to generating interactive stories, the majority of them seem to implement interactivity by allowing

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 260–267, 2012.

users to influence animated story characters [10–12], often by engaging them in conversation. Designers of these systems (e.g. [13, 14]) must trade off between maximizing user interactivity and maximizing story coherence. Systems can maximize interactivity by creating a set of autonomous agents which interact with each other and with the user, however the resulting "story" can lack coherence. Tale-spin [15] was an early example of agent driven story generation. At the other end of the spectrum, maximal coherence can be assured by hand authoring all the episodes in a branching story tree, as was done in the "Choose your own Adventure" series of books (e.g. [16]). Hybrid systems choose various mid-points on this spectrum. Systems such as In-Tale provide coherent guidance to reactive software agents using a declarative story plan [10]. Façade [17] employs a rule-based drama manager to choose from a large pool of pre-authored "dramatic beats," selecting beats which meet certain goals such as level of dramatic tension. In contrast, the Thespian system [18, 19] endows its character agents with weighted goals which allow them to choose appropriate actions if the plot deviates from a pre-authored path.

A very useful system for conceptualizing these trade-offs is provided by Spierling [20]. Spierling discusses four levels of abstraction in story generation. At the highest, most abstract level is the overall plot structure. Below that, in increasing order of specificity, are the individual scene; individual actions/conversations within a scene; and the details of avatars which embody those actions and conversations. Each of these levels can be implemented by software agents with varying ratios of interactivity to authorial control. For example a "scene" could be mandated by the author, while the specific implementation of that scene could be handed down to autonomous agents at the next level.

Our application domain influences our position at each of these levels. First, because children with dyslexia are very skilled at using contextual clues such as pictures to guess the identity of words, visual avatars are not used in the Raft tutor. Second, because our emphasis is on reading motivation, we want to assure that our stories are highly coherent. This leads us toward retaining authorial control over the "scene" and "actions/conversations" level. Instead, we allow user interactivity at the "story" level, by allowing the user to make plot choices which influence which of these pre-written scenes is generated next.

2 The Raft Tutor

A pilot version of the Raft story generation architecture was written in Clips [21], and reported in an earlier paper [22]. That paper gives additional details about how production system rules are used to assemble text. For the purposes of the current study our generation system was re-written in the Drools [23] production system shell. In general, the Raft system maintains a set of data structures whose variables represent the state of the people and locations in the story world. For example, "character" objects include the person's location, goals, mood, role (i.e. "protagonist") and whatever other variables the author deems important to the story. "Location" objects describe important physical states, such as if a door is open or closed. Rules fire if their “if” portion matches the current state of the story world as represented by the “character” and “location” objects. A firing rule will append some text to the developing story, and also change the relevant story variables so that the story world remains consistent with the generated text. If more than one rule matches the current story world state, the story world is forked, and stories are generated for both forks. The production-rule paradigm was chosen because of its ability to match multiple rules to one story state. This provides a natural mechanism for generating multiple story branches, as well as (in future work) multiple lexical realizations of each branch.

```
rule wakeFather
    when
        $protag: character($proLoc: goal == ”wakeFather” ...)
        $protagFather: character($fatLoc: location == ”basement” ...),
    then
        change((object_Story)$protag, ”setLocation”, ”downstairs”);
        change((object_Story)$protag, ”setGoal”, ”takeFatherToRoom”);
        change((object_Story)$protagFather, ”setLocation”, ”downstairs”);
        postTemplate(1, rn, ”Now ”, $proFn, ” was really getting scared.
She reached up and gave her father a hard slap. Now he seemed to be coming
around. He focused his eyes on her for a long moment and said”, $proFn, ” I
don’t feel very well.”);
end
```

Fig. 1. Rule describing an episode

A truncated version of a story-generating rule is shown in Figure 1. This rule fires if the protagonist’s goal is to “wakeFather” and the father is in the basement (among other conditions). The rule outputs some text which is included in the example passage shown in Figure 2, where the variable for protagonist’s name has been instantiated to “Anna.” The rule also updates the story variables for protagonist’s goal and father’s location (several other updates are not shown). In this way, the system can assemble chunks of pre-written text, while assuring that those chunks are not appended inappropriately.

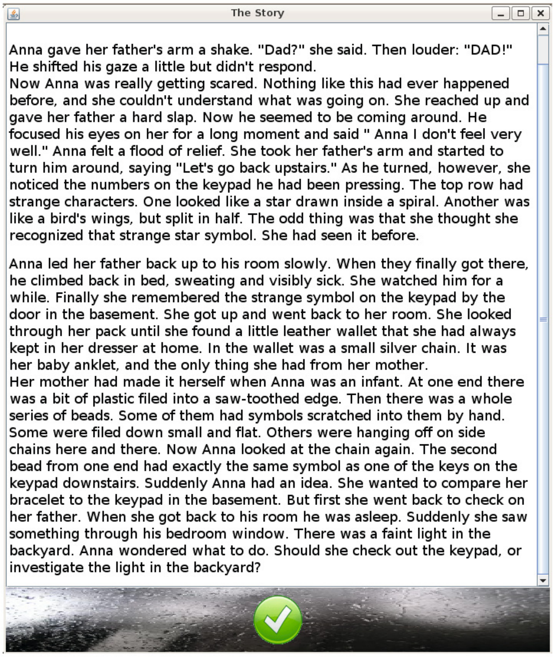
Most assembled scenes end by setting up a plot choice, as shown in the bottom of Figure 2. When the student has finished reading the passage, and clicked the “check mark” button, the system asks for the appropriate user decision in a “What next?” screen (not shown). In our example, the “What next” choices will be “Check out the keypad?” and “Check out the light?” Earlier choices can be revisited using a “WAYBACK” panel. Students can alter earlier plot choices if the story has taken a bad turn, for example if the protagonist gets killed or has a major setback.

The generation system presents its stories on a Viewsonic tablet computer. System data including the users’ plot choices and the stories generated by those choices are logged to a relational database.

3 Study Design

As discussed in Section 1, we hypothesized that interactive text would improve motivation to read among children with dyslexia. To test this hypothesis, we employed a "yoked design" study to compare students who read "interactive" stories to those who read "static" stories generated by the system.

With University of Pittsburgh IRB approval (IRB # PRO11060291), subjects were recruited from among students of the Total Learning Centers (TLC). These students had previously been diagnosed with dyslexia by a school psychologist, and had received further testing and remedial training at TLC. Students received a nominal amount for participation. Before scheduling, subjects were assigned to pairs based on their reading fluency, which had been previously measured using the Gray Oral Reading Test (GORT) [24]. Scheduling limitations dictated that whichever student in the pair could be scheduled first would be assigned to the "interactive" condition, with the effect that the student's exogenous schedules served as our randomization procedure.



Anna gave her father's arm a shake. "Dad?" she said. Then louder: "DAD!"
He shifted his gaze a little but didn't respond.
Now Anna was really getting scared. Nothing like this had ever happened before, and she couldn't understand what was going on. She reached up and gave her father a hard slap. Now he seemed to be coming around. He focused his eyes on her for a long moment and said " Anna I don't feel very well." Anna felt a flood of relief. She took her father's arm and started to turn him around, saying "Let's go back upstairs." As he turned, however, she noticed the numbers on the keypad he had been pressing. The top row had strange characters. One looked like a star drawn inside a spiral. Another was like a bird's wings, but split in half. The odd thing was that she thought she recognized that strange star symbol. She had seen it before.

Anna led her father back up to his room slowly. When they finally got there, he climbed back in bed, sweating and visibly sick. She watched him for a while. Finally she remembered the strange symbol on the keypad by the door in the basement. She got up and went back to her room. She looked through her pack until she found a little leather wallet that she had always kept in her dresser at home. In the wallet was a small silver chain. It was her baby anklet, and the only thing she had from her mother.
Her mother had made it herself when Anna was an infant. At one end there was a bit of plastic filed into a saw-toothed edge. Then there was a whole series of beads. Some of them had symbols scratched into them by hand. Some were filed down small and flat. Others were hanging off on side chains here and there. Now Anna looked at the chain again. The second bead from one end had exactly the same symbol as one of the keys on the keypad downstairs. Suddenly Anna had an idea. She wanted to compare her bracelet to the keypad in the basement. But first she went back to check on her father. When she got back to his room he was asleep. Suddenly she saw something through his bedroom window. There was a faint light in the backyard. Anna wondered what to do. Should she check out the keypad, or investigate the light in the backyard?

Fig. 2. Screenshot a RAFT story episode

The first student in each pair was given the "interactive" version of the tutor, and was allowed to make plot choices that generated a new story. The second student in the pair was then given that same story to read as a "static" text, without plot choices. This allowed textual content to be kept constant between conditions, eliminating the possibility that one group might read more simply because they got more interesting stories.

Students and their parents signed a consent form upon arrival. Students then took a reading motivation questionnaire which was based on the "intrinsic value" section of Pintrich and DeGroot's "Motivated Strategies for Learning" (MSLQ) [25]. Questions were read aloud to the student to remove any impact from poor reading comprehension. Following the motivational pre-test, the students put on a head-set with microphone, and were asked to read a short example story to familiarize them with the system. In both conditions, the system displayed a screen of text and read it aloud using the FreeTTS [26] text to speech system. After listening, the student read it back twice aloud, while the system recorded audio. After reading each passage, students in the "interactive" condition were allowed to make a plot choice which generated the next passage. Static subjects were simply given the next passage. Students were given about 25 minutes for reading in the tutor. The experimenter would then say that the "required" reading period was over, but that there was still a post-test that had to be taken in ten minutes. Students could just sit and rest for those ten minutes, or, if they wished, they could continue to read in the tutor. In both conditions if they chose to continue reading, they still had to read aloud, but didn't have to repeat the text twice. The system then recorded how many "voluntary" pages the students read during the 10 minute period.

Table 1. Reading Results by Condition

Interactive				**Static**			
		Pg. Read				**Pg. Read**	
Age	**Gend**	**Req.**	**Vol.**	**Age**	**Gend**	**Req.**	**Vol.**
11	F	8	5	9	M	12	0
11	M	12	6	11	M	10	0
9	F	5	1	11	M	2	1
14	F	8	1	9	F	7	0
13	M	9	3	13	M	6	3
11	M	6	4	**NA**			

Pre-tests and the warm-up reading took about 25 minutes. Time spent in "required" reading averaged 24.7 minutes in the interactive condition, and 26.3 minutes in the static condition. These numbers include both time spent reading and time spent making plot choices. The "voluntary" reading (or rest) period was 10 minutes. The system was capable of generating around seventy episodes of text, with most episodes ending in a binary plot choice. Our actual users, however, saw very little of this potential story. During the required reading phase of the experiment, students in the interactive condition saw 8 episodes on average, while students in the static condition averaged about 7.4 episodes.

To eliminate the possibility that static students might run out of text if they read more than had been generated by their interactive partner, several additional episodes were automatically generated and added to the end of each of the dynamically generated stories, before presenting them to the static students. As can be seen by comparing the total pages read ("Req." plus "Vol.") between conditions in Table 1, however, this precaution was unnecessary. None of the static students read past the text generated by their dynamic partners.

4 Results

In this section we first report several post-hoc tests used to determine if the pseudo-randomization procedure described in Section 3 resulted in unbalanced conditions. First, t-tests found no significant difference between conditions in pre-test reading motivation (p = .45). Reliability of the motivation instrument was measured by Cronbach's Alpha [27] and found to be a low, but not unacceptable .55. In addition, no significant differences were found in age (p =.41) or in reading fluency (p = .89). Repeating these tests with the non-parametric u-test gives exactly the same pattern of results, with no significant differences between conditions.

External scheduling considerations limited recruitment to 11 students, which prevented our planned paired analysis. We therefore report a post-hoc analysis using the Anova, which provides marginally significant results.

Table 1 shows the number of pages read for each of our 11 subjects. Subjects shown on the left of Table 1 were in the "Interactive" condition, while subjects on the right were in the "static" condition. Two subjects paired by reading ability are shown in the same row. The number of "Required" (Req.) and "Voluntary" (Vol.) pages read are shown under the "Pg. Read" headers.

We used Anova to test for the effect of interactive text on voluntary reading. In an Anova explaining the number of voluntary pages read by experimental condition ("interactive" or "static"), condition turned out to be a significant factor, F(1,9)=5.6 p = .042. This suggests that text type made a significant impact on voluntary reading. From Table 1 we see that voluntary reading was higher with interactive text: students in the interactive text condition read an average of 3.33 voluntary pages, while students in the static text condition read an average of .8 voluntary pages.

5 Discussion and Future Work

This paper has presented preliminary evidence supporting one of the fundamental assumptions underlying the RAFT project, namely that dynamically generated interactive stories will improve motivation to read even in this population of impaired readers. These results suggest that a system with a sufficiently large generation repertoire could lure these children into reading significantly more, and so increase their exposure to connected text. As noted by Szilas and Rety [28], a key to producing such a volume of interactive text will be "combinatorics." We plan to proceed by refactoring our current rule based system to manipulate more elementary narrative structures (e.g. [28]), and to more explicitly follow the functional divisions described by Spierling et al. [20].

After extending the story generator, we plan to revisit this motivational study with more subjects, formal randomization, and motivational questions administered by a second computer, to avoid Hawthorne effects.[1]

Acknowledgments. The first author was supported as a National Library of Medicine (NLM) fellow in the Department of Biomedical Informatics by the following NLM training grant to the University of Pittsburgh: 5 T15 LM007059. We gratefully thank Sheila Carr for helping to pair subjects, and for handling our (unusually difficult) scheduling duties. We also thank Liz Legowski for welcome statistical help, and our anonymous reviewers for many helpful suggestions.

[1] We thank an anonymous reviewer for this suggestion.

References

1. Shaywitz, S.E.: Overcoming Dyslexia: a new and complete science-based program for reading problems at any level. Vintage Books, New York (2003)
2. International Dyslexia Association: IDA Fact sheets: dyslexia basics (2011), http://www.interdys.org/FactSheets.htm
3. National Institute of Child Health and Human Development: What are learning disabilities (2011), http://www.nichd.nih.gov/health/topics/learning_disabilities.cfm
4. Cunningham, A., Stanovich, K.: What reading does for the mind. Journal of Direct Instruction 1, 137–149 (1998)
5. Ivey, G., Broaddus, K.: Just plain reading: A survey of what makes students want to read in middle school classrooms. Reading Research Quarterly 36(4), 350–377 (2001)
6. Pitcher, S., Albright, L., DeLaney, C., Walker, N., Seunarinesingh, K., Mogge, S., Headley, K., Ridgeway, V., Peck, S., Hunt, R., Dunston, P.: Assessing adolescents' motivation to read. Journal of Adolescent & Adult Literacy 50(5), 378–396 (2007)
7. Patall, E., Cooper, H., Robinson, J.: The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. Psychological Bulletin 134(2), 270–300 (2008)
8. Johnson, L., Vilhjalmsson, H., Marsella, S.: Serious games for language learning: How much game how much AI? In: Proceeding of the 2005 Conference on Artificial Intelligence in Education (AIED): Supporting Learning through Intelligent and Socially Informed Technology, pp. 306–313. IOS Press, Amsterdam (2005)
9. Rowe, J., Mott, B., McQuiggan, S., Robison, J., Lee, S., Lester, J.: Crystal island: A narrative-centered learning environment for eighth grade microbiology. In: Proceedings of the AIED 2009 Workshop on Intelligent Educational Games (2009)
10. Riedl, M.O., Stern, A.: Believable Agents and Intelligent Story Adaptation for Interactive Storytelling. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 1–12. Springer, Heidelberg (2006)
11. Spierling, U., Grasbon, D., Braun, N., Iurgel, I.: Setting the scene: playing digital director in interactive storytelling and creation. Computers & Graphics 26, 31–44 (2002)
12. Taraum, P., Figa, E.: Knowledge-based conversational agents and virtual storytelling. In: Haddad, H., Omicini, A., Wainwright, R.L., Liebrock, L.M. (eds.): Proceedings of the, ACM Symposium on Applied Computing (SAC 2004), pp. 39–44 (2004)
13. Mateas, M., Stern, A.: Towards integrating plot and character for interactive drama. In: Working notes of the Social Intelligent Agents: The Human in the Loop Symposium. AAAI Fall Symposium Series, pp. 113–118. AAAI Press, Menlo Park (2000)
14. Szilas, N.: IDtension: a narrative engine for interactive drama. In: Proceedings of the Technologies for Interactive Digital Storytelling and Entertainment (TIDSE) Conference, vol. 3, pp. 187–203 (2003)
15. Meehan, J.R.: Tale-spin, an interactive program that writes stories. In: Proceedings of the Fifth International Joint Conference on Artificial Intelligence, pp. 91–98 (1977)
16. Montgomery, R.A.: The Abominable Snowman. Chooseco, LLC, Waitsfield (1982)
17. Mateas, M., Stern, A.: Integrating plot, character and natural language processing in the interactive drama Façade. In: Proceeding of the Technologies for Interactive Digital Storytelling and Entertainment (2003)

18. Si, M., Marsella, S., Pynadath, D.: Thespian: using multi-agent fitting to craft interactive drama. In: Proceedings of the fourth International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2005, pp. 21–28 (2005)
19. Cavazza, M., Charles, F., Mead, S.: Emergent situations in interactive storytelling. In: Proceedings of the, ACM Symposium on Applied Computing, SAC 2002, pp. 1080–1085 (2002)
20. Spierling, U.: Interactive digital storytelling: Towards a hybrid conceptual approach. In: de Suzanne, C., Jennifer, J. (eds.) Changing Views: Worlds in Play: Proceedings of the 2005 Digital Games Research Association Conference, p. 11. University of Vancouver, Vancouver (June 2005)
21. Giarratano, J., Riley, G.: Expert Systems: Principles and Programming. PWS Publishing Co., Boston (1994)
22. Ward, A., Crowley, R.: Story assembly in the R^{2}aft dyslexia tutor. In: Proceedings 6th Workshop on Innovative Use of NLP for Building Educational Applications (ACL-HLT Workshop), pp. 130–135 (June 2011)
23. Bali, M.: Drools JBoss Rules 5.0 Developer's Guide. Packt Publishing Ltd., Birmingham (2009)
24. Marlow, A., Edwards, R.: Test review: Gray oral reading test, third edition (GORT-3). Journal of Psychoeducational Assessment 16, 90–94 (1998)
25. Pintrich, P., DeGroot, E.: Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology 82(1), 33–40 (1990)
26. Walker, W., Lamere, P., Kwok, P.: Freetts: A performance case study. Sun Microsystems Laboratories (2002)
27. Gliem, J., Gliem, R.: Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales. Midwest Research to Practice in Adult, Continuing and Community Education (2003)
28. Szilas, N., Rety, J.H.: Minimal structures for stories. In: Proceedings of the 1st ACM Workshop on Story Representation, Mechanism and Context (SRMC 2004), pp. 25–32 (October 2004)

Integrating Backchannel Prediction Models into Embodied Conversational Agents

Iwan de Kok and Dirk Heylen

Human Media Interaction, University of Twente
{i.a.dekok,heylen}@utwente.nl

Abstract. In this paper we will present our design for generating listening behavior for embodied conversational agents. It uses a corpus based prediction model to predict the timing of backchannels. The design of the system iterates on a previous design (Huang et al. [5]) on which we propose improvements in terms of robustness and personalization. For robustness we propose a variable threshold determined at run-time to regulate the amount of backchannels being produced by the system. For personalization we propose a character specification interface where the typical type of head nods to be displayed by the agent can be specified and ways to generate slight variations during runtime.

1 Introduction

One of the greatest challenges in developing embodied conversational agents is managing the flow of conversation. Succesful interaction between humans is achieved by complex coordination between verbal and nonverbal behaviors which together shape the information which is passed on from one interlocutor to the other. The behaviors that need to be displayed depend on the state the conversation is in.

With regards to turn-taking the conversational state of the agent will be modelled among two dimensions. The agent can *have* the turn or not and the agent can *want* the turn or not. These dimensions create four conversational states the agent can be in. Each of these states comes with their own type of actions and behaviors that are appropriate.

When the agent has and wants the turn, the agent will communicate what it has to share with its interlocutors. It will do this until it has shared all his information or until its turn is challenged by an interlocutor. At this time the agent will signal this through turn yielding behaviors. When it has lost the turn, the agent will need to display appropriate listening behavior to signal attendance to and understanding of its interlocutor. As soon as it wants the turn back it will need to display turn claiming behavior.

The design proposed in this paper is for the behavior of an agent in the conversational state that it does not have and does not want the turn. In this state the goal of our agent is to keep the interlocutor motivated in speaking by signalling attendance, understanding and/or appraisal through backchannels.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 268–274, 2012.

Humans succeeding in doing so increase quality of the speaker's speech [10,1], understanding of the speaker's speech by the listener [10,1] and rapport between the interlocutors [4].

This listening behavior is typically a combination of reactive and deliberative behavior. In our study we focus on reactive behaviors. Humans do not consciously plan each listener response, but they occur naturally without much thought. Over the past years several reactive prediction models have been developed to determine the timing of these listener responses. At first the handcrafted rules approach was utilized [16], but nowadays the corpus based machine learning approach has proven to outperform these handcrafted rules [12]. However, most implemented agents and robots still use these handcrafted rules [11,9,2] to time their listening behavior, the exception being [5].

Huang et al. [5] use a Conditional Random Fields model to predict the timing of the backchannels which is learned from a corpus of speaker-listener interactions. The timing of the backchannels is determed by comparing the output of the model to a threshold. The threshold is determined in the development of the model and optimized to optimally reproduce the behavior as observed in the corpus. At the predicted timings they randomly place one of three typical head nods, which were found in their corpus.

The proposed system in this paper intends to improve on this system in terms of robustness and personalization. For robustness we target the way the threshold is determined and used in the system. Due to changing conditions in an interactive system which can influence the output of prediction models, such as audio quality, recognition results of features or different speaking styles of users, a fixed threshold can lead to variable results. For some users the model will predict many backchannels, while for others it will predict hardly any backchannels. We propose a variable threshold determined at run-time to regulate the backchannel rate.

For personalization the proposed system allows for more types of head nods to be produced. It offers an interface to specify a character's typical head nods to define and personalize the created characters. Furthermore, these head nods are used as blue prints on which variations are generated. Depending on the certainty of the model that the prediction is correct, the system will generate a determined head nod on a high certainty prediction and a more shallow backchannel on a low certainty prediction.

In the remainder of the paper a general overview of the proposed system is given and the different components that are introduced there are discussed in more detail.

2 General Overview

Our design (see Figure 1) consist of two main components, the prediction module and the listener response generator. The prediction module monitors the interlocutor through the multimodal input channels of the system. Based on these observations, the model - which is trained on human-human interactions - produces

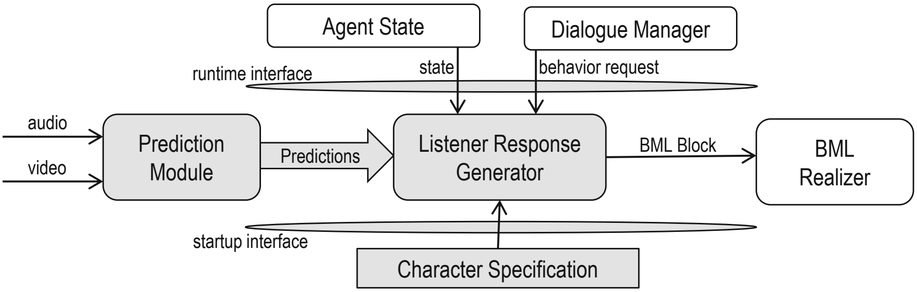


Fig. 1. Overview of the system architecture for generating listener responses. The gray parts and thick arrows are part of the system, while the thin arrows and white parts are part of the outside world, or part of the complete agent architecture.

a prediction value at each time frame, which indicates the appropriateness of a listener response at this time (see Section 4 for more details).

The listener response generator interprets these prediction values and decides at what times and with which form the listener responses are given. It generates BML blocks [15] which are passed on to a BML realizer (see Section 5).

There are two interfaces that can be used to influence the behavior of the system, one at startup and one during runtime. Through the interface at the startup a character's typical behavior can be specified (see Section 3.1). At runtime the agent's state is monitored which influences the behavior and the dialogue manager can request specific behavior from the system (see Section 3.2).

3 Interfacing with the Module

There are two ways to influence the behavior of the module. At the start up of the module the character specification is loaded. At runtime the module monitors the relevant states of the agent and it allows requests for specific behaviors.

3.1 Startup Interface

In our previous work we have seen that listeners differ greatly in amount of responses given, even when interacting in the same context as others [6]. Thus, the rate at which a listener responds is not only determined by the amount of opportunities given by the speaker and the understanding of the listener at these opportunities, but also by the individual characteristics of the listener.

Furthermore, it has been observed that listeners differ in the form they usually use as their listener response. Some listeners frequently use a vocal component in their responses, while others remain silent. Some listeners usually start their head nods upwards, others downwards and also the speed, amplitude and amount of nods per listener response differs between listeners.

To be able to easily generate listeners differing in these aspects with the same listening behavior module, a character specification is loaded at the startup of the module. This character specification constitutes the blueprint for the generated behavior of the listeners. In this specification the frequency of listener responses and the form of typical listener responses for this person are specified.

3.2 Runtime Interface

To allow the agent some control over the generated listening behavior by the listening behavior module a interface at runtime is available.

Through this interface the module monitors the relevant states of the agent. For the first implementation these states are limited to the conversational state (does the agent still neither have nor want the turn?) and the variable understanding. The conversational state determines whether the module produces behavior or not. Understanding has values between 0 and 1, where 0 is no understanding and 1 is full understanding. This state influences the form of the listener responses. Full understanding will produce determined listener responses and no understanding will display misunderstanding behavior.

The second way the module allows interaction is through the request channel. Through this channel requests can be made by the dialogue manager of the agent for a specific listener response. The module will generate this listener response at the first opportunity detected by the prediction module. The request has one of two priority labels; *high* or *low*. In the case of high priority the threshold the prediction value needs to exceed is lowered, while no manipulation of the threshold is performed in the case of low priority. This functionality will be implemented to allow more control from the agent over the generated behavior.

4 Listener Response Prediction Module

The listener response prediction model is responsible for predicting the timing of the generated listener responses. Contrary to the rule based prediction models used so far in embodied conversational agents, our listener response prediction model is a model trained on human-human conversations. Over the years many such models using various machine learning techniques, such as HMM [3], CRF [12,7,13] and SVM [8], have been trained and evaluated and proven to be more accurate than their handcrafted peers. Based on the input features describing the context these models make a prediction on how likely a listener response is at each moment in time. Input features typically used are eye gaze, prosody and lexical features and are derived from audio and video input. All these features can be detected and/or interpreted in real-time and incrementally.

For our implementation we will select an SVM model learned on the MultiLis corpus [8]. The implementation will be such that it is model independent and can easily be replaced by another model, as long as the output is a continuous stream of prediction values, similar to the output depicted in Figure 2.

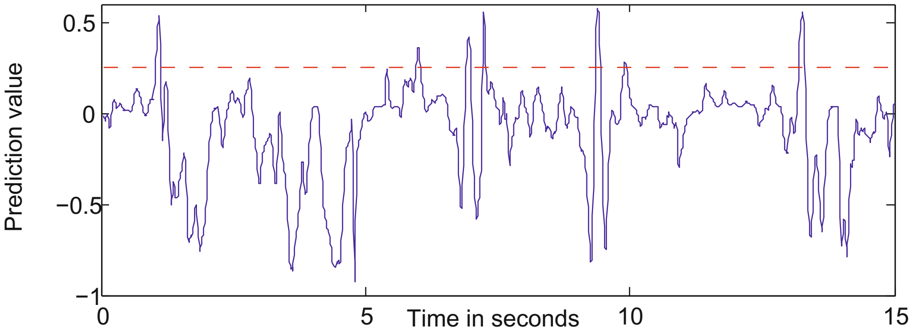


Fig. 2. The output of the prediction module; a prediction value curve

5 Listener Response Generator

The listener response generator is responsible for interpreting the output of the listener response prediction model and generating the appropriate response. The design will enable the module to generate varied, but personalizable listening behavior.

5.1 Timing of Listener Responses

The output of 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 as depicted in Figure 2. From this curve the timing of listener responses can be derived.

At certain times the prediction value curve peaks after a fast increase in prediction value. When the top of this peak exceeds a certain threshold (e.g. the red interrupted line) a listener response is predicted by the model. Selection of the threshold the peaks need to exceed is influenced by the response rate set in the character specification.

To create a more robust system with regards to the amount of backchannels that are generated, we will not select a fixed threshold like Huang et al. [5], but the threshold will be subject to change during an interaction. To regulate the response rate the threshold will slowly decrease as time since the last listener response goes by and increase as soon as a listener response is given by the agent. This will ensure that the system will be able to generate a similar amount of responses under changing conditions and for different speaking styles of users. Furthermore, this will ensure that long periods with no responses are less likely and that listener responses are not performed shortly after each other. This has been found to be perceived as erratic behavior for an agent [14], even though humans do this occasionally. The exact rates and type (linearly/exponentionally/...) of increase/decrease will be determined in the prototyping phase of development.

5.2 Form Selection

When a peak is detected exceeding the threshold, an appropriate behavior needs to be selected. If a request is made for a specific listener response, that listener response is generated (see Section 3.2). If no request is made for a specific listener response, the typical response for the character as specified by its specification (see Section 3.1) is generated.

Even though each individual has a preferred way of giving a listener response, these are not exactly the same. In fact, they vary their typical response slightly each time and at certain times they deviate from their typical response to give a really determined response.

We use the height of the peak in the prediction value curve to generate these variations in the form of the listener response. The higher the peak, the more determined the generated listener response is. This is achieved by increasing the amplitude and/or speed of the movement and/or the intensity of the facial expression. Since wrongly timed vocal listener responses are more often perceived as inappropriate than head nods [14], vocalizations are only added when the peak is high.

6 Conclusion and Future Work

In this paper we have presented our design for generating listening behavior for embodied conversational agents. The systems uses a corpus based prediction model to predict the timing of backchannels. The design of the system iterates on a previous design (Huang et al. [5]) on which we propose improvements in terms of robustness and personalization. For robustness we proposed a variable threshold determined at run-time to regulate the amount of backchannels being produced by the system. For personalization we propose a character specification interface where the typical type of head nods to be displayed by the agent can be specified and ways to generate slight variations during runtime.

For future work we intend to do a subjective evaluation of the system in which we evaluate the different components, by switching certain components and mechanics on or off.

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. Bevacqua, E., McRorie, M., Pammi, S., Pelachaud, C., Schröder, M., Sneddon, I., de Sevin, E.: SAL multimodal generation component with customised SAL characters and visual mimicking behaviour. Tech. rep., SEMAINE Project (2009)
3. Fujie, S., Fukushima, K., Kobayashi, T.: A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In: Proc. Int. Conference on Autonomous Robots and Agents, pp. 379–384 (2004)

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.: Virtual Rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
6. 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 2010. LNCS, vol. 6456, pp. 362–375. Springer, Heidelberg (2011)
7. 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)
8. de Kok, I., Poppe, R., Heylen, D.: Iterative Perceptual Learning for Social Behavior Synthesis. Tech. rep., Centre for Telematics and Information Technology University of Twente (2012)
9. Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling Embodied Feedback with Virtual Humans. In: Wachsmuth, I., Knoblich, G. (eds.) ZiF Research Group International Workshop. LNCS (LNAI), vol. 4930, pp. 18–37. Springer, Heidelberg (2008)
10. 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)
11. 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)
12. 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)
13. Ozkan, D., Sagae, K., Morency, L.: Latent Mixture of Discriminative Experts for Multimodal Prediction Modeling. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 860–868. Association for Computational Linguistics (2010)
14. Poppe, R., Truong, K.P., Heylen, D.: Backchannels: Quantity, Type and Timing Matters. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 228–239. Springer, Heidelberg (2011)
15. Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: 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)
16. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207 (2000)

Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation

David Traum, David DeVault, Jina Lee, Zhiyang Wang, and Stacy Marsella

Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Playa Vista, CA 90094, USA

Abstract. In order to provide comprehensive listening behavior, virtual humans engaged in dialogue need to incrementally listen, interpret, understand, and react to what someone is saying, in real time, as they are saying it. In this paper, we describe an implemented system for engaging in multiparty dialogue, including incremental understanding and a range of feedback. We present an FML message extension for feedback in multipary dialogue that can be connected to a feedback realizer. We also describe how the important aspects of that message are calculated by different modules involved in partial input processing as a speaker is talking in a multiparty dialogue.

1 Introduction

In order to be human-like in their behavior, intelligent conversational agents need to be able to produce a range of feedback to a speaker during a conversation. Human feedback behavior has a number of features, which translate to a set of requirements for satisfactory virtual agent feedback. Some of these requirements are:

- Human feedback is provided in *real-time*, as the speaker is articulating (or having trouble articulating) her utterance. This means that the feedback mechanism can not wait until after the speaker has finished to calculate the feedback.
- Human feedback is also often *specific* [4], so the feedback mechanism requires interpretation and attempted understanding of what the speaker is saying.
- Taken together, these requirements lead to a third one, that understanding be *incremental* - operating on parts of the evolving utterance and computed in real-time before the utterance has been completed.
- Human feedback is also *expressive* [1], indicating aspects of the current mental state of the feedback giver, including beliefs, goals, emotions, and how the developing utterance is related to these. This means the feedback mechanism must have access to the cognitive aspects and be able to do pragmatic reasoning, including reference resolution, to relate the utterance meaning to the agent's mental state.
- Human feedback is sometimes *evocative* [1], trying to create an impression on or response behavior from the observers of the feedback. This includes intended effects on the main speaker, to regulate the content, timing, and amount of detail of what she is saying, as well as intended effects on other observers, such as adoption of beliefs about the feedback giver, or whether they should take a turn next. Evocative feedback means that the feedback mechanism must have access to the communicative goals, plans and intentions of the agents.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 275–288, 2012.

A model of feedback generation that contains each of these features was presented in [35]. In this paper we describe an FML message specification that supports this behavior generation model, as well as an implemented agent dialogue model that can provide the aspects of semantic and pragmatic understanding for specific expressive and evocative feedback in realtime, This model is similar in many respects to that proposed by [24], however that model worked on typed dyadic dialogue rather than spoken input for multiparty conversation, and focused on expressive feedback, while the model below also includes evocative feedback and participation status of participants.

2 Architecture and Message Specification

Our architecture requires the following components to produce incremental feedback:

- A speech recognizer that can produce incremental, word-by-word results, ideally with confidence scores.
- A natural language understanding component (NLU) that produces semantic representations and predictions of final meaning when given a speech recognition output.
- A meta-NLU component that computes confidence estimates given (partial) speech recognition and NLU outputs.
- A vision component that can recognize speaker behaviors such as gaze direction.
- A domain reasoning component that can model beliefs, tasks, plans, and attitudes toward particular topics.
- A dialogue manager that can compute pragmatic effects of communication as recognized by the above input components and update state and calculate communicative intentions.
- A feedback generator that can produce communicative behaviors given the function specifications from the dialogue manager.

In practice, some of these components can be combined in a single software module. Our architecture combines the NLU and meta-NLU components into a single module, and the dialogue manager and domain reasoning is another single module.

In order to pass the needed information from dialogue manager to the feedback generator, we created the XML backchannel feedback message specification in Figure 1. This message type is meant to be part of the SAIBA framework [23], with most of it being FML content [18]. The message also contains aspects that have been developed by earlier processing components. There is one `participant` element for each participant in the conversation. Participant roles are discussed in Section 8. The `conversation-goal` element contains goals related to maintaining and changing participant status. They are discussed in Section 9. The `dialogue-act` element represents the feedback itself – it can be given either as a backchannel (type=listening) or verbally within a turn (type=speaking), although we have not fully implemented the speaking type as of this writing. The `feedback` element contains information about the utterance that is being spoken, and what the agent thinks about it. Attributes of this element and the `partial-text` element are derived by the speech recognizer, and described in Section 4. The `partial-sem` element contains information from the NLU module, and is described in section 5. This information is augmented by contextual information

```
<act>
 <participant id="[character]" role="[role]"/> (minimum 2)
 <fml>
   <conversation-goal>
     <participation-goal goal="[boolean]"/>
     <comprehension-goal goal="[boolean]"/></conversation-goal>
   <dialog-act type="[listening/speaking]">
     <feedback agent="[character]" speaker="[character]"
           utterance="[id]" progress="[integer]" complete="[boolean]">
        <partial-text>[string]</partial-text>
        <partial-sem confidence="[real]">
           <indicators Correct="[boolean]" High="[boolean]"
                Incorrect="[boolean]" Low="[boolean]" MAXF="[boolean]"
                PF1="[boolean]" PF2="[boolean]" PF3="[boolean]"
                WillBeCorrect="[boolean]" WillBeHigh="[boolean]"
                WillBeIncorrect="[boolean]" WillBeLow="[boolean]"/>
           <predicted_nlu><object name="[id]">
                        ... </predicted_nlu>
           <explicit_subframe><object name="[id]">
                        ...  </explicit_subframe></partial-sem>
        <attitude type="[like/dislike]" target="[id]" stance="[leaked/intended]"
                   intensity="real"/>
        <affect type="[emotion]" target="[id]" stance="[leaked/intended]"
                   intensity="real"/>
      </feedback></dialog-act></fml></act>
```

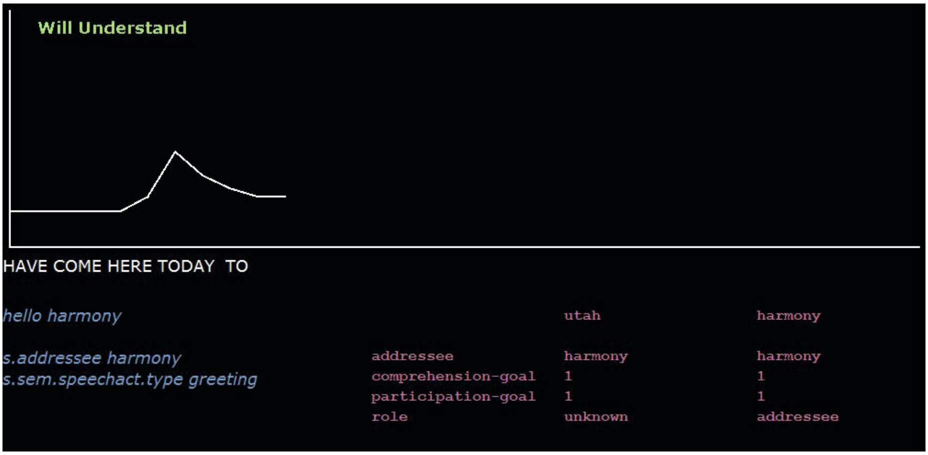
Fig. 1. Feedback Behavior Generation Message Specification

provided by the dialogue manager, as described in Section 6. Finally, the `attitude` and `affect` elements come from the domain model expected utility calculations and emotions, once the dialogue manager has identified the relevant concepts that are being spoken about. The domain model is discussed in Section 7. Finally, in Section 10, we briefly review the feedback behavior generation component that takes as input messages of the form of Figure 1 (more details are provided in [35]).

The specification and components are domain independent, and have been tested in a few different domains. However, to provide more concreteness in examples we present one domain, *SASO4*, described in the next section. Figures 2 and 3, show a visualization of some of the information from the Feedback message, in a graphical form. These figures show a couple of snapshots 2.0 and 4.6 seconds in the progress of a single 7.4 second utterance in the SASO4 domain.

3 Example: The SASO4 Domain

As our development testbed, we situated this work in the SASO4 domain, which extends the scenario described by [29]: *An American Old West town has been freed from a dangerous outlaw, defeated by a U.S. Ranger with the help of Utah, the local bartender. The Ranger and his Deputy must now leave town to pursue their mission elsewhere. But before leaving, they need to recruit a town sheriff, so they offer the job to Utah. He will need resources – e.g., money to buy guns and to hire men – guaranteed before considering the offer. As owner of the saloon, Harmony is an influential woman in town. She will be present in the discussions, pushing forward her own agenda of demands, part of which she cannot discuss in front of Utah and must be dealt with in private by*



HAVE COME HERE TODAY TO

hello harmony

s.addressee harmony
s.sem.speechact.type greeting

	utah	harmony
addressee	harmony	harmony
comprehension-goal	1	1
participation-goal	1	1
role	unknown	addressee

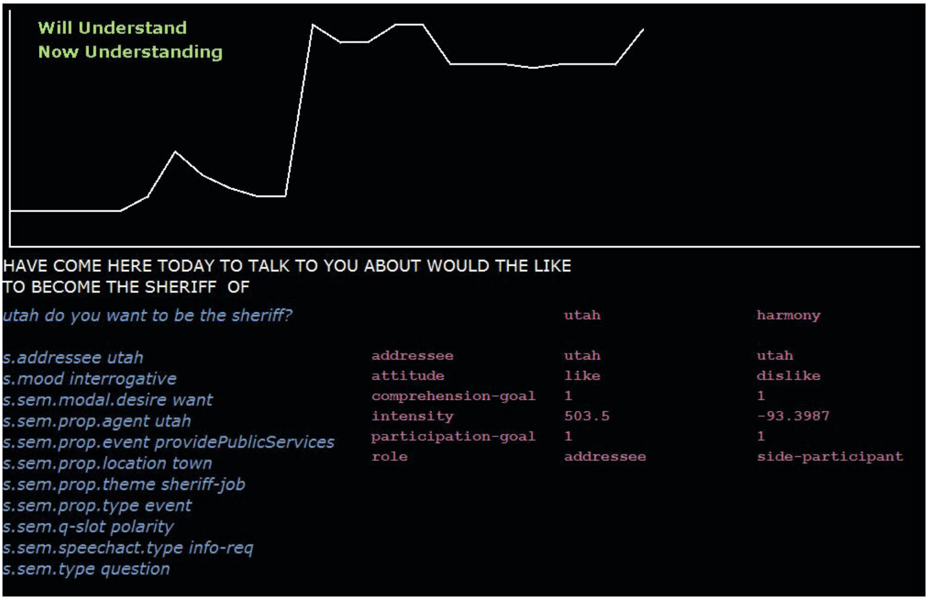
Fig. 2. Visualization of Incremental Speech Processing after 2 seconds

one of the officers. The Ranger and the Deputy have very limited resources, so they must negotiate to reach an agreement by committing as little as possible.

This scenario has many opportunities for feedback, both as the scenario progresses, and within the interpretation of a single utterance. It includes four characters, two played by humans, and two by virtual agents. The scenario starts with no conversation, but the humans could start conversations with one or both agents. It contains situations in which the agent Harmony desires to leave the conversation, and an opportunity for re-entry if she leaves. The agents also have shifting points of view about some of the things discussed as the conversation progresses, e.g. whether Utah will be the Sheriff.

4 Speech Recognition

Our automatic speech recognition (ASR) module is currently PocketSphinx [19]. The ASR is configured with a statistical language model trained on the transcripts in a corpus of user utterances and paraphrases. In the SASO4 scenario, we currently use approximately 1,500 transcripts to train the language model. To enable incremental understanding and feedback based on partial ASR results, after each 200 milliseconds of additional speech from an ongoing user utterance is captured, it is provided to the ASR. The ASR module sets the *utterance, speaker, progress*, and *complete* attributes of the feedback element in the partial message in Figure 1. The *utterance* attribute is a unique id for this session, the *progress* attribute contains the ordinal count of partial interpretations of this utterance. The *complete* attribute signals whether the speaker has stopped speaking. The partial ASR result appears in the `partial-text` element in the feedback message. The partial-text is shown in white in the visualization in Figures 2 and 3.



```
HAVE COME HERE TODAY TO TALK TO YOU ABOUT WOULD THE LIKE
TO BECOME THE SHERIFF  OF
utah do you want to be the sheriff?                                 utah          harmony

s.addressee utah                      addressee                     utah          utah
s.mood interrogative                  attitude                      like          dislike
s.sem.modal.desire want               comprehension-goal            1             1
s.sem.prop.agent utah                 intensity                     503.5         -93.3987
s.sem.prop.event providePublicServices participation-goal           1             1
s.sem.prop.location town              role                          addressee     side-participant
s.sem.prop.theme sheriff-job
s.sem.prop.type event
s.sem.q-slot polarity
s.sem.speechact.type info-req
s.sem.type question
```

Fig. 3. Visualization of Incremental Speech Processing after 4.6 seconds.

5 Semantic Interpretation

We adopt a detailed framework for incremental understanding and confidence estimation that has been developed in [30,31,8,7]. The key components for listener feedback behavior are the semantic frames and confidence indicators that are produced for each partial ASR result. This incremental understanding framework captures utterance meanings using a frame representation, where attributes and values represent semantic information that is linked to a domain-specific ontology and task model [16].

As a user utterance progresses, the incremental NLU component produces two semantic frames. The first frame is a prediction of the meaning of the *complete user utterance* (which may not have been fully uttered yet). This prediction is made using a statistically trained maximum entropy classifier [8]. The `<predicted_nlu>` element in the feedback message specification in Figure 1 provides this predicted frame. Examples of such predictions are also shown in blue in Figures 2 and 3, below the ASR partial result, along with a gloss of the meaning of the frame. For Figure 2 the prediction is that this utterance will be a greeting to harmony. In Figure 3, the prediction has changed to a more detailed frame in which the user is asking Utah if he wants to become the sheriff.

The second type of frame produced by the incremental NLU components is an *explicit subframe* that attempts to capture the explicit meaning of only what the user has said so far, without predicting the complete meaning of the user's full utterance. The identification of this subframe can be performed using related statistical classification techniques [17], and the resulting subframe is given in the `<explicit_subframe>` element in the feedback message specification. (This component can optionally be shown in the visualization, but it is not shown in Figures 2 and 3.)

A third component in this incremental processing framework is a set of boolean-valued confidence indicators that can be used to assess, in intuitive terms, the reliability of the predicted frame for the user utterance [7]. The indicators encompass a range of potentially valuable information about how well an utterance is being understood so far, and how much that understanding may improve as the user continues speaking.

All the indicators are ultimately defined in relation to an F-score metric which can generally be used to assess NLU performance. The F-score calculation looks at precision and recall of the attribute-value pairs (or frame elements) that compose the predicted and correct (hand-annotated) frames for each partial ASR result. Precision represents the portion of frame elements in the predicted frame that were correct, and recall represents the portion of frame elements in the gold-standard annotations that were predicted.

Table 1. Metrics for incremental speech understanding

Metric	Definition	Metric	Definition	Metric	Definition
High$_t$:	$F_t \geq \frac{1}{2}$	WillBeHigh$_t$:	$F_L \geq \frac{1}{2}$	PF1$_t$:	Correct$_t \vee ($Incorrect$_t \wedge$ WillBeCorrect$_t)$
Correct$_t$:	$F_t = 1$	WillBeCorrect$_t$:	$F_L = 1$	PF2$_t$:	High$_t \vee ($Low$_\mathrm{t} \wedge$ WillBeHigh$_t)$
Incorrect$_t$:	$F_t < 1$	WillBeIncorrect$_t$:	$F_L < 1$	PF3$_t$:	High$_t \vee ($Low$_t \wedge \neg$MAXF$_t)$
Low$_t$:	$F_t < \frac{1}{2}$	WillBeLow$_t$:	$F_L < \frac{1}{2}$		
		MAXF$_t$:	$F_t \geq F_L$		

Currently we have been using a set of indicators that are defined in Table 1. In this table, F_t is the F-score of the predicted frame at time $t \in 1...L$, for an utterance that contains L 200 millisecond chunks of audio. F_L is the F-score of the final predicted frame for the complete user utterance. We also are using a set of more complex indicators that may indicate appropriate moments for virtual humans to provide positive feedback. These are defined in the right column of Table 1.

All these indicators are included in the `<indicators>` element in the feedback message specification in Figure 1. In Figure 3, the indicators High ("Now Understanding") and WillBeHigh ("Will Understand") are shown in green at the top left (both true at this point). In Figure 2, only WillBeHigh is true, as the system has low confidence about the current guess. The progression of expected F-score (given in the `confidence` attribute of the `partial-sem` element in the message spec in Figure 1) is shown by the white line at the top of the figures - low for Figure 2 and high for Figure 3.

6 Contextual Pragmatics

The semantic representation provided by the NLU component represents the context-free interpretation of the meaning of the utterance. The next step in processing is interpreting this within the current context to provide a pragmatic meaning, along the lines in [32]. For every partial utterance returned by the NLU component, the following are computed independently by each agent:

- updates on the participant structure
- a set of zero or more dialogue acts that have been performed by the utterance

- resolution of named entities to concepts
- resolution of action and state descriptions in the semantic interpretation to relevant states and tasks in the agents's task model (see Section 7).

We describe each of these briefly (except for addressee and participant structure, described in Section 8). First, primitive concepts that are part of the semantic representation are resolved. Primitive concepts include people (e.g. the participants in the dialogue), objects, locations, action types, attributes, and values. For named concepts, there is a simple look-up table (in most cases, the identity mapping), which allows each agent to have a different internal representation from other agents (if desired), and allows multiple semantic terms to refer to the same internal, domain-specific concept. Slightly more complex is the resolution of typed referring expressions, including indexicals ("I", "we", "you", "here", "there"), anaphors ("he", "she", "it", "this", "that"), and noun phrases that do not uniquely identify the concept for the domain (e.g. "the money"). In this case, the process for reference resolution involves looking up the features of the referring expression (e.g., animacy, gender, location, type) and then finding a list of "candidate" concepts that have these features. Then, if possible, disambiguation is performed by preferring those that have been mentioned most recently. If a single best candidate can not be found, this is motivation for an agent to perform a grounding move, giving feedback of sub-optimal understanding. This might take the form of a clarification request asking which of the possible candidates is meant, or a verification request about one of the candidates, or an open-ended question asking for the disambiguation (though the agent might wait if the agent thinks that they will understand later).

There is also a resolution of complex elements such as actions and states. The agents have a representation of a set of relevant states, whose valence are checked whenever new information comes in, including perceptual information, inference from other internal information, or verbal reports from trusted others. These states are represented as triples of *object, attribute, value*, where each of these are more primitive concepts. A given utterance might uniquely refer to an internal state, but might also not match any state or might match more than one (e.g. if one of the three elements is missing or is an under-constrained referring expression). Likewise, actions are represented, with a set of thematic roles, including *agent, theme, location, destination*, etc. A match is made from the consistency of the semantic frame with the task model frame to get a set of action candidates. Then, depending on the tense and modality, additional matches may be made with action instances in the plan or in the causal history of previous events. Recognized states and actions also provide an additional constraint on concept identification – one candidate concept is preferred over another if it leads to a possible match with a state or action and the other does not.

Finally, a set of dialogue acts are identified that are being performed by the speaker in producing the utterance. These include core speech acts, such as assertions, questions, offers, backward acts, such as answers, acceptances, grounding acts, and other acts that influence the information state of the dialogue. In the version of the semantic frames in the partial-sem attribute of Figure 1, the frames are augmented with reference information, though that is not provided in the visualization in Figures 2 and 3.

7 Domain Reasoner

Once the referred to objects and speech acts have been computed, it is possible to relate the speaker's projected sentiment toward the referenced object with the listener's feelings about the topic. The ability of our agents to interact with humans and other agents is based in their understanding of the goals of each party, the actions that can achieve or thwart those goals, and the commitments and preferences agents have towards competing courses of action. To provide this understanding, our agents use an explicit representation of an agent's current mental state concerning past, present, and future states and actions, their likelihood and desirability, and causal relationships between them. This representation is grounded in a planning representations that has been extended to incorporate representations of decision-theoretic reasoning (i.e, probabilities and utilities), representations to support reasoning about beliefs and intentions and a causal history that expresses the relations between past events to the agent's current beliefs, goals and intentions. We call this representation a causal interpretation.

The agent's valence reactions to its comprehension, including the attitude and affect elements depicted in Figure 1, also rely on this causal interpretation. Specifically, the valence reactions are based on a general computational framework for modeling emotion processes, EMA (Emotion and Adaptation) [14,27]. EMA is based on appraisal theories of emotion that argue that emotion arises from a process of interpreting a person's relationship with their environment; this interpretation can be characterized in terms of a set of criteria (variously called appraisal dimensions, appraisal variables or appraisal checks); and specific emotions are associated with certain configurations of these criteria. To represent the agent's relation to its environment, EMA relies on the agent's decision-theoretic plan based representation. The plan represents a snapshot of the agent's current view of the agent-environment relationship, including its beliefs, desires and intentions. This representation changes moment-to-moment in response to internal and external changes. EMA's appraisal of these changes uses fast feature detectors that map features of the plan into appraisal variables. Appraisals thus provide a continuously updated affective summary of its contents. This is particularly relevant to model the valenced reactions to the dynamically evolving comprehension of partial utterances. The appraisal process in EMA maintains a continuously updated set of appraisal values associated with each proposition in the causal interpretation. A partial list of the variables most relevant to the current discussion include:

Desirability: This characterizes the value of the proposition to the agent (e.g., does it causally advance or inhibit a state of utility for the agent). Desirability has both magnitude and sign – it can be positive or negative. This includes propositions that have intrinsic utility for the agent but also propositions that have extrinsic utility by virtue of causally impacting a proposition that has utility.

Likelihood: This is a measure of the likelihood of propositions.

Causal attribution: who deserves credit/blame.

Controllability: can the outcome be altered by actions under control of the agent.

Changeability: can the outcome be altered by some other causal agent.

Each appraised event is mapped into an emotion instance of some type, such as hope or fear, with some intensity, based on the pattern of appraisals. The intensity is calculated in the form of expected utility based on desirability and utility.

EMA also includes a computational model of coping integrated with the appraisal process. Coping determines, moment-to-moment, how the agent responds to the appraised significance of events. Within EMA, coping strategies are proposed to maintain desirable or overturn undesirable events. As opposed to the more reactive nature of appraisal, coping strategies can be seen as more deliberative attempts to enable or suppress the cognitive processes that operate on the causal interpretation.

With this background, we can characterize how the `affect` and `attitude` elements of the message spec in Figure 1 are calculated. The attitude's type attribute is based on the desirability of a referenced task action. The intensity attribute is derived from the calculation of that action's expected utility. The stance attribute distinguishes between expressive feedback from the appraised desirability (termed "leaked" in the attitude element specification), vs. evocative feedback meant to intentionally realize coping strategies, by conveying a specific affect, which may not be what is really felt (termed "intended"). This latter intentional expression of attitudes is not fully implemented yet in the incremental feedback. In Figure 2, there is no attitude shown, since there is no task model element referred to (yet) given the prediction of a greeting act. In Figure 3, we can see that Utah likes the idea of becoming Sheriff with intensity over 500 while Harmony dislikes the idea with an intensity of about -93.

The `affect` element is tied more directly to the results of the appraisal and coping process. In contrast to the `attitude` element, the `affect` element specifies an emotional category such as anger or fear. It can either express a felt (appraised) emotion or intentionally evoke a reaction by portraying an emotion that might not be felt. Although these appraisal and coping responses are implemented in the agent, the pathway of extracting the appraisals and coping responses based on the partial understanding has not yet been fully implemented.

8 Computing Participant Structure

Participant elements in Figure 1 describe the roles played by all scenario characters that are in contact [34]. Some scenario characters may be out of contact for part of the time, such as when they are in another room. Characters can be played by humans or other agents. The dialogue model tracks two types of roles relevant for participation state. First, there is the *conversational role*, which is either *active-participant* for someone who has recently taken an active part in the conversation, e.g. acting as a speaker or addressee of an utterance that is part of the conversation, or *overhearer* if playing a passive role.

The second type of role is the *utterance role*, which is how the character relates to the particular utterance. Utterance roles are *speaker, addressee, side-participant, overhearer,* and *eavesdropper*. The *speaker* utterance role is the speaker of the utterance that feedback is being given about. Our system currently is given this information for human users via the microphone that is used to pick up the speech or in agent messages used to indicate agent speech. Addressees are computed during message processing, following the algorithm in [33]. If an explicit name is used in a vocative, then the NLU will recognize the addressee. Otherwise, if the speaker is gazing at someone, then that character is assumed to be the addressee. Otherwise, contextual information is used, including the previous speaker, previous addressee, and other participant status. Active participants in the conversation who are neither speaker nor addressee are assigned the utterance role of *side-participant*. Overhearers in the conversation (who are not speakers or addressees of the current utterance) are assigned the utterance role of *overhearer*. Finally, observers of the utterance who do not have a role in the conversation are assigned the utterance role of *eavesdropper*. We can see some changes in participant status between Figures 2 and 3. In Figure 2, both characters think Harmony will be the addressee, because the NLU component thinks Harmony will be identified in the utterance. Utah is not sure of his role at this point. In Figure 3, both agents now think Utah is the addressee, because of prior context and lack of explicit signals. Harmony thinks she is a side-participant.

9 Evocative Feedback: Conversational Goals

As described in [35], there are two types of conversational goals considered, *comprehension goals* and *participation goals*. Both are linked to participant roles, and both have internal and evocative aspects. The internal aspects refer to the agent's actual goals: for comprehension goal, whether or not to comprehend the current utterance; for participation goal, whether or not to be an active participant in the conversation. The internal aspect influences the agent's cognition and action selection. For a positive comprehension goal, the agent will expend cognitive resources to listen to and understand the utterance. For a negative comprehension goal, the agent will focus attention on other matters, such as planning next actions or utterances, emotional reasoning, or task execution. For a positive participation goal, the agent will look for opportunities to further the conversation with active conversational behavior. A negative participation goal will lead the agent to disengage, perhaps moving further away and out of contact.

The evocative conversational goals are the intention to influence others beliefs and actions related to the agent's goals. Regardless of the true internal goals, agents may want to evoke in others a belief (and resulting behaviors stemming from such a belief) that they have either the same or different conversational and participation goals. In general, it is the evocative conversational goals that are passed to the feedback behavior generation component.

There are default goals that are norms for the different utterance roles, shown in Table 2. These defaults can be overridden, however, by more specific goals or coping strategies of the agent. For instance, if the agent is an overhearer or eavesdropper who wants to join the conversation more actively, or an active participant who wants to leave

Table 2. Normative Goals for utterance participant roles

Role	Comprehension	Participation
Speaker, Addressee, Side-Participant	Yes	Yes
Overhearer	No	No
Eavesdropper	Yes	No

the conversation (as Harmony does if Utah challenges her for disliking the plan to make him Sheriff), they may adopt a participation goal that is contrary to their current status. This may lead to a similar evocative goal, and behaviors indicating the desired new status. Another example is that overhearers and eavesdroppers who hear an action with a strong intensity will decide to join the conversation more actively as it turns to this subject, and adopt a positive participation goal. Likewise, one might want to maintain status as an addressee or side-participant, and keep participation goals, while something more urgent demands attention, thus a negative comprehension goal. In Figures 2 and 3, both participation and comprehension goals are 1 for both characters. However in the accompanying video, we can see that sometimes Harmony has a participation goal of 0.

10 Behavior Generation: A Review

Here we review aspects of feedback behavior generation, first reported in [35]. The generation of nonverbal listening behaviors is controlled by the NonVerbal Behavior Generator (NVBG, [25]) and specifically by an extension to the knowledge incorporated into NVBG. NVBG receives signals of the form of Figure 1 from the virtual human system's dialog module, as well as signals such as head nods and gaze of other agents from the perceptual processing compoents.

10.1 Behaviors

To inform the knowledge used in NVBG, we turned to existing literature that describes listening behaviors depending on a listener's roles and goal. For addressees, gaze and mutual gaze conveys the intent to participate and comprehend as well as continued attention [2]. Addressees also glance at other side-participants to seek social comparison [10] or avert gaze as a signal of cognitive overload when comprehending speech [2,13]. Various nodding behaviors are used to signal that the addressee is attending [28], comprehending [5,9] or reacting to the speaker [20] and thereby to signal participation and comprehension. Head tilts and frowns are used to signal confusion [5], and various facial expressions signal emotional reactions to the content of the speech.

Side-participants exhibit similar behaviors as addressees. However, they may be less committed to comprehend the current dialog. If side-participants do not care about understanding the speaker's utterance (i.e. comprehension goal is false) but the goal is to maintain the participation status, they use glances toward the speaker [2,15]. The glances here are not to further comprehend but rather to act as a ratified participant. Mimicking or mirroring the speaker's behavior [11,26] are also exhibited, in part to hold his/her current conversation role.

Eavesdroppers have the goal to understand the conversation but their status as anonymous eavesdroppers may be threatened if they openly signal their comprehension. Thus, to maintain that role, they should avoid mutual gaze and suppress, or restrain from showing, reactions to the conversation [10]. Furtive glances at the speaker are occasionally used for better comprehension, but gaze is quickly averted to avoid mutual gaze, to prevent providing visual feedback [3] and signs of attention to the speaker [2,3,22].

Overhearers are modeled as having neither goals for participation nor comprehension and have fewer concerns about the conversation. Gaze aversion from conversation participants is used to prevent mutual gaze [6,12] since gaze may be considered as a request signal to be included into the current conversation [2]. However, in a highly dynamic conversation, an overhearer will have difficulty avoiding attention to, comprehension of, and reactions to the conversation.

In addition to the behaviors associated with the conversation roles, behaviors are also associated with role shifts. One way to signal a change in the conversation role is for behaviors associated with the current role to be avoided and those associated with the new role to be adopted. For example, gazing at the speaker and making mutual gaze signal role shifting from a bystander to a side-participant or an addressee [2,12]. When the role shift involves changes in the participation goal, interpersonal distance is also adjusted by either moving toward or away from the group to join or leave the conversation [21].

10.2 Processing the Signals

Upon receiving input signals from the dialog module, NVBG updates the agent's role and goals and determines whether to generate a role shifting behavior. The role shifting behavior occurs when the agent's updated participation goal differs from the current participation goal. For example, if the agent's current role is overhearer (participation goal is false) and the updated role is addressee (participation goal is true), he will enter the conversation group and generate attendance behavior by gazing at the speaker and nodding. If the agent's participation goal is unchanged, NVBG generates corresponding feedback behaviors depending on the comprehension and current participation goal.

As described in Section **??**, NVBG may also receive affective information. The affective reaction dominates the reactions related to partial understanding of the speaker's utterance: an affective signal will have higher priority than the comprehension information. The affective reactions include behaviors such as smiles for joy and furrowed eyebrows for anger.

To convey the evolving comprehension level in behavior, the confidence attribute, which has range [0.0, 1.0], is used to define three categories of understanding: confusion ([0.0, 0.5)), partial understanding ([0.5, 1.0)), and understanding (1.0). The `MAXF` indicator further determines which specific feedback is generated. Since the partial understanding level may only change slightly between adjacent words, the model processes the dialog signal when the difference between previous and current partial understanding level exceeds a threshold (currently set at 0.2).

11 Conclusion and Future Work

In this paper we have presented a Function Markup Language specification for incremental feedback for multiparty conversation. It takes into account semantic and pragmatic processing and attitude toward the topic of conversation. It supports both expressive and evocative feedback for a variety of conversational roles and goals. It has been implemented, and connected to the behavior realizer developed by [35]. Future work includes linkage to coping strategies for more evocative feedback, as well as evaluating the impact of the feedback on users engaged in the SASO4 and other scenarios.

Acknowledgements. 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, and no official endorsement should be inferred.

References

1. Allwood, J.: Linguistic Communication as Action and Cooperation. Ph.D. thesis, Göteborg University, Department of Linguistics (1976)
2. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press (1976)
3. Argyle, M., Lalljee, M., Cook, M.: The effects of visibility on interaction in a dyad. Human Relations 21, 3–17 (1968)
4. Bavelas, J.: Listeners as co-narrators. Journal of Personality and Social Psychology 79, 941–952 (2000)
5. Brunner, L.: Smiles can be back channels. JPSP 37(5), 728–734 (1979)
6. Callan, H., Chance, M., Pitcairn, T.: Attention and advertence in human groups. Soc. Sci. Inform. 12, 27–41 (1973)
7. DeVault, D., Sagae, K., Traum, D.: Detecting the status of a predictive incremental speech understanding model for real-time decision-making in a spoken dialogue system. In: Proceedings of InterSpeech (2011)
8. DeVault, D., Sagae, K., Traum, D.: Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialogue & Discourse 2(1) (2011)
9. Dittmann, A., Llewellyn, L.: Relationship between vocalizations and head nods as listener responses. JPSP 9, 79–84 (1968)
10. Ellsworth, P., Friedman, H., Perlick, D., Hoyt, M.: Some effects of gaze on subjects motivated to seek or to avoid social comparison. JESP 14, 69–87 (1978)
11. Friedman, H.S., Riggio, R.E.: Effect of individual differences in non-verbal expressiveness on transmission of emotion. Journal of Nonverbal Behavior 6(2), 96–104 (1981)
12. Goffman, E.: Forms of Talk. University of Pennsylvania Press, Philadelphia (1981)
13. Goodwin, C.: Conversational organization: interaction between speakers and hearers. Academic Press, London (1981)
14. Gratch, J., Marsella, S.: A domain-independent framework for modeling emotion. Journal of Cognitive Systems Research (2004)
15. Gu, E., Badler, N.I.: Visual Attention and Eye Gaze During Multiparty Conversations with Distractions. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 193–204. Springer, Heidelberg (2006)
16. Hartholt, A., Russ, T., Traum, D., Hovy, E., Robinson, S.: A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture. In: Language Resources and Evaluation Conference (LREC) (May 2008)
17. Heintze, S., Baumann, T., Schlangen, D.: Comparing local and sequential models for statistical incremental natural language understanding. In: Proceedings of SIGDIAL (2010)
18. Heylen, D., Kopp, S., Marsella, S.C., Pelachaud, C., Vilhjálmsson, H.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)
19. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I.: Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In: Proceedings of ICASSP (2006)
20. Ikeda, K.: Triadic exchange pattern in multiparty communication: A case study of conversational narrative among friends. Language and culture 30(2), 53–65 (2009)

21. Jan, D., Traum, D.R.: Dynamic movement and positioning of embodied agents in multiparty conversations. In: Proc. of 6th AAMAS, pp. 59–66 (2007)
22. Kendon, A.: Conducting Interaction: Patterns of Behavior in Focused Encounters. Cambridge University Press, Cambridge (1990)
23. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., 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)
24. Kopp, S., Stocksmeier, T., Gibbon, D.: Incremental Multimodal Feedback for Conversational Agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 139–146. Springer, Heidelberg (2007)
25. 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)
26. Maatman, R.M., Gratch, J., Marsella, S.C.: 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)
27. Marsella, S., Gratch, J.: Ema: A process model of appraisal dynamics. Journal of Cognitive Systems Research 10(1), 70–90 (2009)
28. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. AAMAS 20, 70–84 (2010)
29. Plüss, B., DeVault, D., Traum, D.: Toward rapid development of multi-party virtual human negotiation scenarios. In: Proceedings of SemDial 2011, the 15th Workshop on the Semantics and Pragmatics of Dialogue (September 2011)
30. Sagae, K., Christian, G., DeVault, D., Traum, D.R.: Towards natural language understanding of partial speech recognition results in dialogue systems. In: Short Paper Proceedings of NAACL HLT (2009)
31. Sagae, K., DeVault, D., Traum, D.R.: Interpretation of partial utterances in virtual human dialogue systems. In: NAACL-HLT 2010 Demonstration (2010)
32. Traum, D.: Semantics and pragmatics of questions and answers for dialogue agents. In: Proceedings of the International Workshop on Computational Semantics, pp. 380–394 (2003)
33. Traum, D.R., Morency, L.P.: Integration of visual perception in dialogue understanding for virtual humans in multi-party interaction. In: AAMAS International Workshop on Interacting with ECAs as Virtual Characters (May 2010)
34. Traum, D.R., Rickel, J.: Embodied agents for multi-party dialogue in immersive virtual worlds. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 766–773 (2002)
35. 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)

Designing Relational Agents as Long Term Social Companions for Older Adults

Laura Pfeifer Vardoulakis[1], Lazlo Ring[1], Barbara Barry[1],
Candace L. Sidner[2], and Timothy Bickmore[1]

[1] Northeastern University, College of Computer and Information Science,
Boston, MA, USA
{laurap,lring,bbarry,bickmore}@ccs.neu.edu
[2] Worcester Polytechnic Institute, Worcester, MA, USA
sidner@wpi.edu

Abstract. Older adults with strong social connections are at a reduced risk for health problems and mortality. We describe two field studies to inform the development of a virtual agent designed to provide long-term, continuous social support to isolated older adults. Findings include the topics that older adults would like to discuss with a companion agent, in addition to overall reactions to interacting with a remote-controlled companion agent installed in their home for a week. Results indicate a generally positive attitude towards companion agents and a rich research agenda for virtual companion agents.

Keywords: relational agents, social interfaces, social dialogue, wizard-of-oz study.

1 Introduction

Studies have demonstrated that a lack of social support can have negative effects on the health and well-being of older adults [1], and older adults who face extreme isolation face significantly higher risks of mortality than their connected peers [2]. A recent meta-analysis estimates that 7-17% of older adults face social isolation and 40% experience loneliness [3] (social isolation refers to minimal contact with others, whereas loneliness refers to the subjective, usually negative, reactions to a person's social experiences [4]).

To address these problems, we are developing a virtual agent that can provide social support and wellness coaching to isolated older adults, in their homes, for months or years. This companion agent will be always on, always available, to provide a range of support interactions including: companionship dialogue, game co-play, exercise and wellness promotion, social activity tracking and promotion, facilitating connections with family and friends, and memory improvement tasks, among others.

To inform the design of this agent's dialogue capabilities, we conducted two field studies to determine what older adults would want to talk about with an in-home companion agent.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 289–302, 2012.

2 Related Work

2.1 Social Technologies for Older Adults

Many researchers have explored technologies that provide social activity scaffolding for older adults. In a longitudinal field study, Plaisant, et al., investigated shared, symmetric access for family calendars, as a way for remote, inter-generational family units to stay in touch and improve awareness surrounding daily activities [5]. Wearable and stationary devices that promote multimedia sharing with family and friends have also been designed to improve the social-connectedness of isolated adults [6].

Technologies designed specifically to provide companionship for older adults are an area of recent research. Leite, et al., developed a robotic companion designed for game co-play [7]. Wada, et al., have examined non-conversational therapeutic robots, and Klamer, et al., have examined the health benefits of in-home robots [8,9]. Cavazza, et al., explored the challenges surrounding a conversational agent companion that is able to intelligently ask about a user's day [10]. To explore how agents might be more useful than found in [9], this work undertakes a larger sample of participants in advance of full technology in participants' homes

2.2 Wizard of Oz Methodologies

In a Wizard of Oz (WOZ) study, a user interacts with a computer that is not autonomous, but rather one that is remotely-controlled by another human (often unbeknownst to the user) [11]. This technique is frequently used to explore human-computer interactions that are not possible with current technologies, such as full speech generation and understanding. WOZ methods have been used to explore companion agents, but only in single lab-based sessions [12]. Dow, et al., propose a new design for controlling embodied characters, one that blends both machine and human control [13]. We utilize this approach in the present work.

2.3 Relational Agents

Relational agents are autonomous, embodied agents designed to form relationships with their users by building trust, rapport, and therapeutic alliance over time [14]. These agents are typically designed as computer-animated, humanoid agents that simulate face-to-face dialogue with their users. Relational agents have been successfully used in health interventions, including several designed specifically for older adults [15]. When designing agents to promote social connectedness, relational agents provide several affordances. The agents are autonomous, since family, friends, and caregivers may not be available at all times.. The agents are conversational, because older adults with limited computer literacy are familiar and comfortable with this interaction format. Finally, the agents are relational, in that they are designed for companionship and long-term continual use, and thus can adapt to the changing nature of the socio-emotional relationship users have with them.

3 Preliminary Exploration: Eldercare Companion Volunteers

Our initial approach to understanding how elders might interact with companion agents was to meet with human role models: volunteers who provide periodic visitation to isolated older adults. We collaborated with a non-profit organization in Boston that manages a network of trained volunteers who provide support and assistance to elders and adults with disabilities. Members of our research staff first went through the orientation and training that is provided to new volunteers. We then conducted interviews with four volunteers and accompanied two of them on home visits to their elder "recipients".

The volunteers we interviewed were all women in their 20s (all trainees that we met were also female), and they all described their relationships with their recipients as friendships rather than service relationships. Volunteers visited their recipients once a week for approximately 1-2 hours. Recipients ranged in age from 60 to 97 and were mostly (75%) female. All had mobility and other health problems, keeping three of them mostly at home except when their volunteers took them for walks during visits.

According to the volunteers, the recipients do most of the talking during visits, with storytelling by the elder taking up a significant portion of most interactions. When they are visiting in the elders' homes, the televisions are typically turned on, and chat topics include: storytelling, small talk (weather, etc.), topics occasioned by the television (during co-watching), reports of recent events and future plans ("relationship continuity" behaviors [16]), sports, the recipient's health, and the recipient's family. Two of the volunteers reported that their recipients craved more social contact with their family and friends, but that they didn't want to impose, so rarely initiated contact.

4 WOZ Study: What Do Older Adults Want to Talk about with a Companion Agent?

To further understand how older adults would want to interact day-to-day with an in-home agent companion, we developed a virtual conversational agent that could be placed in the home and be remotely controlled by a researcher for a one-week duration. Since our primary objective was to understand the range of topics that older adults would want to talk about, we designed a research platform in which users could interact with the agent using unconstrained speech and nonverbal behavior.

4.1 The Remote Wizard of Oz System

The system runs on a dedicated computer in an older adult's home and is connected to the Internet. The agent talks using synthetic speech and synchronized nonverbal behavior, while the older adult converses using natural speech and non-verbal behavior that is captured via the computer's integrated microphone and webcam (Figure 1). The real-time audio and video of the older adult are streamed to a Wizard of Oz station, where a research assistant controls the agent responses by choosing preselected utterances and/or animation commands from the control-station software, or by manually typing utterances which are transmitted to the agent for real-time synthesis and animation.

Wizard commands are sent to the agent using an XML command language over a TCP/IP connection. Commands include specifications for spoken utterances, along with coordinating nonverbal behavior (hand gestures, eyebrow raises, head nods, posture shifts, gaze-aways) and facial displays of affect. Nonverbal behavior is generated using BEAT [17] or manually specified by the Wizard. Live audio and video of the participant was streamed to the Wizard using the Skype4Com[1] API, and archived using VodBurner[2]. All interface actions taken by the Wizard were also logged with timestamps.

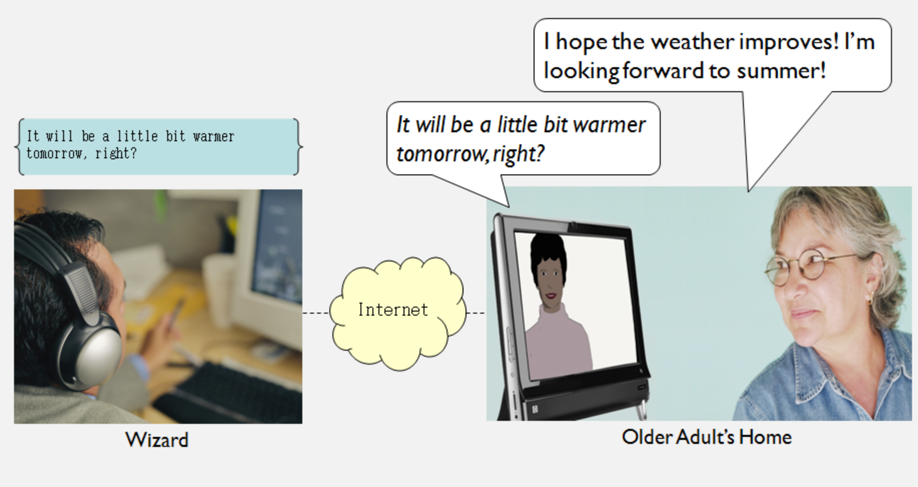

I hope the weather improves! I'm looking forward to summer!
It will be a little bit warmer tomorrow, right?
It will be a little bit warmer tomorrow, right?
Internet
Wizard
Older Adult's Home

Fig. 1. Wizard-Agent Setup

4.2 Methods

Participants were recruited via an online job recruiting site. In order be eligible, participants needed to be 55 or older, speak English, live alone, and have a high-speed Internet connection. A research assistant met participants in their home to obtain informed consent, collect baseline measurements, install the agent computer and connect it to the Internet via the participant's existing network connection. Participants went through a simple introductory conversation with the agent with the research assistant, to make sure they were comfortable with the experience.

Participants were told that they could have daily conversations with the agent during a pre-scheduled 90-minute time window (when the wizard would be standing by). At the end of the week, a research assistant revisited the older adult in their home, to

[1] http://developer.skype.com/accessories/skype4com
[2] http://www.vodburner.com/

administer final measures, conduct a semi-structured interview about their experience, and collect the study computer. Study measures included socio-demographics, and the UCLA Loneliness Questionnaire [18] at intake, and an Agent Satisfaction Questionnaire at termination.

Four research assistants played the role of the Wizard. In keeping with the exploratory nature of the study, Wizards were given no instructions regarding what they should talk about with participants, only that they should have a "conversation".

Privacy and Ethical Issues. Since the agent computer could be remote-controlled to begin video streaming by the wizard, participants were told how to tell the camera was active (an illuminated LED), and how to cover the camera if they wanted to ensure privacy. Although active deception is commonly used in Wizard of Oz studies so that participants think they are interacting with a fully automated system, we felt that carrying on such deception in a participant's home, over an extended period of time, was unwarranted. Participants were told in advance that the agent was not autonomous, but rather remote-controlled by a person at all times.

4.3 Participants

Twelve older adults (10 women, 2 men), aged 56-73 (*m*=62) participated in the week-long study. Participants were mostly Caucasian; two were African-American. Participants were generally well-educated (all but one had some college) and came from diverse working backgrounds. Five participants were retired. Participants scored between 26-53 (*m*=38.6, *sd*=8) out of a possible 80 on the UCLA loneliness measure, indicating that most participants reported low levels of loneliness.

4.4 Results

Participants had between 1 and 5 conversations with the agent (*m*=3.5), with conversations lasting between 1.95 to 122.31 minutes (*m*=28.33, *sd*=20.73).

Conversational Topics. Audio from all agent-participant dialogues was coded for high-level topics of conversation, along with the start and end of each topic boundary. A preliminary list of topics was created by consensus of the researchers following preliminary review of the dialogues. Coders added topics to the list if they felt that none of them adequately described a dialogue segment they were reviewing. In total, 70 distinct topics were discussed during the 41 agent-participant interactions (Appendix I).

We find that the agent-participant conversations were highly individualized and that topics varied greatly, ranging from discussions of *Family* and *Friends* to *Music*, *News* and *Fashion.* Fifty-nine percent of all topics were not discussed by more than one participant (Table 1).

Despite this, there were many topics in common across participants. Table 2 presents examples of the most common topics. The three topics discussed by nearly all participants (other than greetings and farewells) were: *Family*, *Weather* and *Story-*

telling. Discussion of *Future Plans* and asking *Questions to the Agent* also took place by more than half of the participants, ranging from inquiries about the agent's functionality to questions about its development trajectory and future applications. We also examined topics that were common across multiple conversations and found that, *Storytelling*, *Weather*, *Future Plans,* and *Family* were brought up in at least half of all Agent-Participant conversations.

Table 1. Agent-ParticipantConversation Information

Participant	Num Conversations	Avg Conv. Length (Minutes)	Top Topics	Time Spent on Topic
1	4	41.62	Storytelling Miscellaneous Food	26.30% 19.13% 6.84%
2	4	12.34	Miscellaneous Report Future Plans	31.12% 14.16% 10.20%
3	5	14.87	Storytelling Wizard of Oz Future Plans	14.62% 13.19% 10.08%
4	4	17.33	Storytelling Future Plans Opinions	60.12% 9.97% 6.68%
5	5	50.05	Television Greeting Storytelling	12.45% 12.44% 9.70%
6	3	22.66	Sports Agent Weather	41.44% 11.95% 6.14%
7	2	24.55	Travel Daily Activities Habits	15.36% 11.94% 11.34%
8	3	20.10	Questions to Agent Storytelling Goodbye	12.31% 12.03% 11.22%
9	3	17.53	Questions to Agent System Greeting	27.96% 13.99% 9.33%
10	4	26.28	Storytelling Wellness Family	21.56% 16.30% 13.18%
11	4	47.59	Storytelling Agent Exercise/Wellness	18.50% 16.01% 14.21%
12	1	54.77	Family Agent Miscellaneous	35.10% 21.38% 9.39%

Table 2. Examples of frequent conversation topics (*Tanya* is the name of the agent)

Topic	Example
Family	*"I'm the oldest in my family … I have a younger sister …"* – P7 *"I had to mail my grandson his weekly letter..."* –P10
Weather	*"I'm doing well – I just came back and it's freezing out! I had to go out and do a bunch of errands and it's so cold out!"* –P10
Storytelling by elder	*"Would you like me to tell you about working on my Great-Aunt's tobacco farm when I was a kid? …"* – P1
Future Plans	P11: *"Would you like to talk again tomorrow?"* Agent:*"Yes I would."* P11: *"So would I."* Agent: *"What time are we on for?"* P11: *"Well the afternoon, …"*
Questions to the Agent	*"Tanya, did the computer school design you? Or whose project are you?"* –P8 *"Do you have facial expressions, Tanya?...oh, a smile, great!" - P11*

Conversation Topics of Specific Importance to Older Adults. Several topics were identified that are of particular importance to the design of companion agents for older adults (Table 3).

Activity identification and planning. Participants discussed activities as past events, new activities and future plans. While all participants mentioned lifestyle activities (e.g. reading, walking, seeing friends) those who scored as the least lonely (P11, P9) demonstrated more activity planning (Table 3. 1a & 1b). Some planning statements included specific details connoting commitment, such as picking up bus schedules or reaching consensus with activity partners, while other planning statements expressed positive or negative sentiment about an event, either in anticipation or reflection. Studies in psychology and neuroscience have demonstrated the broad health benefits of cognitive enrichment activities and physical exercise for aging adults [19,20]. While a generic increase in activity improves health, amplified benefit is obtained by tailoring for engagement [21], variety of cognitive demand [22] and framing health messages in interactive systems for older adults [23]. Personalization of activity planning by virtual agents to best support older isolated adults requires detailed re-search into activity planning habits of older adults. As virtual agents are engaged in long-term interactions with users, enabling detailed user models, activity recommendations can be honed in support of the greatest individual health benefit.

Character strength disclosures and attitudes toward aging. Participants offered repeating statements revealing their character strengths [24]. Attitudes toward aging were less explicit than character strength disclosures (Table 3, 2a & 2b). Distancing from negative attributes of aging was more prevalent than direct statements about positive aging. Three participants distanced themselves from "old people" who were sedentary or ruminative about their physical ailments. Identification of positive and negative attitudes towards aging would present an opportunity for intervention. Longevity studies show that a positive attitude toward aging (e.g., that aging offers wisdom and more free time rather than memory loss and loneliness) increases life expectancy by 7.5 years on average [25].

Family history and social ties. Our connections to others can be expressed in many forms, from personal narratives to calendars and to-do lists (Table 3, 3a & 3b). Participants recounted stories about family and friends providing fodder for reconstructing their social networks. Personal narratives included self-explanation of physical proximity, frequency of interactions, and social support akin to network connections in the covey model [26]. Connectedness of some participants was closely linked to community-based, scheduled events. For older isolated adults, being able to understand and utilize networks of support can mitigate isolation [27]. Six participants explicitly defined others as sources of and recipients of help, further defining the roles of people in their social network. Virtual agents may be to help older adults create new social ties and maintain existing ones to meet their health needs.

Table 3. Topics Important for Older Adults

Topic	Example Utterances
1a. Activity identification	*"When I'm traveling I enjoy shopping, …" - P11* *"Now it's golf, which is a lot easier for me. Well not to do well in but at least to participate in" - P6*
1b. Activity planning	*"...once it gets cold, it's a whole different kind of a flow in terms of planning and travel".- P5* *"Maybe that could be my goal...to make sure I go to the dancing tonight. Is that okay" - P11*
2a. Character strengths	*"I have to be on the move." (Vitality)- P1* *"I went to a fundraiser for charity to raise money for an orphanage." (Altruism)- P9*
2b. Attitudes toward aging	*"… some seniors have nothing better to do than to just sit around and just gossip and you know." - P1* *"Being retired is new to me. That's why I roam around so much." - P2* *"I think in this country unlike other countries older people aren't as valued and aren't as much a part of the community" - P9*
3a. Family and friend histories	*"My mother's sister was married to a man in western mass and they had a truck farm". - P1*
3b. Social ties	*"I just lost my dog Sam who is a Lab at age 13 about six months ago and he was my best pal. - P8* *"It is interesting I don't know them particularly but I think we feel a commitment to each other in the sense that if something happens I'd feel comfortable calling any of them saying I'd need help, and they'd be right there, even though we don't socialize. - P11*

Participant Reactions to the Agent. Participants reported high levels of satisfaction with the agent and indicated that they were comfortable having her in their home (Table 4).

Table 4. Agent Satisfaction Measures and Scores

Question	Anchor 1	Anchor 7	Mean (SD)
How satisfied were you with Tanya?	Not at all	Very satisfied	6 (1.09)
How much would you like to continue working with Tanya?	Not at all	Very much	5.36 (1.68)
Would you rather have talked to a person than Tanya?	Definitely prefer a person	Definitely prefer Tanya	4.08 (1.78)
I feel comfortable having Tanya in my home.	Disagree completely	Agree completely	5.7 (1.05)

We also conducted in-person, semi-structured interviews with participants to further explore their experience with the in-home agent. These interviews were audio-recorded, transcribed, and coded for themes.

All participants had something positive to say about their experience with the agent (Tanya); four participants (P1, P2, P10, P12) had extremely positive reactions. For many, Tanya provided a sense of companionship and support.

> *"Yeah and I thought that I was going to cry because it was like losing a friend after talking to her for so many days …."* –P1

> *"I was very pleasantly surprised to find that there was such a connection to what I knew was actually a computer generated human being … It did not feel like fantasy land although I didn't have the delusion that I was really talking to a human person there. I mean I was and I wasn't but I felt a connection and as I told you before I feel that there was an accountability built in there. And support."* – P2

Eight participants (P1, P3, P4, P5, P6, P7, P8, P12) reported some negative comments regarding their experience with Tanya. Most of the negative reactions had to do with the lack of realism, the static nature of the interactions and the simplicity of Tanya's abilities. A few participants simply did not feel a connection to Tanya, and one participant (P8) reported that the interactions with Tanya made her feel worse, because they made her realize that she lacked the human interactions and the friendships that she desired for her life.

Privacy. Four participants (P1, P4, P5, P6) expressed no privacy concerns with the agent in their home. On the other hand, 7 participants expressed strong privacy concerns (P3, P6, P7, P8, P9, P10, P12). These concerns mostly revolved around the use of the webcam and uncertainty about whether or not they were recorded. Another factor that increased concerns about privacy was that the computer screen was on at all times, though dimmed most of the time. One participant (P6) ended up turning the computer to the side when it wasn't in use, in order to prevent the webcam from having any possible view of his home. Two other participants (P1, P12) stated that they strategically placed the computer in a location where the camera would only be able to view a very small space of their home.

The Wizard Effect. Participants reported that throughout the study, they were cognizant of the Wizard of Oz component. For a few (P1, P2, P4, P9), the wizard component was in the background, and they viewed their experience as interacting with Tanya. For several others, the wizard component was in the foreground, and for some, not knowing who was *truly* behind the interaction caused anxiety.

> *"Well, rationally, I knew that there was a person controlling Tanya but it didn't feel like that."* – P2

> *"I didn't know if it was one or more people behind the scenes. It made me uncomfortable that I didn't know who was listening or watching."* – P6

Always On. Finally, we asked participants about their potential desire to interact with the agent throughout the day, instead of during a restricted time frame. While many participants found it convenient to have a specific interaction time, a few expressed positive reactions to interacting with the agent freely throughout the day. However, many of those participants also cautioned that they would want a sense of control over the interactions and the ability to turn the system off, if necessary.

> *"I would just like to make sure that there is an understanding – such as, when you call someone on the phone and they tell you that this is not a good time to talk, you can call back at a time that is good to talk. The thing for me is that if [the agent was] here all the time I would like that accessibility to be able to have the companionship all the time, but I would like to make sure that it is set up so that I don't have to rearrange my schedule to talk to her. I would like to be able to start and stop talking whenever I want to."* – P5

5 Conclusion

We found high levels of acceptance of and satisfaction with the in-home social support agent by older adults in the WOZ study, with many participants stating that it provided them with a sense of companionship. Across both field studies, we found that elders would like to tell stories to and discuss the weather, their family, and their future plans with a live-in companion. *Storytelling* is particularly interesting because it is the topic that elders in both studies spent the most time on. In the WOZ study, participants spent between 1.8 and 43.87 minutes (m = 16.98, sd = 15.98) telling stories to the agent. This indicates that the ability of agents to share in a storytelling experience would be valued and utilized by older adults. We also found that discussion of topics important for the social support of elders—including *Activity Planning*, *Attitudes Towards Aging*, and *Social Ties*—may require especially nuanced dialogue, although WOZ participants did volunteer much of this information on their own.

As discussed in Section 4, this work does have limitations. The in-home video recording utilized for WOZ purposes made eight participants somewhat uncomfortable, thus, the data collected might not be representative of completely anonymous conversations with an agent.

Despite this limitation, these studies provide a research agenda of dialogues to emulate in companion agents designed to provide social support for older adults. Our next steps involve implementing and testing an autonomous companion agent that is able to conduct many of these conversations without the support of a human Wizard, integrating information from the Internet (weather conditions, sports scores) and sensors (motion, vision, prosody) to develop a system that is able to provide adaptive, tailored social support over months or years of operation.

Acknowledgments. We thank Connor Westfall, Colin Shaughnessy, Bridgette Collado, and Ashley Kline for their assistance with this study. This material is based upon work supported by the National Science Foundation under Grant No. 1012086. 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.

References

1. Bassuk, S., Glass, T., Berkman, L.: Social Disengagement and Incident Cognitive Decline in Community-Dwelling Elderly Persons. Annals of Internal Medicine 131(3), 165–173 (1999)
2. Berkman, L., Syme, L.: Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda county residents. American Journal of Epidemiology 109(2), 186–204 (1979)
3. Dickens, A., Richards, S., Greaves, C., Campbell, J.: Interventions targeting social isolation in older people: a systematic review. BMC Public Health 11(1), 647 (2011)
4. Grenade, L., Boldy, D.: Social isolation and loneliness among older people: issues and future challenges in community and residential settings. Aust. Health Review 32(3), 468–478 (2008)
5. Plaisant, C., Clamage, A., Hutchinson, H., Bederson, B., Druin, A.: Shared family calendars: Promoting symmetry and accessibility. ACM Trans. Comput-Hum. Interact. 13(3), 313–346 (2006)
6. Chen, C.-Y., Kobayashi, M., Oh, L.: ShareComp: sharing for companionship. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, Portland, OR, USA, pp. 2074–2078. ACM (2005)
7. Leite, I., Mascarenhas, S., Pereira, A., Martinho, C., Prada, R., Paiva, A.: ”Why Can’t We Be Friends?” An Empathic Game Companion for Long-Term Interaction. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 315–321. Springer, Heidelberg (2010)
8. Wada, K., Shibata, T.: Robot Therapy in a Care House - Change of Relationship among the Residents and Seal Robot during a 2-month Long Study. In: The 16th IEEE International Symposium on Robot and Human interactive Communication, RO-MAN 2007, August 26-29, pp. 107–112 (2007)
9. Klamer, T., Ben Allouch, S.: Acceptance and use of a social robot by elderly users in a domestic environment. In: Pervasive Computing Technologies for Healthcare, PervasiveHealth (2010)

10. Cavazza, M., Raul Santos de la, C., Turunen, M.: How was your day?: A companion ECA. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada, vol. 1, pp. 1629–1630. International Foundation for Autonomous Agents and Multiagent Systems (2010)
11. Kelley, J.F.: An iterative design methodology for user-friendly natural language office information applications. ACM Trans. Inf. Syst. 2(1), 26–41 (1984)
12. Bradley, J., Benyon, D., Mival, O., Webb, N.: Wizard of Oz experiments and companion dialogues. In: Proceedings of the 24th BCS Interaction Specialist Group Conference, Dundee, United Kingdom, 2010, pp. 117–123. British Computer Society (2010)
13. Dow, S., Mehta, M., MacIntyre, B., Mateas, M.: Eliza meets the wizard-of-oz: blending machine and human control of embodied characters. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, 2010, pp. 547–556. ACM (2010)
14. Bickmore, T., Caruso, L., Clough-Gorr, K., Heeren, T.: 'It's just like you talk to a friend' relational agents for older adults. HCI and the Older Population 17(6), 711–735 (2005)
15. Bickmore, T., Schulman, D., Yin, L.: Maintaining Engagement in Long-term Interventions with Relational Agents. Applied Artificial Intelligence: AAI 24(6), 648–666 (2010)
16. Gilbertson, J., Dindia, K., Allen, M.: Relational Continuity Constructional Units and the Maintenance of Relationships. Journal of Social and Personal Relationships 15(6), 774–790 (1998)
17. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: SIGGRAPH 2001, Los Angeles, CA, pp. 477–486 (200)
18. Russell, D., Peplau, L.A., Curtrona, C.E.: The revised UCLA lonelines scale: concurrent and discriminant validity evidence. Journal of Personality and Social Psychology 39, 472–480 (1980)
19. Teri, L., Lewinsohn, P.M.: Modifcation of Pleasant and Unpleasant Events Schedule for use with elderly. Journal of Consulting and Clinical Psychology 50, 444–445 (1982)
20. Mahncke, H.W., Bronstone, A., Merzenich, M.M.: Brain plasticity and functional losses in the aged: scientific bases for a novel intervention. In: Aage, R.M. (ed.) Progress in Brain Research, vol. 157, pp. 81–109. Elsevier (2006)
21. Sheldon, K.M., Lyubomirsky, S.: How to increase and sustaing positive emotion: The effects of expressing gratitude and best possible selves. The Journal of Positive Psychology 1, 73–78 (2006)
22. Carlson, M., Parisi, J., Xia, J., Xue, Q., Reobok, G., Bandeen-Roche, K., Fried, L.P.: Lifestyle Activities and Memory: Variety be the spice of life. Journal of the International Neuropsychological Socieity 18, 286–294 (2012)
23. Bickmore, T., Schulman, D., Yin, L.: Maintaining engagement in long-term interventions with relational agents. Applied Artificial Intelligence: AAI 24(6), 648 (2010)
24. Peterson, C., Seligman, M.E.P.: Character strengths and virtues: a handbook and classification. Oxford University Press, Inc., New York (2004)
25. Levy, B.: Improving memory in old age through implicit self-stereotyping. Journal of Personality and Social Psychology 71(6), 1092–1107 (1996)
26. Antonucci, T.C., Akiyama, H.: Social Networks in Adult Life and a Preliminary Examination of the Convoy Model. Journal of Gerontology 42(5), 519–527 (1987)
27. Hooyman, N., Kiyak, H.: Social Gerontology: a multidisciplinary perspective, 9th edn. (2009)

Appendix I: Full List of Conversational Topics

Topic	Num Distinct Participants	Avg duration (seconds)
Agent	3	139.62
Books	4	161.14
Boston	2	56.88
Boston/New England	6	92.46
Computers and Older Adults	1	250.69
Daily activities	5	56.48
Education	1	40.95
Exercise and wellness	3	172.33
Family	11	150.02
Fashion	1	37.13
Fitness	1	59.26
Food	5	153.75
Friends	5	91.72
Future plans	9	53.39
Goodbye	12	43.92
Greeting	12	66.52
Habits	5	44.48
Health	4	59.00
Hobbies	1	17.02
Job	1	109.87
life lessons, morals, ethics	3	135.26
Loneliness	1	36.80
Medical	3	119.38
Miscellaneous	7	94.07
Miscellaneous (articles)	1	161.08
Miscellaneous (cartoons)	1	76.14
Miscellaneous (casino)	1	138.13
Miscellaneous (children)	2	168.64
Miscellaneous (christmas)	1	48.32
Miscellaneous (colors)	1	72.85
Miscellaneous (computers)	1	48.57
Miscellaneous (current events)	1	56.91
Miscellaneous (flashmob)	1	327.72
Miscellaneous (halloween)	1	21.35
Miscellaneous (holidays)	1	238.07
Miscellaneous (internet)	1	109.64
Miscellaneous (plants)	1	71.41
Miscellaneous (poker)	1	104.99
Miscellaneous (Richmond)	1	93.53
Miscellaneous (smiling)	1	41.18
Miscellaneous (weekend)	1	36.31

Movies	3	192.99
Music	1	69.68
New England	1	35.80
New England/Boston	1	122.26
News	1	77.94
Opinions	5	59.13
Participation in Research	1	35.70
Personal	4	58.36
Personal history	3	56.42
Pets	2	167.54
Politics	2	120.75
Questions	2	39.53
Questions for the agent	8	67.67
Report	6	60.61
Research on Computer Agents	1	135.63
Sports	6	180.15
Storytelling	10	161.71
Surfing Internet	1	41.93
Technology	3	80.70
Television	4	102.84
Thanksgiving	1	18.48
Travel	6	64.49
Weather	11	40.23
Wellness	5	87.81
Wellness Follow-up	1	76.69
Wellness: Goal Setting	1	114.74
Work	4	109.00
WOZ	5	78.61
WOZ: Ideas for use of system	2	80.66

A Cognitive Model for Social Role Compliant Behavior of Virtual Agents

Jeroen de Man[1], Annerieke Heuvelink[2], and Karel van den Bosch[2,⋆]

[1] VU University Amsterdam
j.de.man@vu.nl
[2] TNO
{annerieke.heuvelink,karel.vandenbosch}@tno.nl

Abstract. This paper presents research on how to model the characteristics of social groups into the constituent members of that group. A (virtual) person can belong to different social groups simultaneously (e.g. family, religious community; war tribe, etc). Each group has their own characteristics, such as common goals or a set of norms, which (partly) determine the behavior of the individuals. We developed a method to generate behavior of virtual characters as a function of the social groups they belong to. This is achieved through calculating plan utilities by taking into account the social groups, personal preferences, and the situational context. The method is tested using a military house-search scenario, revealing that our virtual characters acted in accordance with their social groups, even in the face of conflict between groups, by expressing behavior relevant to one or more of their social roles.

1 Introduction

The norms and values shared within a social group shape for a significant amount the behavior of its members. For example, a western male adult behaves differently in a family setting than when in a football stadium, and yet again different when in church. Furthermore, the role that a person holds within a social group also affects how someone behaves, or is expected to behave. For example, in a football club the captain behaves differently than a substitute player. Sometimes, an individual may find itself in the context of two (or more) different social groups whose norms conflict. For example, if the father brings his children along to the stadium, his mates are likely to experience other behavior of him than they normally do. The situation compels the father to weigh and balance the norms and expectations of both groups. How the father will behave is determined by many factors, like the importance that the social group attaches to particular behavior, the value that the individual assigns personally to displaying the particular behavior et cetera. Although it is difficult to predict the resulting behavior precisely, it is always a function of weighing profits and losses.

In order to understand someones behavior, it is necessary to know the current social group(s) that person belongs to, the role of the individual within the group,

[⋆] With thanks to co-supervisors Tibor Bosse and Jurriaan van Diggelen.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 303–310, 2012.

and the norms that apply to the specific group and role. When one is unfamiliar with the applicable norms, it can be hard to understand why someone behaves in the way he or she does. Erroneous interpretations of someones intentions may easily arise. And in some circumstances, this may have serious consequences. An example is the military. Current military missions are often staged in faraway countries with non-western cultures and unfamiliar social groups. Yet, when confronted with individuals, or a group of people, it is imperative for commanders and their teams to interpret their behavior accurately and timely (McFate, 2005). There is a growing awareness that training plays an important role in preparing the soldier for missions in unfamiliar settings (Muller et al., 2011). In this paper we present work on the development of a model that generates the behavior of socially compliant intelligent agents that can be used for such training.

2 Background Research

Tajfel (1972) introduced the concept of social identity as 'the individual's knowledge that he belongs to certain social groups together with some emotional and value significance to him of this group membership', which served as the beginning of the social identity theory. Self-categorization theory (Turner et al., 1987) explains group behavior by stating that people can categorize themselves at different levels of abstraction. Most important here is the level that defines social identity; the ingroup-outgroup level. Categorization at this level evokes the so-called process of *depersonalization*, which 'brings self-perception and behavior in line with the contextually relevant ingroup prototype' (Hogg & Terry, 2000). It is thus possible for individuals to behave not according to their own personality, but according to some *prototype* of a particular (social) group. As people fulfill a particular role in a social group they 'rapidly internalize social norms about what their roles entail' (Sunstein, 1996). Hofstede & Hofstede (2004) point to a problem forthcoming from this view: 'As almost everyone belongs to a number of different groups and categories at the same time, we unavoidably carry several layers of mental programming within ourselves, [..] The mental programs from these various levels are not necessarily in harmony.' The model we propose is designed to handle this issue arising from combining multiple social groups.

Culture, a particular type of social group, has already been incorporated in various cognitive models. One line of research models culture by means of norms and obligations (Conte et al., 1999). Castelfranchi et al. (2000) proposed an architecture in which agents are able to communicate, adapt and violate norms. Unfortunately, there is a lack of research in 'intelligent violation of norms' (Castelfranchi et al., 2000). The Cultural Cognitive Architecture proposed by Taylor & Sims (2009) uses *schemas* and *appraisal theories of emotion* in their model. A downside of this architecture is that it currently only consists of a theoretical framework. FAtiMA is an architecture incorporating emotions and personality (Dias & Paiva, 2005) and has been expanded to incorporate culture using *symbols* and *rituals* (Aylett et al., 2009; Mascarenhas & Paiva, 2010), however the method does not generalize well to social groups.

Solomon et al. (2008) describe a model of Culturally Affected Behavior (CAB). Here, culture is modeled as a network of actions affecting mental states. A mental state has a current utility (to what extent it is valid in the given situation) and an intrinsic utility that reflects the importance of the mental state to the agent. E.g. a mental state *is-observant-of-Islam* has a high intrinsic utility to an agent representing a pious Muslim, but a low intrinsic utility to an agent representing a western atheist. The action to offer alcohol to a Muslim agent would decrease this agent's view that the person offering him alcohol is observant of Islam. Based on the current utilities of an agent, a Socio-Cultural Satisfaction (SCS) is calculated whereby intrinsic utilities are used as a weight for the respective state. The work of Bulitko et al. (2008) shows that a variety of factors, such as emotion or personality, can be incorporated in determining an agent's actions using this method. Below, we propose a model where this method is used to create social compliant behavior.

3 Social Compliant Behavior Model

In this paper our model of social compliant behavior is illustrated in a military context where a soldier needs to learn how to enter a house and to address the occupants. In this scenario soldiers closely interact with people of different cultures and customs. The occupants are modeled using a set of prototypes, in this example related to a Muslim culture. The occupants can be either a family, a group of militant soldiers, or a combination thereof.[1]

3.1 Prototypes for Roles in Social Groups

Social group and *role* are the main concepts of our model. An agent may belong to one of more social groups simultaneously (e.g. Islam, family) by fulfilling a role within each social group (e.g. muslima, head of family). These roles produce the agent's behavior through so-called (role) prototypes. Fig. 1 shows an example, the various concepts are explained in the following paragraphs.

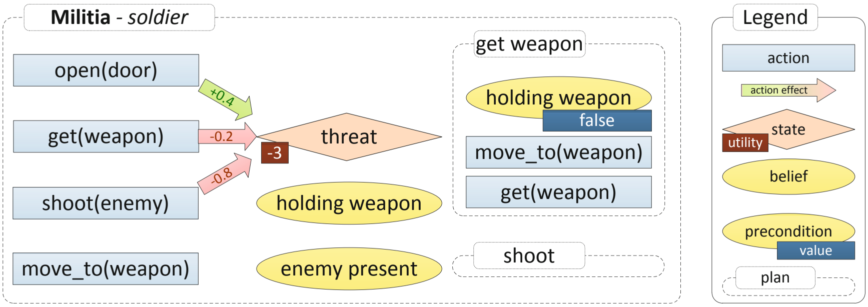


Fig. 1. Visualisation of the prototype of the role *soldier* of the social group **Militia**. The 'get weapon' plan is unfolded; the 'shoot' plan is folded.

[1] For a more detailed description of the model's functioning we refer to Man (2011).

Beliefs. Beliefs represent an individual's knowledge of the world (Rao & Georgeff, 1995). For example, a Muslim woman has a belief expressing whether she is veiled or not; a militant soldier has a belief whether or not he is carrying a weapon (see Fig. 1: a soldier has beliefs about `holding weapon` and `enemy present`). Which beliefs are required by each prototype are defined beforehand, however the actual value or content of beliefs can change at any time.

States. States are the driving force behind plan selection and are therefore very important in the model. The term *state* is first used in the *Culturally-Affected Behavior project* where it is described to 'decode states of the world [..] and have intrinsic utility values that represent the relative importance that the human behavior model has for the state weighed against other states' (Solomon et al., 2008). Like proposed by Bulitko et al. (2008) a state 'has some intrinsic utility / concern-value to the agent'. Thus, states carry information about the world and their presence (or absence) have a value to the agent.[2] The difference with beliefs - that directly link to information about the world - is that states constitute more abstract mental concepts, e.g. compare the belief `enemy present` with the state `threat` in Fig. 1. In our model, states are not defined for individual agents, but for roles of social groups. For example, decency is very important for all Muslim women living in Islamic countries. Not wearing a veil in the presence of non-family members evokes a feeling of indecency that should be avoided. Thus, the prototype of the role "woman" for the social group "Islam" has a state `decency` with a high intrinsic utility.

Plans. Every prototype contains its own set of plans. A plan consists of a sequence of actions and (optional) preconditions. The `get weapon` plan of a soldier shown in Fig. 1 consists of the actions `move_to(weapon)` followed by `get(weapon)`. Plans can have a precondition added, e.g. `holding weapon(false)` as getting a weapon is not needed when the agent already has a weapon.

Action Effects. Plans are composed of multiple actions for which certain effects are expected. One type of expected effect is a change in one or multiple states and is modeled as an action effect (see the arrows denoting the action effects in Fig. 1). For example, the plan to open the door contains the action `open(door)`. By performing this action, we expect the current utility of the agent's state curiosity to decrease as it will find out what is on the other side. Similarly, some actions are expected to increase the current utility of a state; a Muslim woman veiling herself increases her decency. The effects that actions may have on the current value of states can vary in degree. For example, for a militia man grabbing a weapon may decrease his feeling of threat somewhat, but shooting the enemy is likely to decrease his feeling of threat even more.

Observation Functions. Action effects model the *expected* effects of actions, but the *actual* effects may be different. For instance, whether the threat for a

[2] Note that a state does not refer to a particular configuration of the agent, unlike other approaches where a state describes a particular configuration of information.

militia man indeed decreases depends on whether he has eliminated the enemy altogether. An agent establishes the actual effects by means of its observation functions. Observation functions are more detailed and context dependent than action effects. Observation functions are specific for a particular role. Spotting an enemy soldier is very threatening to persons belonging to a militia group, but not (so much) for civilians. Observation functions thus affect states. Moreover, they also affect beliefs. For example, if the observation is made that the door is being opened, the belief that the door is closed needs to be adjusted.

3.2 Agents as a Combination of Prototypes

The previous section described the components of a prototype denoting a role within a social group. An agent is a collection of any number of roles. In addition, an agent contains modifiers to regulate the importance of the different roles as explained below. Furthermore, agents are preconfigured by an initial set of beliefs and states, whose content or value can change over time. During the run of a scenario, the observation functions of the different prototypes adjust these values in real-time. The following paragraphs explain these concepts in more detail.

Prototype Importances. An agent can be related to various social groups via different roles. However, just as any person does not feel equally connected to all of his or her social groups, an agent can also differentiate between each of its roles. This is modeled by defining a set of *static* modifiers that describe the relative importance of each prototype for the agent. A second type of modifiers are the *dynamic* modifiers. The motivation for adding these modifiers comes from the process of *depersonalization* (see Sec. 2 Background Research) that denotes that context affects the importance of roles. To model this, dynamic modifiers are calculated for each social group based on the number of people present belonging to that particular group. The static and dynamic modifiers regulate the current importance of the different prototypes for an agent, which affects the resulting agent behavior. The importance of a particular prototype relative to the other roles is determined by calculating a weighted average of the static and dynamic modifiers for each group.

Observation Processing. An agent processes input (e.g. observations) to respond to its environment. A prototype makes inferences on states denoted by the state transitions, and inferences on beliefs denoted by belief transitions. These transitions express how a belief or state should change for a particular prototype. For example, the belief `door_open` becomes true; or `curiosity` decreases by 0.8. For each prototype such a set of transitions follow from the observation functions. This could result in different, or even conflicting transitions when two or more prototypes are relevant. The actual transitions applied to the current beliefs and states are established using the prototype weights. If conflicting beliefs are derived, the belief transition of the group with the highest weight is processed. For state transitions, a weighted average of the effects is added to the current state value.

Plan deliberation. Deliberation starts with the formation of a list of plans that can be executed at that particular point in time. Each plan may have one or more associated preconditions. Every time the agent starts its plan deliberation, each precondition is checked against the current beliefs. Only plans for which every precondition is met are considered in the deliberation process.

Plans contain actions for which action effects on states may be defined in prototypes. Remember that action effects refer to expected outcomes and do not necessarily constitute actual effects. Plans are prioritized according to expected results. The agent first simulates all action effects for each applicable plan, thereby calculating simulated values for the states taking the prototype weights into account. To evaluate the effects of a plan, a comparison is made between the initial state values and the expected values when executing the plan. For each prototype an agent belongs to, the predicted outcomes of plans are scored. Then, an over-all comparison is conducted, taking the weights of the prototypes into account. The plan having the best outcome is selected.

4 Proof of Concept

The model has been implemented using Jadex (Pokahr et al., 2005) and a small scale evaluation study was conducted as a proof of concept. Three different scenarios were developed to investigate whether prototypes can be used to create social role compliant agent behavior. The context is house-searching by western soldiers in a culturally unfamiliar setting (Islamic), with different compositions of people in the house. The player is a western soldier; the agents represent members of different groups having varying roles. In the first scenario, the group in the house is composed of family members. In the second scenario, a militant soldier and his family occupy the house. In the third scenario, only militant soldiers are present in the house. Within these scenarios, each agent should act according to its roles, taking the context (e.g. other agents) into account.

To instantiate the behavior model for this context, we considered the social groups family, militia and Islam with roles such as child, soldier or muslima. Important states in this context are for example threat with a high intrinsic utility for a soldier and decency with a high intrinsic utility for an Islamic woman. Afterwards, for each role relevant plans such as opening the door or getting a weapon were identified. For each of the actions in these plans, the expected effects on the various states were modeled as being low, medium or high. The following observations were made.[3]

First, the model was able to produce behavior consistent with the agent's social groups and roles, even when agents belonged to multiple groups and served multiple roles. This can be illustrated, for example, by the behavior of the woman Muslim agent. When she hears the knock on the door, she wants to open the door to satisfy her curiosity. However, being an Islamic woman, opening the door would be indecent for her to do. This creates a conflict. The model acknowledges the conflict and uses the relative importance of her roles and states, leading her

[3] For a more detailed description of the implementation and the results see Man (2011).

to decide not to open the door. However, the Muslim man is not restrained by this cultural norm and opens the door. Second, agents' behavior is affected by multiple roles. For example, the militant soldier agent together with its family (second scenario) responds to the door knock by first retrieving a weapon and afterwards opening the door. Retrieving a weapon comes forth from his role in the militia group, while opening the door is the default human response. This demonstrates that the model generates behavior affected by multiple roles. Third, results show that agents' behavior is affected by context. When the militia soldier is accompanied by other militia members (third scenario), it responds by retrieving a weapon and taking cover instead of opening the door. So, in a hostile context, the agent adapts its behavior accordingly.

5 Conclusion

Soldiers in foreign missions need to be able to interact with local people while taking their (cultural) norms into account. These communication skills can be trained in virtual environments, provided that the virtual characters behave according to the norms of their social groups. This paper presents research on how to model the characteristics of social groups into the constituent members of that group. We developed a method to generate behavior of virtual characters as a function of the roles they fulfill in various groups. This is achieved by defining role prototypes reflecting what is important for a particular social group, what plans are available to change the current situation, and what effects are expected from particular actions. An agent, being a combination of role prototypes, uses this information for calculating plan utilities, taking into account the social groups, personal preferences, and the situational context.

The model was implemented in Jadex and tested using a house-search scenario. We found that our virtual characters acted in accordance with their social groups, even in the face of conflict between groups, and were able to express behavior relevant to one or more of their social roles. It is not claimed that the model presented here has sociological or psychological validity. Although concepts and processes have been based on research in those areas, the current implementation has been constructed only at face value. In order to develop valid models, it is needed to utilize more data from relevant studies and theories, validate the created models in human subjects research and experiment with larger implementations in a variety of domains.

The proposed model opens up various paths for future research: personality could be implemented as a separate set of intrinsic utilities and action effects; emotion could be incorporated by relating the current values of states to different emotions which in turn may affect the intrinsic utilities of particular states; or agents could learn from differences between expected and actual effects. The model is obviously not complete, but the architecture allows developing models that produce social compliant behavior in virtual agents. Even with just a few social groups, a large variety of agents can be created by making different combinations. This, in itself, is an important step in developing a wide range of scenarios for interactive training in socio-culturally appropriate behavior.

References

Aylett, R., Vannini, N., Andre, E., Paiva, A., Enz, S., Hall, L.: But that was in another country: agents and intercultural empathy. In: Proc. of AAMAS 2009, Richland, SC, pp. 329–336 (2009)

Bulitko, V., Solomon, S., Gratch, J., van Lent, M.: Modeling culturally and emotianally affected behavior. In: Proc. of AIIDE 2008, Stanford, California, US, pp. 10–15 (2008)

Castelfranchi, C., Dignum, F., Jonker, C., Treur, J.: Deliberative Normative Agents: Principles and Architecture. In: Jennings, N.R. (ed.) ATAL 1999. LNCS, vol. 1757, pp. 364–378. Springer, Heidelberg (2000)

Conte, R., Falcone, R., Sartor, G.: Introduction: Agents and norms: How to fill the gap? Artificial Intelligence and Law 7, 1–15 (1999)

Dias, J., Paiva, A.C.R.: 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)

Hofstede, G., Hofstede, G.J.: Cultures and Organizations: Software of the Mind, 2nd edn. McGraw-Hill, New York (2004)

Hogg, M., Terry, D.: Social identity and self-categorization processes in organizational contexts. Academy of Management Review 25(1), 121–140 (2000)

de Man, J.: Composing agents from role prototypes of social groups. Master's thesis, VU University, Amsterdam, Netherlands (2011), http://www.few.vu.nl/~jmn300/

Mascarenhas, S., Paiva, A.: Creating virtual synthetic cultures for intercultural training. In: CATS 2010 (2010)

McFate, M.: The military utility of understanding adversary culture. Joint Forces Quarterly 38, 32–48 (2005)

Muller, T., van den Bosch, K., Kerbusch, P., Freulings, J.: LVC training in urban operation skills. In: Proc. of EURO SISO/SCS 2011, The Hague (2011)

Pokahr, A., Braubach, L., Lamersdorf, W.: Jadex: A BDI reasoning engine. In: Multi-Agent Programming: Languages, Platforms and Applications, vol. 15, pp. 149–174. Springer US (2005)

Rao, A.S., Georgeff, P.G.: BDI agents: From theory to practice. In: Proc. of ICMAS 1995, pp. 312–319 (1995)

Solomon, S., van Lent, M., Core, M., Carpenter, M., Rosenberg, M.: A language for modeling cultural norms, biases and stereotypes for human behavior models. In: Proc. of BRIMS 2008 (2008)

Sunstein, C.R.: Social norms and social roles. Columbia Law Review 96(4), 903–968 (1996)

Tajfel, H.: Social categorization. In: Moscovici, S. (ed.) Introduction a la Psychologie Sociale, vol. 1, pp. 272–302. Larousse, Paris (1972)

Taylor, G., Sims, E.: Developing Believable Interactive Cultural Characters for Cross-Cultural Training. In: Ozok, A.A., Zaphiris, P. (eds.) OCSC 2009. LNCS, vol. 5621, pp. 282–291. Springer, Heidelberg (2009)

Turner, J., Hogg, M., Oakes, P., Reicher, S., Wetherell, M.: Rediscovering the social group: A self-categorization theory. Blackwell, Oxford (1987)

A Cognitive Social Agent Architecture for Cooperation in Social Simulations

Jackeline Spinola[1,2] and Ricardo Imbert[1]

[1] Facultad de Informática, Universidad Politécnica de Madrid,
Campus de Montegancedo, s/n, 28660, Boadilla del Monte, Madrid, Spain
[2] Faculdade de Engenharia Elétrica e de Computação, Universidad Estadual de Campinas,
Barão Geraldo - CP 6101, 13083-852, Campinas - SP – Brazil
j.spinola@alumnos.upm.es, rimbert@fi.upm.es

Abstract. We present a cognitive agent architecture to allow its agents to exhibit social behaviors influenced both by rational and cognitive motivators. Currently, our focus is on cooperation and our objective is the observation or emergence of non-scripted behavior in social cooperative environments. We present the results of applying our architecture in the design of virtual agents in which they spontaneously cooperate to achieve their individual and social goals.

Keywords: Social Simulation, Cognitive Architecture, Cognitive Agents, Roles, Social Interaction, Cooperation.

1 Introduction

In the last three decades a growing interest in the use of the agent paradigm and computer simulation in the research of social science aspects has been observed [1, 2, 3, 4]. According to Gilbert [5], cognitive agent architectures seem to be an appropriate alternative when trying to simulate systems that resemble social processes like cooperation and lately they have been showing their capacity to enlighten our understanding on many social interaction processes [6, 7, 8, 9].

In a previous work we have presented a mechanism combining cognitive and rational elements on the generation of behaviors for reactive-deliberative agents [10]. We have proposed an agent architecture, COGNITIVA, easy to adapt to new contexts following an incremental development strategy and we have shown that the inclusion of cognitive motivators does not compromise the agent reasoning process [10, 11].

Lately, to extend the architecture, the capabilities of these agents have been reviewed in order to include social abilities. Given the complexity and diversity of social behaviors, our current focus on the development of the social level is on a particular behavior, the cooperative behavior. Cooperation is obtained when agents can act together to a common end [12]. Despite the fact that cooperation can be achieved by designing agents with an appropriate rational processes [13], for those areas interested in social simulations to understand or represent social phenomena, this approach can be considered very weak and limited. We claim that our cognitive motivators can

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 311–318, 2012.

enrich the usually intricate cooperative processes with intrinsic aspects of social life, such as self-interest or uncertainty of success in an interaction, and will enable us the observation or emergence of non-scripted behavior in social cooperative environments.

In the following section we begin reviewing the main characteristics of the original COGNITIVA. Then we present the new elements and processes introduced to allow agents to cope with social cognitive-influenced behaviors. In section 4 we describe an evaluation scenario in which social agents show decentralized and autonomous abilities for degrees of cooperation. We also present some preliminary results and comments. We conclude with a final discussion and future work ideas.

2 The Cognitive Architecture

COGNITIVA [10-11] is an integrated architecture that follows a continuous perception-cognition-action process. It receives its inputs through an Interpreter, which filters the information received according to three perspectives: reactive, deliberative and social. The reactive layer is triggered when it is necessary to give rapid response to changes perceived by the agent. It occurs when certain situations require an immediate action, when there is no time for planning. For other situations, agents have the ability to propose goals, their purposes in the environment, and to design plans to achieve them. A Goal Generator generates agents' goals and is also responsible for maintaining and checking the validity of the Goal set. A Planner is responsible for designing the plans that allow agents to achieve their goals.

All these outputs are collected by a Scheduler, which determines the appropriate order to execute reactions or plans, according to their priority and compatibility. The Reasoner is the component that decides how the plan will be carried out when there are several ways to achieve it. Finally, the effectors execute the associated actions.

One of the main components of the architecture is called Belief. Beliefs include everything an agent knows (or believes it knows) about: (i) rooms, surroundings, the agent needs to know, (ii) objects, devices, the agent has to manipulate or be aware of, and (iii) other agents in the system, both human, human-driven and synthetic, including information about itself. Beliefs are characterized as Defining Characteristics (DC) and Transitory States (TS). DC describe the general attributes of a subject (room, object or agent) and the values of which are normally immutable. Examples of DC are the agent identification and, more important from the cognitive point of view, its personality traits, values that influence the way an agent behaves in a system. TS specify the current state of a subject and vary continuously throughout time. Within the TS, it is possible to include some beliefs such as physical states and emotions.

3 COGNITIVA New Social Layer

The new social layer of our architecture extends agents capabilities since agents can act together and rely on others to achieve their goals or improve their performance. The social layer lies in the central concept of Role, a common approach in the agent community. Although there are some similarities with some of these approaches [15], in COGNITIVA, Role is based on two well-known theories: Role Theory and Activity Theory. Role Theory [16, 17] most relevant propositions about social behavior are the division of labor in society, interaction among heterogeneous specialized positions called roles. Activity Theory, is a way of explaining human activity and behavior in social contexts [18, 19, 20, 21]: "An activity is carried out by a human agent who is motivated by the solution of a problem or an intention, mediated by tools and in collaboration with others" [22]. The division of labor includes the roles and the tasks each individual carries out within the community in which he participates and acts.

In COGNITIVA, 'role' is an abstract concept that defines the position of the agent within the social system. Its associated behaviors are intended to achieve a set of goals. It encapsulates what an agent can do and what it can provide other agents with within its society. A Role is composed of a set of tasks closely related to the role. If one plays a role of a 'teacher', for example, related tasks would be 'to give classes', 'to prepare exams', etc. Each task is defined by some properties: (i) scope (whether or not a task can be done for others), (ii) attributes (any parameter agent needed to execute a task), (iii) list of actions and (iv) the expected results. The roles also demand coordination actions to verify their cardinality or to allow simultaneity between roles.

3.1 Cognition-Influenced Social Agents

The way agent acts is influenced by its cognitive motivators and it is particularly important in a social environment. Personality traits and emotions are used as utility functions when an agent has to decide about its behavior and its strategies of interaction in the social system.

From the designer point of view, the election of the agent's cognitive motivators will depend on the desired system and its objectives. It is worth pointing out that providing different instances of the same agent with different values for their personality traits will make them to have different "personalities", what will cause that their behavior regarding the same situations can be slightly or completely different. Consequently, create a variety of individuals with coherent but diverse behaviors is as easy as providing them with different personality trait values.

Considering for instance, the case of cooperation, we can use the agent's personality traits to determine its degree of cooperation: agents will decide whether or not to interrupt their activities to cooperate. The reason is that, although cooperation can bring interesting benefits, the agent should never lose its autonomy to decide its actions. A less cooperative agent, for example, will give more priority to its own agenda and would not interrupt them to serve other(s). Conversely, a cooperative one will be much more likely to help others, even if it involves some personal effort. In the case of the emotions, for example, the emotion of fear, it is normally a sign that the individual faces (or believes he will face) a situation he evaluates as dangerous. Under these circumstances, if he has to choose, the most common decision would probably be the one that the individual would evaluate as having the higher likelihood of achievement of what he wants or less risk or chance of failure.

4 Evaluation Scenario

In order to evaluate the new architecture, a simple and classic scenario is presented: a prey-predator environment. This context, in addition to being a rich test bed for experimenting with the social capabilities of the agents, allows us to evaluate the influence of cognitive motivators in the agent reasoning process, to what extent their behavior has been enriched as a result of the presence of these motivators and the emergence of different degree of cooperative behaviors.

In the proposed scenario, zebras and lions have to cohabit a virtual savanna. The main objective is for the zebras to survive. The purpose of the simulation is to verify that individuals endowed with cooperative behaviors and the ability to obtain benefits that goes beyond their individual capacities will be motivated to employ them spontaneously. Zebras will have to maintain a collective vigilance, i.e., that at least one of the zebras will assume a sentinel role at any given moment. Vigilance is an important activity since early predator detection can prevent zebras being captured by surprise and killed. As it cannot be carried out simultaneously with their private activities, the cooperative approach allows the zebras to conciliate their activities.

According to Bednekoff [23], an animal will decide whether to assume a sentinel role based on two variables: its energy reserves (a severe hunger may make animals switch their priorities) and its associates' actions: believing that any of them is already monitoring the environment will cause zebras' fear to decrease. The expectancy that a lion may appear is enough to make the zebras want to monitor their environment.

Considering the cooperative perspective, we have specified that zebras are able to perceive others and to communicate with them to pursue collective vigilance. Their behavior will be influenced by personality traits like 'cooperativeness', emotions like 'fear', or physical states like 'hunger' and 'thirst'. Although internally all zebras code are identical, the configuration of different values for the cognitive motivators should produce different behavior regarding the same situation. The implementation of COGNITIVA used for this simulation defines those values using five possible fuzzy linguistic labels: <nothing>, <slightly>, <moderately>, <enough> and <absolutely>.

Reactions are of the type 'fleeing from a predator', while some of the zebras' main goals are: 'to be safe', 'to be fed' and 'to be rested'. Zebras can perform two roles: 'sentinel' and 'regular member' although young zebras cannot play the sentinel role. The sentinel role includes two tasks: 'to monitor the environment' and 'to emit an alarm call' for warning conspecifics of predators. The regular member role includes 'eating', 'drinking', 'resting', and also 'to monitor the environment', to allow isolated agents also be able to watch the environment. There is no explicit definition of the number of sentinels. Some simulation parameters are: as time passes, zebras' physical needs (hunger, thirst, tiredness) increase and when an animal perceives its surroundings are monitored, either by itself or by a conspecific, it feels safer and its fear decreases, otherwise, fear increases [23]. The conflict between zebras' goals emerges when, for instance, the hunger of a sentinel zebra has increased too much and the goal 'to be fed' is triggered. Should this occur, the zebra's personality traits and emotions will influence in the way the zebra manages to achieve its goals. If a zebra can count on another to monitor the environment, they can switch roles and it carries out its new goal of being fed. If the zebra does not attain any cooperation, its hunger will continue to rise to the point it will be pushed to abandon the sentinel role to eat, in spite of letting the herd vulnerable. When zebras find out no one is monitoring the environment, fear begins to increase urging the need for replacing the lost sentinel.

4.1 Preliminary Results

With this initial evaluation, we have obtained some preliminary results that are worth commenting on. Regarding the relationship between cooperation and number of individuals, we have found out that even in simulations where all zebras are <absolutely> cooperative, with a low number of zebras (≤ 3), cooperation is difficult to be obtained and most of the time each individual acts as if it were alone in the system. However, as the number of members on the herd increases, collective vigilance emerges and zebras benefit from the social monitoring. An equilibrium point in which some zebras are eating/drinking/resting while other are monitoring the environment is achieved after some time and a direct reward of this cooperative social behavior is a monitored environment that would decrease the probability of zebras being captured by surprise.

The sentinel role performed among a sample of 6 <absolutely> cooperative zebras (a number that has shown to be adequate for achieving cooperation without overloading the system with potential sentinels) can be seen in Fig. 1. The horizontal lines represent the period each zebra is monitoring the environment. The grey areas represent time interval when none of the zebras are doing it. In the beginning of this simulation run, zebras get hungry, thirsty or tired at the same time, since they were started from similar initial physical conditions. In this situation, all of them abandon their sentinel role to consider their physical needs, finding sometimes hard to find other individual available to carry out this role. After a while, as these needs stabilize, they tend to cooperate and the performance of the sentinel role is spontaneously distributed almost uniformly among them. On average, each zebra has monitored the environment 16,13 % of their time (standard deviation of 2.34 % in 10 simulation runs) and the environment has been monitored 89 % of the simulation time.

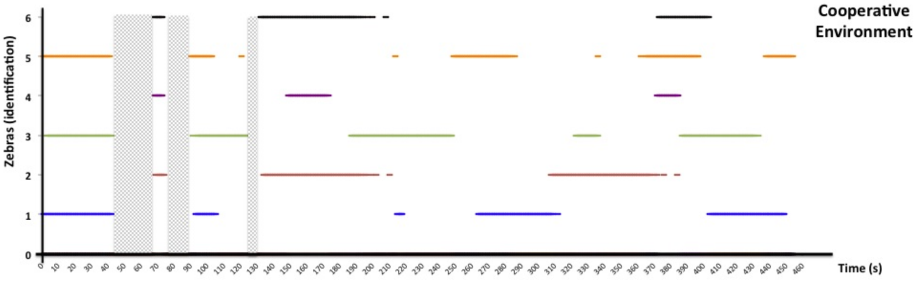


Fig. 1. Variation of sentinel role through time for 6 cooperative zebras

Conversely, when we have increased the variety of zebras regarding their personality traits, from <nothing> cooperative to <absolutely> cooperative, we have observed a different but coherent behavior in the system: although they all receive similar number of requests for adopting their sentinel role, those zebras whose personality traits lead them to a greater degree of cooperation carry out the work of monitoring longer than the others. They have monitored the environment an average of 15,08 % of their time and the standard deviation has increased to 5,34 %. The environment has been monitored 74 % of the simulation time.

The same pattern in the variation of the sentinel role has been verified for simulations with a larger number of zebras, experimenting with 12 y 18 zebras. For a larger number of zebras (> 22), they start colliding physically in the scenario, influencing the execution of their plans. Since it could affect the results, they have been discarded.

Through these results, it was possible to verify that individuals with same goals but distinct personality traits and emotions can behave differently, showing a degree of cooperative behavior. It was also possible to compare the resulting cooperative behavior of the different herds. By means of changes in their cognitive motivators values, different social behaviors have emerged without any central coordination. The simulation has also allowed us to exam the best zebra configuration to minimize the risk of being captured by surprise, what would increase the possibilities of escape in case of a predator approach.

5 Discussion and Future Work

In the architecture presented, the cognitive motivators that influence the agent reasoning processes has allowed us to introduce uncertainty on the agent cooperative behavior, similar to our everyday experience. Throughout a heterogeneous agent configuration we could observe non-scripted behavior in a social simulation. It is a crucial step that enables our architecture to be used for computational emulation of societies for the representation or study of social phenomena.

Detractors to the consideration of a cognitive dimension along with the rational process argue that it provides little benefit. It is true that, in many occasions, pure rational behaviors following centralized, well-defined optimized algorithms may offer better solutions. Nevertheless, for some cases, efficiency is not the only valid measure for the quality of the results. In cases of dynamic/unpredictable environments, where there is not a clear a priori solution, this cognitive alternative can be even more useful.

Another common critique is the impression of loss of control because of the cognitive influence. When we have included several selfish individuals in our experiment, the whole herd has been put in risk more often due to the unwillingness of those individuals to cooperate in pursuit of the social welfare. However, no loss of control can be assumed in that situation since the individuals were just adjusting the priority of their goals in a different way, maintaining the same individual and global purposes.

Our interest in the short term is to keep on analyzing the cooperative capabilities of our agents and also explore new social interaction strategies, such as competition or negotiation, in order to complete and improve COGNITIVA. Furthermore, we aim at making agent cognitive behavior more believable when compared with pure rational behavior in human-agent interaction or for the simulation of social processes. User testing will be used to exam this proposal.

Acknowledgments. First author is funded by Spanish Ministry of Foreign Affairs and Cooperation through the Spanish Agency of International Cooperation for Development.

References

1. Axelrod, R.: Advancing the art of simulation in social sciences. In: Conte, R., Hegselmann, R., Terna, P. (eds.) Simulating Social Phenomena, pp. 21–40. Springer, Berlin (1997)
2. Moretti, S.: Computer Simulation in Sociology: What Contribution? Social Science Computer Review 20(1), 43–57 (2002)
3. Garlick, M., Chli, M.: The effect of social influence and curfews on civil violence. In: 8th international Conference on Autonomous Agents and Multiagent Systems, vol. 2, pp. 1335–1336. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2009)
4. Troitzsch, K.G.: Perspectives and challenges of agent-based simulation as a tool for economics and other social sciences. In: 8th international Conference on Autonomous Agents and Multiagent Systems, vol. 1, pp. 35–42. International Foundation for Autonomous Agents and Multiagent Systems, Richland (2009)
5. Gilbert, N.: When does social simulation need cognitive models? In: Sun, R. (ed.) Cognition and Multi-Agent Interactions: From Cognitive Modeling to Social Simulation, pp. 428–432. Cambridge University Press, New York (2005)
6. Grand, S., Cliff, D., Malhotra, A.: Creatures: Artificial Life Autonomous Software Agents for Home Entertainment. In: First International Conference on Autonomous Agents, pp. 22–29. ACM, New York (1997)
7. Franklin, S.: Autonomous Agents as Embodied AI. Cybernetics and Systems 28(6), 499–520 (1997)
8. Rickel, J., Johnson, W.L.: Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition and Motor Control. Applied Artificial Intelligence 13, 343–382 (1999)
9. Kim, Y.D., Kim, Y.J., Kim, J.H., Lim, J.R.: Implementation of Artificial Creature based on Interactive Learning. In: FIRA Robot World Congress, pp. 369–374 (2002)
10. Imbert, P.R., de Antonio, A.: When Emotion Does Not Mean Loss of Control. In: Panayiotopoulos, T., Gratch, J., Aylett, R., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 152–165. Springer, Heidelberg (2005)
11. Imbert, R., de Antonio, A.: Using Progressive Adaptability against the Complexity of Modeling Emotionally Influenced Virtual Agents. In: Baciu, G., Lin, M.C., Lau, R.W.H., Thalmann, D. (eds.) CASA 2005, Hong Kong, China, pp. 145–150 (2005)
12. Tuomela, R.: Cooperation: A Philosophical Study. Philosophical Studies Series 82 (2000)
13. Bratman, M.E., Israel, D.J., Pollack, M.E.: Plans and resource-bounded practical reasoning. Computational Intelligence 4(4), 349–355 (1988)
14. Arnellos, A., Vosinakis, S., Anastasakis, G., Darzentas, J.: Autonomy in Virtual Agents: Integrating Perception and Action on Functionally Grounded Representations. In: Darzentas, J., Vouros, G.A., Vosinakis, S., Arnellos, A. (eds.) SETN 2008. LNCS (LNAI), vol. 5138, pp. 51–63. Springer, Heidelberg (2008)
15. Cabri, G., Leonardi, L., Ferrari, L., Zambonelli, F.: Role-based software agent interaction models: a survey. Knowledge Eng. Review 25 4, 397–419 (2010)

16. Biddle, B.J.: Recent Developments in Role Theory. Annual Review of Sociology 12, 67–92 (1986)
17. Merton, R.K.: The Role-Set: Problems in Sociological Theory. The British Journal of Sociology 8(2), 106–120 (1957)
18. Leont'ev, A.N.: The Problem of Activity in Psychology. Soviet Psychology 13(2), 4–33 (1974)
19. Leont'ev, A.N.: Activity, Consciousness, and Personality. Prentice-Hall, Englewood Cliffs (1978)
20. Bannon, L.J., Bødker, S.: Beyond the interface. Encountering artifacts in use. In: Carroll, J.M. (ed.) Designing interaction: Psychology at the Human-computer Interface, pp. 227–253. Cambridge University Press, Cambridge (1991)
21. Engeström, Y.: Activity theory and individual and social transformation. In: Engeström, Y., Miettinen, R., Punamäki, R.-L. (eds.) Perspectives on Activity Theory, pp. 19–38. Cambridge University Press, Cambridge (1999)
22. Ryder, M.: Spinning Webs of Significance: Considering anonymous communities in activity systems (1998), `http://carbon.ucdenver.edu/~mryder/iscrat_99.html`
23. Bednekoff, P.A.: Coordination of safe, selfish sentinels based on mutual benefits. Annales Zoologici Fennici 38, 5–14 (2001)

A Model for Social Regulation of User-Agent Relationships

Sandra Gama, Gabriel Barata, Daniel Gonçalves, Rui Prada, and Ana Paiva

INESC-ID and Instituto Superior Técnico,
Univesidade Técnica de Lisboa, Lisbon, Portugal
{sandra.gama,gabriel.barata}@ist.utl.pt
{daniel.goncalves,rui.prada,ana.paiva}@inesc-id.pt

Abstract. Conversational agents have been subject of extensive research. An increasingly wider number of such agents simulate affective behavior in order to convey familiarity and increase believability. Nevertheless, the evolution of social relationships among people occurs gradually and the degree of intimacy associated with such relationships regulates people's behaviors. Similarly, we must take into account the progressive growth of relationships when modeling user-agent interaction. In this paper we present a model that regulates the development of user-agent relationships, articulating the Social Penetration Theory with personality modeling. User tests showed that gradual relationship building achieved through the implementation of our model makes an agent more interesting, while increasing its believability, engagement and fun.

Keywords: Conversational agent, socially intelligent agents, user-agent relationships, social regulation

1 Introduction

Embodied artificial agents have become popular over the last decade [5] [4]. A wide number of such agents rely on user-agent conversation [9] [15] [24], since it plays a major role in these interactions [7]. It has been proven that affect is very important in the creation of relationships among people and even other species [18]. Agents with affective behavior, besides being more believable, are likely to have increased probabilities of building social-emotional relationships with users [5]. The simulation of affective behavior, however, does not simply rely on written dialogue, but it takes advantage of other modalities as well, such as facial and body expressions. Interaction with agents that provide such features is potentially richer and more satisfactory [3]. Affect must be simulated with regard to believability so that relationships between the user and the agent are created and maintained. As a result, we need to take into account the development of social relationships among humans, which goes through several stages and occurs in a gradual, progressive way. In order to do so, we have combined results derived from the Social Penetration Theory [1] with personality modeling based on the Five Factor model [10] to regulate the development of user-agent relationships.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 319–326, 2012.

We present a model that allows user-agent relationships to evolve in a natural, progressive way. Agents that implement this model have the potential to engage users in longer interactions, while maintaing believability.

This paper is organized as follows. In the next section we discuss some research that situates our approach. We then present our model for social regulation of user-agent affective relationships and briefly describe the user-agent interaction. Lastly, we present and discuss the results of user tests.

2 Related Work

Given the importance of conversation in building user-agent relationships [7] and the role that affect plays in creating agents that are both believable and engaging [3], considerable research has been conducted on affective artificial agents. However, the expression of emotion must be carefully taken into account when creating a conversational agent [6]. Particularly, an agent that has an associated model of emotions is likely to better understand the user, thus adapting its responses accordingly [13]. A popular model of emotions for agents is the OCC model [17], according to which emotions are the result of the agent's interpretation of events, other agent's actions and object's features, as well as the agent's reaction to these aspects. Many studies base their work on this model. In order to create a dynamic behavior, the FLAME model [13] represents emotions by intensity through fuzzy logic and, regarding emotional states and behaviors, it maps events and expectations accordingly. Another interesting model for agent behavior is PAR [2], which takes into account the agent's own actions, as well as other agents', and allows acting, planning and reasoning on these actions. It combines the OCC model for emotion analysis and generation with The Five Factor Model of personality traits (OCEAN) [10]. Even though these models allow the simulation of affective actions and study emotion expression to improve relationships, they do not model any form of either social behavior regulation of emotion or gradual relationship development. Actually, several research studies rely on social regulation mechanisms for relationships using the Social Penetration Theory. One such study is Cassel and Bickmore's research [8], which takes into account several concepts underlying this theory as a strategy to obtain collaboration. Another example is Schulman and Bickmore's [22] conversational agent that persuades users to perform physical exercise. A strategy is followed in which superficial topics are discussed, followed by slight self-disclosure by the agent and self-disclosure eliciting. Then, only after empathic actions are performed and conversation status is assessed does a persuasive dialogue take place. All aforementioned models and systems either research the user-agent affective relationship in some way or explore the development of more personal relations with regard to a particular practical goal. However, none of them articulates personality modeling and affect with social regulating mechanisms to create and further develop relationships. We have attempted to bridge these two very important aspects of social behavior with a model that associates the regulation of social relationship gradual building with personality and emotion.

3 A Model for Social Regulation of User-Agent Relationships

In order to endow social user-agent relationships with human-like, gradually developing relationships, our model relies on social regulation of social connections which, consequently, restricts affect expression as well. Our approach thus consists of an articulation between a perception-action paradigm [21] and the Social Penetration Theory [1]. Regarding the first, it is inspired in the studies performed by Rodrigues et al. [21], which is grounded both in the Perception Action Model (PAM) [19] and in Vignemont and Singer's research [11], stating that the agent has to choose an action regarding the perception it builds upon input stimuli. As a result, that action also causes changes in the agent's surrounding environment. These changes are processed, leading to new actions, making up an interaction cycle. The Social Penetration Theory [1] describes the gradual development of social relationships. We have adopted this theory to create a representation for relationship evolution over time. To do so, we took two different definitions into account: *Affinity* and *Intimacy*. The first is related to the establishment of aspects in common during superficial interaction and is more associated with initial stages of a relationship. As for *Intimacy*, it consists of disclosure and exploration of deeper subjects in conversations and is of uttermost importance to the development of deeper relationships. These two concepts are used in our model for the simulation of Social Penetration Theory's four stages [25]: we have defined two variables, aff (affinity) and int (intimacy), that model each of these concepts. Both scores after an interaction (aff_{t+1} and int_{t+1}) are the sum of the values for these variables before the interaction (aff_t and int_t) and values associated with the interaction ($aff_{interaction}$ and $int_{interaction}$). The user chooses a conversation option for interaction, which is assigned values for both affinity and intimacy. As a consequence, despite initial both these variables' initial values being equal to 0, they may assume negative values. Regarding relationship evolution, we modeled each stage according to the aforementioned concepts of *Affinity* and *Intimacy*, following the Social Penetration Theory [1] and the underlying stage definition [25]. Each stage of a relationship has an associated numerical threshold value both for *Affinity* and *Intimacy*. Taking into account Social Penetration Theory's four states, the relationship is on a stage i ($i \in \{1, 2, 3\}$) if the current *Affinity* value (aff) is higher than or equal to the threshold associated with the current stage ($affT_i$) and lower than the threshold for the next stage ($affT_{i+1}$). It should also be verified that the present *Intimacy* value (int) is higher or equal than the current stage's intimacy threshold ($intT_i$) and lower than the threshold for the next stage ($intT_{i+1}$). Since there are no stages beyond stage 4, we have created a special condition for this level, so that stage computation does only take into account the current stage's thresholds ($affT_i$ and $intT_i$). Relationship stage modeling is then used for social regulation. It is actually the main basis for action decision. In fact, our computational model for a socially regulated agent follows Social Penetration Theory's [1] principles associated with each relationship stage when making decisions on which actions to perform. For instance, it is not until the second stage that the model allows the agent to perform disclosure.

On the other hand, on the first and second stages, the agent often displays a polite smile. Regarding physical closeness representation, there are three different proximity frames. The agent's visual representation on the first stage consists of its full body, while at the second stage we can see a closer representation, where it is depicted approximately from its waist up. Regarding further stages, the agent's face is zoomed in, representing increased proximity.

The ways in which people perceive and react are affected by personality. Even though certain particular behaviors may change over time, personality itself remains almost constant over one's lifetime [2]. The Five Factor Model of personality traits [10] has been generally accepted [26]. It represents a taxonomy that captures individual psychological traits and describes the human personality as consisting of five traits: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. We relied on this model to build both the agent's and the user's personality model. A numerical value is assigned to each trait, following a 5-point scale, ranging from 1 (lowest) to 5 (highest). When creating the personality model for a conversational agent that aims at building an evolving relationship with the user, we have defined high scores for all the traits except *Neuroticism*. User personality is taken into account when performing agent decision making. At the beginning of each interaction, since there is no *a priori* available information on the user's personality, all personality traits are assigned an initial score of 3 points out of the aforementioned 5-point scale. As conversation takes place, the user's personality model is iteratively updated. Each option selected by the user to verbally interact with the agent is assigned a tuple of personality traits' values (o, c, e, a, n), ranging from 1 to 5 points, corresponding to the intensity of the traits that are expressed in that interaction. For instance, if the user selected an option where she replied to the user *You are welcome. I'll always be here for you.*, corresponding to a strong agreeableness (while it does not contribute to other factors), the interaction resulting tuple would be $(3, 3, 3, 5, 3)$, with a resulting score of 5 for agreeableness and a 3-point score for all other traits. Personality is updated regarding both these values and the assumptions from the previous model.The previous trait value T_t is weighted with the interaction trait value T_{int}, regarding the relationship stage S. The deeper the relationship is, the less impact a single interaction has upon it, as stated by Altman and Taylor [1].

Our model consists of five main modules: **(i) User Personality Evaluation** takes the user's chosen verbal interaction as input and updates the user personality model regarding the current interaction. It takes into account the valence of the answer regarding all personality traits, as well as resulting affinity and intimacy scores; **(ii) Empathic Appraisal** while also taking the user written interaction as input, this module creates a set of candidate user emotions (happiness, sadness, fear, disgust, surprise, anger, strong happiness and neutral state) that may be associated with the interaction option that has been selected; **(iii) Social Evaluation** as aforementioned, regulates the development of relationships. This module is central, since it regulated merged information from both the user personality model and the set of candidate user emotions to infer the current user emotion. It does so by assigning probability functions to candidate emotions, regarding numerical values of each personality trait, and then choosing the best candidate; **(iv) Agent Emotion Evaluation** processes the current agent's emotion regarding the current relationship status and user emotion, taking into account the agent's personality model; and **(v) Action Decision** makes a decision on which actions to perform, both verbally and visually, taking into account both the current relationship status and both the user's and agent's current state of emotions. However, actions are not limited to written verbal expression. Actually, regarding the fundamentally social and emotional characteristics of relationships [5] and the fact that people respond to social cues from a computer in a similar way to other people's, even if unconsciously [20], we created a model enables the agent to visually represent affect. Since facial expressions are a powerful way to convey emotion [18], we modeled the *six basic expressions* [12]: *happiness*, *sadness*, *surprise*, *anger*, *disgust*, *fear*, as well as the neutral expression and an expression of strong happiness. Furthermore, since the representation of proximity increases the closeness felt by the user [16] [3], our model supports the three aforementioned conversational frames, that are directly related to the relationship's current status of intimacy.

4 Evaluation

We have implemented an agent that is built upon our model, which interacts through written dialogue and expresses both facial expressions and physical proximity. The agent takes the initiative of interacting by prompting the user with a simple phrase. The user then chooses a verbal response out of a list of verbal interactions. The agent reasons upon this answer by updating both the user's model of personality and the relationship's stage and it then infers the user's current emotional state. Finally, it generates a response that is expressed in both a verbal and visual way, to which the user again responds, continuing the interaction cycle.

We created two test conditions. The first consisted of the interaction with an embodied virtual agent that implemented a version of our model without the social regulation component being active. On the second test condition, the user interacted with a visually similar agent where our model was fully integrated. As stated, the objective of this research was to study the impact of our model in the development and regulation of a user-agent relationship. In particular, we intended to study three particular interaction aspects: believability, engagement and fun. To do so, we designed a questionnaire to be filled in at the end of each user test. Along with a small number of profiling questions, and given that friendship is a particularly relevant type of social relationship, we used some questions from an adapted version of the McGill Friendship Questionnaire [23] to infer engagement and fun. In order to study believability, we also created a set of questions comparing user-agent conversation to interaction with other human beings. The resulting questionnaire was subject to validation with 5 users before performing further tests. At the evaluation stage, we started tests by briefly presenting the agent to each test subject, while verbally and visually explaining how to interact. Afterwards, we allowed users to freely interact for ten minutes. Participants were then asked to fill in the questionnaire. We had a total of 30 participants, 15 for each test condition. All subjects were university students, 11 (36.67%) of whom were female and 19 (63.33%) male. Furthermore, 24 (80%) subjects were aged between 18 and 25, while the remaining 6 (20%) belonged to the age group between 26 and 35 years old. Regarding the three model evaluation aspects we have taken into account (believability, engagement and fun), all measured aspects display general higher values when comparing both test conditions. Average believability increased from 3.04 ($\overline{x} = 3.04$, $\sigma = 0.56$) to 3.82 ($\overline{x} = 3.82$, $\sigma = 0.55$), while engagement from 3.73 ($\overline{x} = 3.73$, $\sigma = 0.47$) to 4.49 ($\overline{x} = 4.49$, $\sigma = 0.53$) and fun from 3.38 ($\overline{x} = 3.38$, $\sigma = 0.47$) to 4.11 ($\overline{x} = 4.11$, $\sigma = 0.50$). Looking more closely at the results, a Shapiro-Wilk test showed evidence against normality, suggesting the adequateness of a Kruskall-Wallis test. We were able to conclude that social regulation does in fact have a great impact on either believability, engagement or fun. In particular, when concerning believability, the model version with social regulation (Mdn = 3.67) differs significantly from the model without this feature (Mdn = 3.00) ($U = 187.50$, $p < 0.05$, $z = -3.09$). As for engagement, social regulation also seems to have a great impact, since the version that displays this feature (Mdn = 4.67) is significantly different from the one who does not (Mdn = 3.67) ($U = 189.00$, $p < 0.05$, $z = -3.15$). Regarding fun, there is a significant difference between the condition where social regulation is taken into account (Mdn = 4.00) and the scenario one where this feature is not active (Mdn = 3.33) ($U = 190.00$, $p < 0.05$, $z = -3.19$). This corroborates the previous general conclusions confirming that the social regulation component of our model has a great impact in all aspects we have taken into account, which validates our hypothesis that social regulation plays an important role on user-agent interaction.

5 Conclusions and Future Work

The popularity of conversational virtual agents has increased over the years. The exploration of emotion is such agents, besides improving user-agent relationships, increases believability. However, when regarding the nature of social relationships, we must take into account several particularities, such as gradual development over time. We have created a model that regulates the evolution of relationships, articulating the Social Penetration Theory [1] with the Five Factor personality model [10], allowing a user-agent relationship to naturally unfold. We have performed user tests with an agent implementation of our model, which have shown promising results, ascertaining that our model increases believability, while engaging users in a positive, engaging and fun interaction experience. One very interesting aspect to take into account in the future would be to implement memory mechanisms to further enhance interaction over time, since we already provide social mechanisms that will potentially keep users engaged in interaction for a longer period of time. Such a research study would also provide us with the means to adapt our model to more human-like, longer-term interactions.

References

1. Altman, I., Taylor, D.A.: Social Penetration: The Development of Interpersonal Relationships. Holt, Rinehart and Winston (1973)
2. Badler, N., Allbeck, J., Zhao, L., Byun, M.: Representing and parameterizing agent behaviors. In: Proceedings of Computer Animation, pp. 133–143. IEEE (2002)
3. Bickmore, T., Picard, R.: Subtle expressivity by relational agents. In: Proceedings of the CHI 2003 Workshop on Subtle Expressivity for Characters and Robots (2003)
4. Bickmore, T., Schulman, D.: Practical approaches to comforting users with relational agents. In: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, pp. 2291–2296. ACM (2007)
5. Bickmore, T.W.: Relational agents: Effecting change through human-computer relationships. PhD thesis, Massachusetts Institute of Technology (2003)
6. Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction (TOCHI) 12(2), 293–327 (2005)
7. Campos, J.: May: my memories are yours an interactive companion that saves the users memories (2010)
8. Cassell, J., Bickmore, T.: Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. User Modeling and User-Adapted Interaction 13(1), 89–132 (2003)
9. Cavazza, M., Brewster, C., Charlton, D., Smith, C.: Domain knowledge and multimodal grounding. Technical report, Citeseer (2007)
10. Costa, P.T., McCrae, R.R.: Revised neo personality inventory (neo pi-r) and neo five-factor inventory (neo-ffi). Psychological Assessment Resources, Odessa (1992)
11. De Vignemont, F., Singer, T.: The empathic brain: how, when and why? Trends in Cognitive Sciences 10(10), 435–441 (2006)
12. Ekman, P., Friesen, W.V.: Manual for the facial action coding system. Consulting Psychologist Press, Palo Alto (1978)
13. El-Nasr, M.S., Yen, J., Ioerger, T.R.: Flamefuzzy logic adaptive model of emotions. Autonomous Agents and Multi-agent Systems 3(3), 219–257 (2000)
14. Elliott, C.D.: The a ective reasoner: A process model of emotions in a multi-agent system. Ph D Thesis Northwestern University (1992)
15. Hakulinen, J., Turunen, M., Smith, C., Cavazza, M., Charlton, D.: A model for flexible interoperability between dialogue management and domain reasoning for conversational spoken dialogue systems. In: Fourth International Workshop on Human-Computer Conversation, Bellagio, Italy (2008)
16. Klein, J., Moon, Y., Picard, R.W.: This computer responds to user frustration: Theory, design, and results. Interacting with Computers 14(2), 119–140 (2002)
17. Ortony, A., Clore, G.L., Collins, A.: The cognitive structure of emotions. Cambridge Univ. Pr. (1990)
18. Picard, R.W.: Affective Computing. The MIT Press, Cambridge (1997)
19. Preston, S.D., De Waal, F.B.M., et al.: Empathy: Its ultimate and proximate bases. Behavioral and Brain Sciences 25(1), 1–20 (2002)
20. Reeves, B., Nass, C.: How people treat computers, television, and new media like real people and places. CSLI Publications and Cambridge University Press (1996)
21. Rodrigues, S.H., Mascarenhas, S.F., Dias, J., Paiva, A.: I can feel it too!: Emergent empathic reactions between synthetic characters. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009, pp. 1–7. IEEE (2009)

22. Schulman, D., Bickmore, T.: Persuading users through counseling dialogue with a conversational agent. In: Proceedings of the 4th International Conference on Persuasive Technology, p. 25. ACM (2009)
23. Souza, L.K.: Amizade em adultos: adaptação e validação dos questionários mcgill e um estudo de diferenças de gênero (2006)
24. Ståhl, O., Gambäck, B., Hansen, P., Turunen, M., Hakulinen, J.: A mobile fitness companion (2008)
25. Taylor, D.A., Altman, I.: Communication in interpersonal relationships: Social penetration processes. Sage Publications, Inc. (1987)
26. Wiggins, J.S.: The five-factor model of personality: Theoretical perspectives. The Guilford Press (1996)

Using Collaborative Discourse Theory to Partially Automate Dialogue Tree Authoring

Charles Rich and Candace L. Sidner

Worcester Polytechnic Institute
Worcester, MA, USA
{rich,sidner}@wpi.edu

Abstract. We have developed a novel methodology combining hierarchical task networks with traditional dialogue trees that both partially automates dialogue authoring and improves the degree of dialogue structure reuse. The key to this methodology is a lightweight utterance semantics derived from collaborative discourse theory, making it a step towards dialogue generation based on cognitive models rather than manual authoring. We have implemented an open-source tool, called Disco for Games (D4g), to support the methodology and present a fully worked example of using this tool to generate a dialogue about baseball.

Keywords: cognitive models for behavior generation, conversational and story-telling agents.

1 Introduction

Although the ultimate goal of most intelligent virtual agents research is to generate dialogue interaction from a rich underlying cognitive model, many current agents, for example in video games, are still developed using traditional dialogue tree authoring technology. Dialogue trees, however, have have many problems. We have developed a novel methodology that mitigates some of these problems by combining dialogue trees with hierarchical task networks. The benefits of our methodology include easier development and maintenance of large dialogues, greater reuse of dialogue authoring investment, and partial automation of dialogue generation. Most importantly, because our methodology is based on collaborative discourse theory, it is a step along the road toward more cognitively based dialogue interaction. The methodology is implemented in an open-source tool obtainable from the authors.

To quickly ground our discussion, Fig. 1 shows an example of the kind of dialogue interaction we are concerned with. This dialogue involving baseball games and news is taken from our current research project building a relational agent for isolated older adults.[1] Other dialogue interactions in the project involve

[1] See http://www.cs.wpi.edu/~rich/always.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 327–340, 2012.

1 *Which team are you rooting for?*
2 • *Yankees.*
3 • *Red Sox.* ⇐
4 Really? But they aren't so great at winning the World Series!
5 • I bet you are a Yankees fan. ⇐
6 • Ah, but who cares? They play great ball!
7 No, I'm just joking with you.
8 • Oh. ⇐
9 • That's too bad, it would be more fun if you were!
10 *Did you catch Thursday's game?*
11 • *Yes.* ⇐
12 • *No.*
13 • *I don't want to talk about baseball anymore.*
14 What did you think of it?
15 • Awesome game! ⇐
16 • It was boring.
17 • We really needed that win.
18 Yeah, it was very intense. Great game.
19 • Yeah. ⇐
20 *I wonder how the Red Sox are doing in the standings. Should I check?*
21 • *Yes.*
22 • *No.* ⇐
23 • *I don't want to talk about baseball anymore.*
24 *Do you want to hear some recent baseball news?*
25 • *Yes.* ⇐
26 • *No.*
27 *Ok, I have several interesting stories here. The first one is about injuries. Would*
28 *you like to hear it?*
29 • *Yes.* ⇐
30 • *No.*
...
31 *Got time for another story?*
32 • *Yes.*
33 • *No.* ⇐
34 Well, that's about all. I sure like talking about baseball with you!
35 • Me, too. ⇐
36 • Let's talk again tomorrow.

Fig. 1. Example menu-based interaction (⇐ is user selection). *Italic lines* are automatically generated in the form shown in Fig. 2, with color added by the rules in Fig. 10.

1 *What is the Baseball favoriteTeam?*
10 *Shall we achieve LastGameDialogue?*
13 • *Let's stop achieving Baseball.*
20 *Shall we achieve BaseballBody by standings?*
24 *Shall we achieve BaseballBody by news?*
27 *Shall we achieve BaseballNews?*
31 *Shall we achieve BaseballNews again?*

Fig. 2. Default versions of indicated lines in Fig. 1 before color added by rules in Fig. 10

diet and exercise counseling and calendar event scheduling. Because these are goal-directed interactions, chatbot technology was not appropriate.

Notice that Fig. 1 is a menu-based interaction, so that the system needs to generate both the agent's utterances and the user's menu choices.

1.1 The Problems with Dialogue Trees

The traditional dialogue tree approach to implementing such interactions entails manually authoring a tree of all possible agent utterances and user responses. For example, Fig. 3 shows the dialogue tree that is used to generate the subdialogue in lines 4–9 of Fig. 1. Notice that only some of the lines in the dialogue tree actually appear in Fig. 1, due to the user's menu choices. Now, imagine the much larger dialogue tree required to represent the interactions resulting from all possible user choices in Fig. 1.

The main advantage of such dialogue trees is that they give the author direct and complete control over exactly what is said during the interaction. Furthermore, with the addition of typical advanced features, such as conditionals, side effects, goto's and computed fields, such dialogue trees can be quite flexible in terms of implementing the desired control flow in a particular application.

The main disadvantage of dialogue trees is that they are very labor intensive, both for initial authoring and subsequent modification and reuse. Our methodology addresses this disadvantage by partially automating dialogue generation. To preview our results, we automatically generated the semantic content of the 21 italicized lines out of the total 36 lines in Fig. 1. (Of these 21 generated lines, the 7 lines in Fig. 2 required additional manual effort to add "color" as described in Section 6).

Furthermore, dialogue trees are an unsatisfying solution from the standpoint of artificial intelligence research. In comparison, our methodology explicitly models (some of) the goals of an interaction and the meanings of (some) utterances relative to those goals.

Our methodology arose out of two important observations about dialogues such as Fig. 1. First, most dialogues are hierarchically structured collaborations, even if they include only utterances and no actions. What this means is that the overall dialogue has some goal, e.g., discussing baseball, which is decomposed

4 Really? But they aren't so great at winning the World Series!

- 5 I bet you are a Yankees fan.
  7 No, I'm just joking with you.
  - 8 Oh.
  - 9 That's too bad, it would be more fun if you were!
    Ok, from now on I'm a Yankees fan.
    - Great!
- 6 Ah, but who cares? They play great ball!

Fig. 3. Dialogue tree underlying lines 4–9 of Fig. 1, as indicated

10 LASTGAME: Did you catch {. . . }'s game?
11 • Yes
GOTO THINK
12 • No
GOTO STANDINGS
13 • I don't want to talk about baseball anymore.
GOTO . . .
14 THINK: What did you think of it?
. . .
GOTO STANDINGS
20 STANDINGS: I wonder how the {. . . } are doing in the standings. Should I check?

Fig. 4. Tags and goto's needed if a traditional dialogue tree were used to represent lines 10–20 of Fig. 1. The {. . . } indicate computed fields (see Section 6).

into subgoals (subdialogues), such as discussing the last game, checking the standings, and so on, recursively.

Second, we observed that many of the lines in a typical dialogue have to do with what you might call the "control flow" within this hierarchical structure. For example, the user's choice in lines 1–3 will control which of two introductory subdialogues they enter. Similarly, the user's choice in lines 10–13 will control whether or not they enter a subdialogue regarding the last game or whether they end the overall baseball dialogue entirely. Eventually, of course, the conversation gets down to (sub-. . .)subdialogues that consist entirely of application-specific content, such as lines 4–9.

In the traditional dialogue tree approach, both the hierarchical structure and the control flow is collapsed into the same representation together with the application-specific content. Collapsing this information together causes many problems. To start, this approach requires tags and goto's to express control flow branches and joins. For example, Fig. 4 shows the pattern of tags and goto's that would be needed in a traditional dialogue tree to represent the control flow in lines 10-20. This kind of "goto programming" is well-known to be error prone, especially when the logic is being frequently modified. Furthermore, such tangled webs of goto's make it difficult to reuse parts of the dialogue in other situations.

1.2 A Hybrid Methodology

Our solution to the problems with dialogue trees has been to evolve a hybrid methodology in which we use a hierarchical task network (HTN) to capture the high-level task (goal) structure and control flow of a large dialogue, with relatively small (sub-)dialogue trees attached at the fringe of the HTN. As we will see in detail below, making the hierarchical task structure of the dialogue explicit makes it possible to automatically generate much of the interaction shown in Fig. 1. The high-level task structure is also the part of the dialogue that most often gets reused. Furthermore, because all of the subdialogues, such as Fig. 3, are at the fringe of the HTN, there is no need for goto's—all of the subdialogues "fall through" to the control structure of the HTN.[2]

To summarize our methodology:

– We start by laying out the hierarchical goal structure of the dialogue.
– Then we formalize the goal structure and control flow as an HTN.
– Next we add application-specific subdialogues at the fringe of the HTN.
– We iteratively test and debug the hybrid representation.
– Finally, we add color to the automatically generated utterances as desired (e.g., the difference between the lines in Fig. 2 and Fig. 1).

At the end of this process we often have a high-level goal structure that can be reused in other similar applications. For example, when we recently started building a basketball dialogue, we found that we could reuse the baseball HTN structure by substituting different subdialogues at the fringe. Bickmore, Schulman and Sidner [2] also experienced a high degree of reuse in applying a version of this methodology to health dialogues.

This methodology is supported by a tool, called Disco for Games (D4g), which is an extension of Disco, the open-source successor to Collagen [12, 13]. In the remainder of this paper, we explain in detail how each line in Fig. 1 is generated by D4g. First, Section 2 describes Disco's HTN formalism and how we have extended it in D4g to add dialogue trees at the fringe. Next, Section 3 describes Disco's lightweight utterance semantics, based on collaborative discourse theory, which is the key to the automatic generation of both agent utterances and user menus. Section 4 describes a small set of general rules that account for all of the automatically generated dialogue in the example (Section 5). Finally, Section 6 describes how application-specific formatting rules are used to add color.

2 Hierarchical Task Networks

Fig. 5 shows a diagrammatic summary of the HTN and dialogue tree hybrid structure underlying the interaction in Fig. 1. In this diagram, we follow the common simplifying convention of omitting nodes when a task has only a single recipe (decomposition) or a recipe (decomposition) has only a single step.

For the executable representation of HTN's we use the ANSI/CEA-2018 standard [10], on which Disco is based. Lines 1–17 of Fig. 6 show the XML syntax in ANSI/CEA-2018 for specifying the `Baseball` task, which is the toplevel goal of the dialogue, and its four steps (subgoals): `intro`, `lastGame`, `body` and `closing`. HTNs in this formalism include tasks with named and typed inputs, such as `favoriteTeam` (a `Team`) and `lastDay` (a `Day`) and outputs, and one or more recipes (`<subtasks>`)that decompose each non-primitive task into a sequence of (possibly optional or repeated) steps. Optionality and repetition are expressed together by the `minOccurs` and `maxOccurs` attributes of a step, both of which default to 1.

[2] Readers with a knowledge of the history of programming languages will recognize this as analogous to the argument for "structured programming" over goto's.

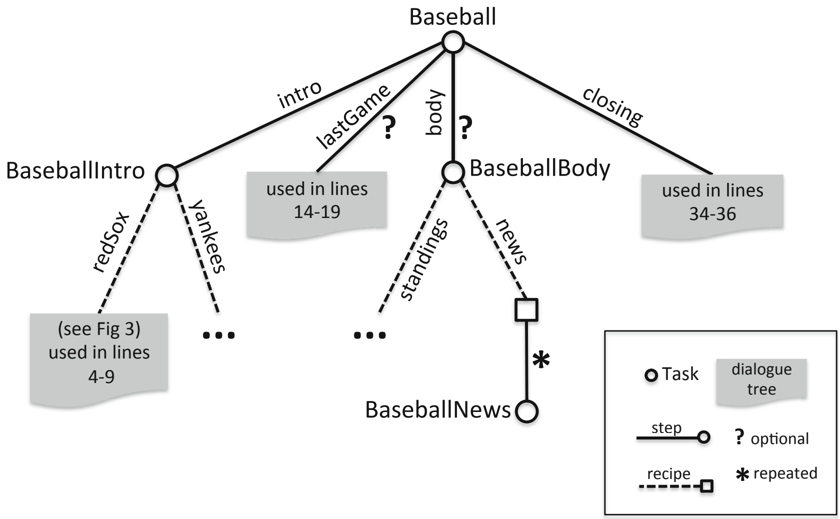


Fig. 5. Hierarchical task network underlying example interaction in Fig. 1

```
<task id="Baseball">
 <input name="favoriteTeam" type="Team"/>
 <input name="lastDay" type="Day"/>
 <subtasks id="talk">
  <step name="intro" task="BaseballIntro"/>
  <step name="lastGame" task="LastGameDialogue" minOccurs="0"/>
  <step name="body" task="BaseballBody" minOccurs="0"/>
  <step name="closing" task="ClosingDialogue"/>
 </subtasks>
</task>

<task id="BaseballIntro">
 <subtasks id="redSox">
  <step name="intro" task="RedSoxIntroDialogue"/>
  <applicable> $Baseball.favoriteTeam==Team.ENUM.redSox </applicable>
 </subtasks>
 ...
</task>

<agent id="RedSoxIntroDialogue" text="Really? But they aren't...">
 <user text="I bet you are a Yankees fan.">
  <agent text="No, I'm just joking with you.">
   <user text="Oh."/>
   <user text="That's too bad, it would be more fun if you were!">
    <agent text="Ok, from now on I'm a Yankees fan.">
     <user text="Great!"/></agent></user></agent></user>
 <user text="Ah, but who cares? They play great ball!"/>
</agent>
```

Fig. 6. Part of ANSI/CEA-2018 and Disco for Games (D4g) specification of Fig. 5

ANSI/CEA-2018 also supports the use of JavaScript to specify preconditions and postconditions of tasks and the applicability conditions of recipes. All of these conditions use a three-valued logic, where the JavaScript null value represents unknown. For example, the `<applicable>` element on line 14 selects the appropriate introductory subdialogue when the user's favorite team is the Red Sox.

Lines 18–26 of Fig. 6 are a straightforward XML encoding of the dialogue tree in Fig. 3, which is the leftmost dialogue tree on the fringe of the HTN in Fig. 5. The syntax used in these lines is transformed by D4g's XSLT preprocessor into appropriate ANSI/CEA-2018 specifications that cause the structure of the dialogue tree to unfold properly when executed in Disco. Thus, from a D4g author's point of view, both the HTN and the dialogue tree portions of the specification can be conveniently intermixed in a single file.[3]

3 Utterance Semantics

The key to automatically generating dialogue from the HTN portion of our hybrid representation is a lightweight semantics for dialogue utterances derived from Sidner's artificial language for negotiation [15] based on collaborative discourse theory [6, 8].

Collaborative discourse[4] theory is fundamentally an *interpretation* theory. It views dialogue as being governed by a hierarchy of tasks (goals) and a stack-like focus of attention and explains how to interpret an utterance (by either participant) as contributing to or changing the current task. Three fundamental ways that an utterance can contribute to a task are to:

1. provide a needed input (`Propose.What`)
2. select the task or a subtask to work on (`Propose.Should`), or
3. select a recipe to achieve the task (`Propose.How`).

In Disco, these three fundamental types of contribution are formalized, respectively, in the semantics of the first three builtin utterance types shown in Fig. 7 along with their default formatting. The semantics of these utterances also includes Sidner's model of the negotiation of mutual beliefs via proposal, acceptance and rejection [15]. Understanding these semantics is very important for the dialogue designer because, as we will see in the next section, they are the link between the HTN structure and the automatically generated dialogue utterances.

For example, when a dialogue participant utters a `Propose.What`, it means (in part) that the speaker:

– believes the proposition that the *input* to the *task* is *value* and
– intends that the hearer believe the same thing.

[3] D4g also supports transferring control to an HTN from inside a dialogue tree, so that HTN's and dialogue trees can in fact alternate in layers. However, we do not use this feature very often.

[4] In this work we consider only two-participant discourse, i.e., dialogue.

Propose.What(*task*, *input*, *value*)	*The task input is value.*
Propose.Should(*task*)	*Let's achieve task.*
Propose.How(*task*, *recipe*)	*Let's achieve task by recipe.*
Accept(*proposal*)	*Yes.*
Reject(*proposal*)	*No.*
Ask.What(*task*, *input*)	*What is the task input?*
Ask.Should(*task*)	*Shall we achieve task?*
Ask.How(*task, recipe*)	*Shall we achieve task by recipe?*
Ask.How(*task*)	*How shall we achieve task?*
Propose.Stop(*task*)	*Let's stop achieving task.*
Ask.Should.Repeat(*task*)	*Shall we achieve task again?*

Fig. 7. The main builtin Disco utterance types and their default formatting

Similarly, uttering a Propose.Should proposes a task or subtask, such as an optional step, to work on. Uttering a Propose.How proposes a recipe to use. If the hearer Accept's one of these proposals, then mutual belief in the respective proposition is achieved. The hearer can also Reject a proposal.

Utterances can also be questions. In Sidner's framework, questions are modeled as proposals by the speaker that the hearer provide information. For example, when a dialogue participant utters an Ask.What (see Fig. 7), it means (in part) that the speaker intends that the hearer respond by uttering a Propose.What that provides a *value* for the specified *task* and *input*. The other three utterance types starting with Ask have analogous semantics.

Disco implements these utterance semantics in its dialogue interpretation algorithm. Basically, to interpret a new utterance, the algorithm visits every live[5] task node in the HTN tree and asks the question, "Could this new utterance contribute to this task?" If the answer is yes, Disco attaches the new utterance as a child of the task node; otherwise it marks the utterance as "unexplained." (For more details about Disco's dialogue interpretation algorithm see [12].)

4 Generation Rules

Disco treats dialogue *generation* as the algorithmic inverse of interpretation. In other words, Disco visits every live task node in the current HTN tree and asks the question, "What are the possible utterances that *could* contribute to this task?" The answers to this question are the generation candidates.

Fig. 8 shows the overall functional flow of dialogue generation in Disco in more detail. Starting on the left, the first step is to apply the general generation rules described in this section to the current dialogue state, yielding a set of candidate utterances for either the agent or the user (depending on whose turn it is in the interaction). Each generation rule produces zero or more candidate utterances. These candidate utterances are then sorted according to heuristic priorities (see

[5] A task is live if and only if its precondition is not false and all of its predecessor steps, recursively up the tree, have been successfully completed.

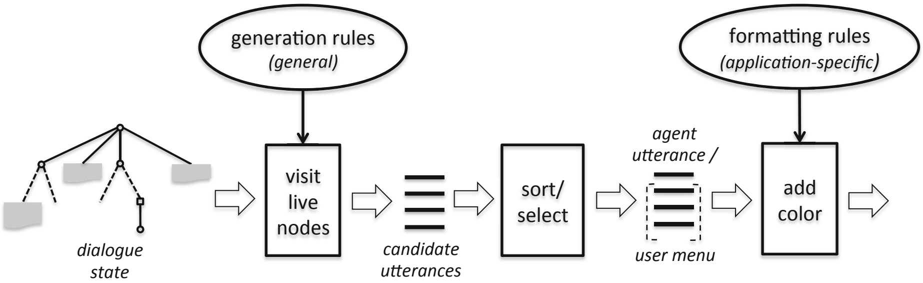


Fig. 8. Functional flow of dialogue generation in Disco

```
define AskWhatRule (task)
  foreach input in inputs(task)
    if input does not have a value
      then return { new Ask.What(task, input) }
  return {}
```

Fig. 9. Pseudocode for `Ask.What` generation rule (applied to live task nodes)

below). If utterances are being generated for the agent, then the highest priority candidate is chosen for the agent to speak; if utterances are being generated for the user menu, then the (perhaps truncated) candidate list is used to populate the user menu. Finally (see Section 6), optional application-specific formatting rules are applied to add color to some utterances, as desired.

Twelve general rules generate the content of all the lines in Fig. 1. These same rules are used in all of the dialogue applications we have built so far. There is one rule for each of the eleven utterance types in Fig. 7, plus an additional rule that generates the appropriate agent utterance or user menu entries from a dialogue tree on the fringe, when it is live.

Each generation rule is implemented as a small Java method that is applied by the generation algorithm to each live task node in the HTN tree as described above. For example, Fig. 9 shows pseudocode for the rule that generates `Ask.What` utterances. Notice that this rule will not return a question for a particular input if that input already has a value (which could happen either via dialogue or some other system process). In general, the generation rules only return candidate utterances when the relevant information is not already known.

Some rules behave differently depending on whether candidates are being generated for the agent or the user. For example, when the `Propose.What` rule is being executed for the user and the input type is an enumerated datatype, the rule returns a set of utterances (menu choices) that includes a `Propose.What` for each possible value.

Currently, because our agent is always subordinate to the user, the rules for `Accept` and `Reject` only generate candidates for the user menu. The inputs to these rules are task nodes that are dynamically added to the dialogue state whenever the agent makes a proposal (including questions). Notice that the default formatting of Accept and Reject is simply "Yes" or "No," but that the underlying semantics includes the proposal being accepted or rejected.

The heuristic priorities used to sort candidate utterances have the most impact in agent generation, since only the topmost candidate is actually uttered by the agent. (In the case of the user, the priorities only affect the order in which the menu choices are displayed.) Currently, each of the twelve general rules has a fixed priority—we do not tweak priorities for a particular dialogue. Although these priorities are not yet grounded in any cognitive theory, they do follow a logical order of design decisions. For example, the Ask.What rule has a higher priority than the Ask.How rule, since the properties of the input to a task may affect the best recipe choice.

5 The Example Revisited

We can now revisit the example in Fig. 1 and explain how all of the italicized automatically generated lines are produced (at least in uncolored forms shown in Fig. 2). Starting with the agent utterances:

- Line 1 is generated by the application of the Ask.What rule to the favoriteTeam input of the Baseball task (see Fig. 6, line 2). Notice that no agent question is generated for the lastDay input of Baseball, because this input has already been bound as part of the system initialization.
- Line 10 is generated by the application of the Ask.Should rule to the LastGameDialogue step of Baseball (see Fig. 6, line 6).
- Line 13 is generated by the Propose.Stop rule, which only returns an utterance when applied to the toplevel goal of the dialogue (Baseball in this case). This rule provides the user with a convenient menu option for exiting the whole dialogue. This rule has some internal heuristics for when to return an utterance, depending on the dialogue state, so that the exit option is not always offered.[6]
- Lines 20 and 24 are generated by the application of the Ask.How rule to the standings and news recipes for BaseballBody (see Fig. 5).
- Line 27 is generated by the application of the Ask.Should rule to BaseballNews (see Fig. 5).
- Line 31 is an Ask.Should.Repeat, which is a variant of Ask.Should that is generated whenever the task being proposed is the second or subsequent instance of a repeated step, such as BaseballNews. The reason for this variant is to facilitate attaching a different formatting rule, as we will see in the next section.

Regarding the automatically generated user menu choices in Fig. 1:

- All of the Yes and No menu choices are generated by the Accept and Reject rules.

[6] Our agent currently automatically accepts all proposals by the user.

– Lines 2 and 3 are generated by the application of the `Propose.What` rule to the `favoriteTeam` input of the `Baseball` task. As mentioned above, this rule checks for the special case of enumerated datatypes, in which case it generates a `Propose.What` for each possible value. Furthermore, the default formatting for these utterances is simply the printable string for the data value. Enumerated datatypes in Disco are declared as JavaScript objects with an `ENUM` field (see Fig. 6, line 14).

In summary, we have now seen how all of the content in the example interaction is generated by the application of the general rules described above to the dialogue structure in Fig. 5. In the next section, we will see how the differences between the lines in Fig. 1 and in Fig. 2 are achieved.

6 Adding Color with Formatting Rules

The main reason why authors like dialogue trees is that it allows them to creatively tailor their use of language to the character and the narrative context—what we call "adding color" to the dialogue. In other words, the dialogues don't read like they were generated by a computer.

In the methodology we have evolved, we have found it best to postpone adding color until late in the authoring process. First, we develop and debug the HTN, such as Fig. 5, that represents the goal hierarchy and desired control flow between the fringe subdialogues. At the end of this phase, we have a working interaction that looks like Fig. 1, except with the corresponding lines from Fig. 2 appearing instead. Then we add color as desired via formatting rules. In this example, there were seven lines that needed color.

Formatting rules in Disco are specified in a Java properties file, which is organized as one key/value pair per line with an equal sign separating the key from the value, as shown in Fig. 10. Each key ends in `@format` with a prefix describing the type of utterance to which it is to be applied. For example, the rule in line 1 of Fig. 10 applies to all occurrences of `Ask.What` in which the *task* is `Baseball` and the *input* is `favoriteTeam`. The rule in line 10 applies to all occurrences of `Ask.Should` in which the *task* is `LastGameDialogue`, and so on. When a rule is applied, the value part of the rule is substituted for the default formatting of the corresponding utterance.

```
 1 Ask.What(Baseball,favoriteTeam)@format = Which team... rooting for?
10 Ask.Should(LastGameDialogue)@format = ...{$Baseball.lastDay}'s game?
13 Propose.Stop(Baseball)@format = I don't... about baseball anymore.
20 Ask.How(BaseballBody,standings)@format = I wonder... Should I check?
24 Ask.How(BaseballBody,news)@format = Do you want to hear... news?
27 Ask.Should(BaseballNews)@format = Ok, I have several interesting...
31 Ask.Should.Repeat(BaseballNews)@format = Got time for another story?
```

Fig. 10. Formatting rules applied to indicated lines in Fig. 2

The rule on line 10 of Fig. 10 illustrates the use of computed fields, which is an important feature of the formatting system (that is also available for utterances that appear in dialogue tree). Curly brackets {...} in an utterance enclose arbitrary JavaScript code that is executed during the final formatting process to compute a string to substitute at that point in the utterance. In line 10, this feature is used to retrieve the value of the `lastDay` input of the most recently created instance of `Baseball`.

A commonly used special case of computed fields that is supported in Disco formatting rules (but not illustrated in this example) is using a vertical bar `|` to separate a set of alternative variations. For example, if `Accept@format` were set to `Ok|Yup|Sure`, the formatting system would systematically use these variations instead of all the Yes's in Fig. 1.

7 Related Work

Both HTNs [5] and dialogue trees [4] are very well known and commonly used techniques. HTNs have been used by others, such as Bohus and Rudnicky [3] or Smith, Cavazza et al. [16], for generating dialogue, but the goal of these efforts has been to eliminate dialogue trees, rather than coexist with them, as we have. Orkin et al. [9] have tried to eliminate manual dialogue authoring by applying data mining techniques to crowdsourced data to automatically create HTNs, which are then used to generate new dialogues.

This paper has grown out of our own previous work in several ways. In [11], we first described the idea of discourse generation as the algorithmic inverse of discourse interpretation and introduced the model of applying rules (called "plugins" in that paper) to the live nodes of an HTN to generate dialogue candidates, as shown in the first step of Fig. 8. We first described D4g in [7], although the emphasis in that work was on combining actions and utterances in a single representation, whereas this paper concerns itself entirely with utterances. The DTask system [1, 2], also an extension of ANSI/CEA-2018 and Disco, used HTNs with adjacency pairs (a single agent utterance with a user response menu) at the fringe, instead of complete dialogue trees, as in D4g. That work also explored the reuse advantages of HTNs in dialogue.

8 Conclusion

We have demonstrated, using an example baseball dialogue, how combining hierarchical task networks with dialogue trees greatly improves the authoring process as compared to using dialogue trees alone. Our methodology is supported by open-source tool, called Disco for Games (D4g), that is available by contacting the authors.

We have used D4g to author similar dialogues on other topics, such as family and weather, with similar positive experiences in terms of the number of automatically generated lines. (We do not quote the fraction of automatically generated lines here as a statistic, because this number is easily skewed by the number of lines in the subdialogues at the fringe of the HTN.)

Looking toward the future, we see D4g as a step along the road toward totally automatic generation of dialogue. We expect to continue to extend the set of semantically specified utterance types (the current set is already in fact larger than shown in Fig. 7), which along with additional generation rules, will increase the amount of automatically generated content in dialogues. For example, we are interested in revisiting the automatic generation of tutorial dialogues, as we did in [14].

Acknowledgements. We would like to thank Fred Clinckemaillie and JJ Liu for their help with the baseball dialogue.

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0811942 and 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.

References

1. 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)
2. Bickmore, T., Schulman, D., Sidner, C.: A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. J. Biomedical Informatics 44, 183–197 (2011)
3. Bohus, D., Rudnicky, A.: The RavenClaw dialog management framework: Architecture and systems. Computer Speech and Language 23(3), 332–361 (2009)
4. Despain, W.: Writing for Video Games: From FPS to RPG. A. K. Peters (2008)
5. Erol, K., Hendler, J., Nau, D.: HTN planning: Complexity and expressivity. In: Proc. 12th National Conf. on Artificial Intelligence, Seattle, WA (July 1994)
6. Grosz, B.J., Sidner, C.L.: Plans for discourse. In: Cohen, P.R., Morgan, J.L., Pollack, M.E. (eds.) Intentions and Communication, pp. 417–444. MIT Press, Cambridge (1990)
7. Hanson, P., Rich, C.: A non-modal approach to integrating dialogue and action. In: Proc. 6th AAAI Artificial Intelligence and Interactive Digital Entertainment Conf., Palo Alto, CA (October 2010)
8. Lochbaum, K.E.: A collaborative planning model of intentional structure. Computational Linguistics 24(4), 525–572 (1998)
9. Orkin, J., Smith, T., Roy, D.: Behavior compilation for ai in games. In: Proc. 6th AAAI Artificial Intelligence and Interactive Digital Entertainment Conf., Palo Alto, CA, pp. 162–167 (October 2010)
10. Rich, C.: Building task-based user interfaces with ANSI/CEA-2018. IEEE Computer 42(8), 20–27 (2009)
11. Rich, C., Lesh, N., Rickel, J., Garland, A.: A plug-in architecture for generating collaborative agent responses. In: Proc. 1st Int. J. Conf. on Autonomous Agents and Multiagent Systems, Bologna, Italy (July 2002)

12. Rich, C., Sidner, C.: Collagen: A collaboration manager for software interface agents. User Modeling and User-Adapted Interaction 8(3/4), 315–350 (1998); reprinted in Haller, S., McRoy, S., Kobsa, A. (eds.): Computational Models of Mixed-Initiative Interaction, pp. 149–184. Kluwer Academic, Norwell (1999)
13. Rich, C., Sidner, C., Lesh, N.: Collagen: Applying collaborative discourse theory to human-computer interaction. AI Magazine 22(4), 15–25 (2001)
14. Rickel, J., Lesh, N., Rich, C., Sidner, C.L., Gertner, A.S.: Collaborative Discourse Theory as a Foundation for Tutorial Dialogue. In: Cerri, S.A., Gouardéres, G., Paraguaçu, F. (eds.) ITS 2002. LNCS, vol. 2363, pp. 542–551. Springer, Heidelberg (2002)
15. Sidner, C.L.: An artificial discourse language for collaborative negotiation. In: Proc. 12th Nat. Conf. on Artificial Intelligence, Seattle, WA, pp. 814–819 (August 1994)
16. Smith, C., Cavazza, M., Charlton, D., Zhang, L., Turunen, M., Hakulinen, J.: Integrating Planning and Dialogue in a Lifestyle Agent. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 146–153. Springer, Heidelberg (2008)

Authoring Rules for Bodily Interaction: From Example Clips to Continuous Motions

Klaus Förger, Tapio Takala, and Roberto Pugliese

Department of Media Technology,
School of Science,
Aalto University,
Espoo, Finland
{klaus.forger,tapio.takala,roberto.pugliese}@aalto.fi

Abstract. We explore motion capture as a means for generating expressive bodily interaction between humans and virtual characters. Recorded interactions between humans are used as examples from which rules are formed that control reactions of a virtual character to human actions. The author of the rules selects segments considered important and features that best describe the desired interaction. These features are motion descriptors that can be calculated in real-time such as quantity of motion or distance between the interacting characters. The rules are authored as mappings from observed descriptors of a human to the desired descriptors of the responding virtual character. Our method enables a straightforward process of authoring continuous and natural interaction. It can be used in games and interactive animations to produce dramatic and emotional effects. Our approach requires less example motions than previous machine learning methods and enables manual editing of the produced interaction rules.

Keywords: animation, motion capture, bodily interaction, continuous interaction, authoring behavior.

1 Introduction

Virtual characters are common in modern games and their bodily motions can reflect the emotions and attitudes between characters. Real-time motion synthesis and synchronization with external stimuli enables making the characters interactive. The possibilities for using bodily interaction have increased as even consumer level sensor technology allows capturing bodily motions. Sometimes a motion does not mean much out of context, but can have a lot of meaning if displayed as a synchronized reaction to an action [1]. A good example is a virtual character taking a step backwards in isolation versus taking a step backwards as a reaction to aggressive behavior of another character.

Expressive motion based interaction between two virtual characters is possible with a library of recorded and annotated motions. For example, if a character moves in a way that was annotated as angry, then another character could react

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 341–354, 2012.

by selecting a motion that was annotated as scared. Similar interaction between a human and a virtual character requires ability to evaluate the style of previously unseen motions in real-time. Only part of the emotional content in a motion library can be annotated in advance as it can vary depending on the context of the motion.

In this paper we explore how motion captured examples of human interaction can be used in authoring interaction rules for virtual characters. These rules allow real-time generation of expressive behaviours in a continuous interaction loop. The idea is that there should be no frozen pauses during an interaction sequence as can happen in task-based approaches, but instead all idle moments could be used to reflect the attitudes of the participants. Continuous interaction scheme could be also used for subtle control over the style of motion. For example a walking character could immediately react to an observed aggressive action by changing the style of the walk from neutral to careful. The amount of visible aggression could be continuously mapped to the amount of carefulness.

We concentrate on the case where humans and virtual characters have equal amount of information from each other. The virtual characters observe humans through features that characterize different qualities of human motion. The features, from now on referred as motion descriptors, can be calculated in real-time. We show how using example motion pairs makes authoring of interaction rules a straightforward process. The example motions also help avoiding impossible combinations of motions descriptors. Moreover, we show that selective use of motion descriptors allows solving the curse of dimensionality, that arises from modeling human motion simultaneously with several descriptors.

We next present related work, and then describe our implementation in three parts. First is the calculation of motion descriptors. The second is the interaction rule authoring where observed descriptors are mapped to descriptors of desired reactive behaviours. The last part is motion synthesis that turns the descriptors to actual motions of a virtual character. In the fourth section, we present a use case of the process of authoring behaviours. The last sections contain discussion, conclusions and future work.

2 Related Work

Real-time motion synthesis can be done by creating a motion graph from captured motions and playing one motion segment after another according to the graph [2]. Furthermore, if the captured motions are annotated, it is possible to control the motion synthesis [3]. This can happen by selecting motion segments from the graph that correspond to attributes used in the annotation. We use motion synthesis that is based on a motion graph. Our graph includes information about the motion style that is used in controlling the motion synthesis. Therefore, it is similar to the metadata motion graphs, which have been used for synchronizing human motion with beats in streaming music [4].

Manual annotation of motion can be very time consuming. Therefore automatically calculated motion descriptors for human motion have been developed

[5,6]. The descriptors can measure for example the amount of motion, acceleration or qualities of the pose of a character. Similar values have been also calculated from the relational motion of hands representing two entities such as small animals [7]. We use motion descriptors for annotation of recorded motion in a motion graph and real-time motions. We also extend the relative descriptors from relations between hands to the case of relations between two human characters.

Earlier systems that allowed interacting with virtual creatures were often targeted at goal oriented interaction [8] or interaction scripted with if-else clauses [9]. A more fluid model of interaction was allowed by a probabilistic method that uses pairs of recorded actions and reactions to learn how to react to human movements [10]. A similar system based on example motions has been used for teaching cleaning robots socially acceptable motion styles [11]. Our method for defining interaction fits in between these older methods as it takes advantage of example action-reaction motions and manual definitions.

The importance of usability has been noted in earlier works that present tools for authoring behaviours of virtual creatures [12,13]. These tools assume that a range of low level behaviours such as wandering, following and actions that display emotional states are available. The tools allow applying the low level behaviours to crowds and joining them together to form more complex patterns. Our method could be used for authoring the said low level behaviours. One requirement for the earlier tools has been that using them should not require coding experience or understanding complex models that govern the behaviours. This requirement is taken into account in the design of our method.

Publications related to virtual characters include works on embodied conversational agents (ECAs) [1]. Considerable effort has been taken towards a unified Behaviour Markup Language (BML) that can be used when creating ECAs [14]. These works mainly view bodily motion as a way to make verbal conversations more believable. In this paper, we consider varied situations beyond conversations, and study non-verbal interaction where bodily motion is the only channel of communication. We also extend the scope of behaviours from friendly and believable characters to ones that could be considered anti-social and annoying as those can be required in for example games containing dramatic sequences.

A proposal has been made to extend the BML from describing the behavior of a single virtual human to the case of continuous interaction between two characters [15]. Continuous interaction is a core aspect of our paper, but we have a different point of view on what part of motion we want to control. The proposal suggests developing an XML based approach that could be useful in defining and controlling discrete reactions and gestures. However, we concentrate on motion style that we abstract with motion descriptors which have a continuous range of values that vary from frame to frame. For this reason we use a more continuous control mechanism.

Our work builds on earlier work about defining interaction rules using motion descriptors [16]. The earlier system showed that Radial Basis Functions (RBF) [17] can be used to map the input motion descriptors to output descriptors. The output descriptors where then used to control a motion synthesis engine. These parts created a virtual character that reacts to observed human motions. The system used only two input and output descriptors and therefore had limited capability to create interesting interaction. In this paper, we show that using more than two descriptors allows much more varied interaction. However, it also introduces the curse of dimensionality as the number of the combinations of the descriptors grows exponentially compared to the number used descriptors. That in turn forced us to find new ways to create the interaction rules.

3 Implementation

Our system is based on an interaction loop where two characters can observe each other and react to the actions they witness. The steps from the observed motions to the synthesized motion are shown in Figure 1. First, the values of the input descriptors are calculated from the observed motion of another character (or human) and from the character's own motion. Secondly, the input descriptor values are mapped according to the interaction rules into the desired output descriptor values. Then the motion synthesis engine creates a motion that fits the output descriptor values as well as possible. Next, the newly synthesized motion can be observed by another virtual character or by a human.

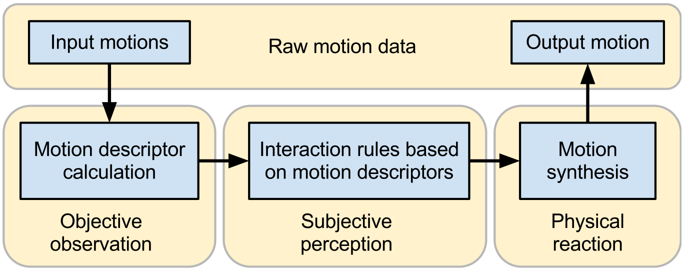


Fig. 1. Model of tasks performed by an interactive virtual character system and the flow of information in the system

3.1 Motion Descriptors

In our implementation, a motion descriptor is a function that takes as input the observed 3D motion data and calculates a value between zero and one for each frame of the motion. Ideally, the descriptors would tell all about the motion style of an action and ignore unimportant aspects. In practice, the descriptors are limited to what can be easily defined mathematically and calculated in real-time. The descriptors act as objective measures that are not affected by any internal states of the characters.

In this work we concentrate on behaviours that include standing, walking, jumping and generally moving around on a flat floor. For these types of motions, relevant motion descriptors for an isolated character include Quantity of

Motion (QoM) [6], turning left/right and moving backwards/forwards. Examples of motions corresponding to high and low values of the descriptors are shown in Figure 2. The QoM estimates the total amount of energy in the motion and is calculated as the sum of instantaneous velocities of all body parts. We also use a variant of QoM called non-transitional QoM (NtQoM) that estimates the energy used only for body language or other expressive motions, disregarding locomotion.

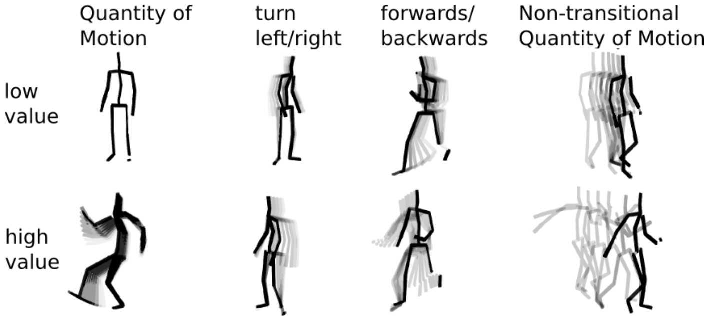


Fig. 2. Examples of motion descriptors calculated from an isolated character

Another class of descriptors are relational descriptors, i.e. those that compare motion of two characters. Of these we use the distance between the characters, their facing angle, and approach/retreat as illustrated in Figure 3. The facing angle is normalized to be zero for face-to-face characters and one when a character's back is turned towards the other. The values of approach/retreat are linearly relational to the velocity along the direction to the other character. The extreme values of approach/retreat are normalized in order to get zero when the character is moving towards the other character at 7 m/s and one when moving away at the same speed. The value 7 m/s is reasonable limit for bodily interaction as a forceful jump without a run can have approximately that velocity. All the other descriptors are also normalized in a similar manner. For the distance there is only one common value for both characters, but the characters have their own values for the facing angle and approach/retreat.

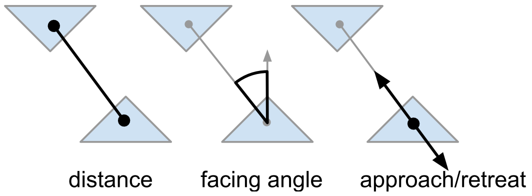


Fig. 3. Relational motion descriptors between two characters

We used a commercial motion tracking system (NaturalPoint OptiTrack) with 24 cameras working at 100 frames per second to capture the motions from which we calculated the descriptors. We were also able to calculate all the used descriptors from the data Microsoft Kinect provides using OpenNI library. However, the lower accuracy and higher amount of errors in the Kinect data made using it impractical. Especially calculating QoM from the noisy data is unreliable.

While the motion descriptors estimate different aspects of the motion, they are not independent from each other. Some dependencies are direct. For example, if QoM is zero, it limits all other descriptor values to those corresponding to no movement. Another type of dependency is dynamic. For example, between QoM and facing angle all value combinations are physically possible, but changing facing angle becomes impossible if the QoM remains zero. Because of these physical constraints, it is useful to pick descriptor values from recorded motions. A randomly selected set of descriptor values might be impossible to synthesize into motion of a virtual character.

3.2 Interaction Rule Authoring

The interaction rules define the reactions of a virtual character to observed motions. The reactions can vary depending on the own position and motion of the virtual character. A rule can reflect physical properties such as being strong or weak. Also, the mental state of the character, such as sadness or aggressiveness, can be built in the behavior produced with a rule.

In practice, the rules are used for projecting a frame of an observed motion to a desired reaction. Here we consider the frame to include positional information and instantaneous velocities of body parts. Our method uses motion descriptors to abstract the frames of input and output motions. In an earlier publication the mappings were created with manually defined example point pairs [16]. A mapping consisted of one point in the input corresponding to one point in the output space. After the mappings for the rule were defined, projecting a point from the input to the output was done with Radial Basis Functions (RBF), that is a sparse data interpolation method [17]. This process of creating a rule required filling the input space evenly with points along every dimension.

The standard RBFs approximate a function from a high dimensional space to a single dimension using example points were the output value is predefined [17]. In practice, if the point that is being projected is close to only one of the example points, the output will have the value that is linked to that example point. Should the point be in the middle of two example points, the output would be an average of the linked output values weighted by the inverse of distances to those example points. The case of projecting from a high dimensional space to another high dimensional space with RBFs requires only repeating the standard case for each of the output dimensions [17]. The RBFs were chosen as the interpolation method as it is a simple approach to code and cheap to calculate.

The projection using point pairs and RBFs works well up to two descriptor dimensions, but faces the curse of dimensionality in the combinatorial sense with a higher number of dimensions. This means that the amount of point pairs required to cover the input space grows exponentially with the number of dimensions. High dimensionality also hinders visualizing and editing points as three spatial dimensions is the limit on human vision. Growing dimensionality also makes motion synthesis harder as each descriptor dimension sets new requirements for the produced motions.

When experimenting with high dimensional interaction rules, we observed that creating the mappings rarely requires using more than three dimensions simultaneously. In fact, many interesting and useful mappings require only using one or two dimensions, but the set of required dimensions varies between mappings. This observation lead us to the conclusion that the problems caused by a high dimensionality could be solved by allowing the author of the rules to select which dimensions are relevant to each of the mappings individually.

Mathematically, our solution requires pairing each set of descriptor values with a scaling vector and a modification to the projection done with RBFs. The scaling vector indicates how important the related descriptor dimensions are. The input data we need to consider includes the point we want to project p and the example point pairs indexed as $[1...k...K]$. A point pair consists of an input point i_k and the related input scaling vector s_k, an output point g_k and the related output scaling vector u_k. The data we want to calculate is the output descriptor values o and the output scaling vector h. During real-time interaction p is the observed motion, o is the desired reaction and h tells how the output descriptors should be prioritized in the motion synthesis.

Next, we go through the changes needed in the standard case of RBF. Let us consider an input space with dimensions indexed as $[1...n...N]$ and an output space with dimensions indexed as $[1...m...M]$. The scaling vector s_k of an example input point i_k affects the calculation of the distance d_k between the point p, that is being projected, and the point i_k as follows:

$$d_k = \left| (i_k - p) \begin{bmatrix} s_{k_1} & 0 & \dots & 0 \\ 0 & s_{k_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_{k_N} \end{bmatrix} \right| . \tag{1}$$

If the scaling vector s_k is all ones, then the distance calculation returns to the unmodified case. If one of the values in the scaling vector is zero, then that input descriptor dimension is effectively ignored by the mapping. The output scaling vectors $u_{1...K}$, which are paired with example output descriptor values $g_{1...K}$, enable the creation of mappings that can be independent also on the output side of the projection. In practice, this means that the influence of a mapping (a point pair) can be limited to only part of the output descriptors $o_{1...M}$. The values for the example output scaling vectors $u_{1...K}$ and the example output points $g_{1...K}$ of the mappings are used in the RBF interpolation resulting in the output descriptor values $o_{1...M}$ as follows:

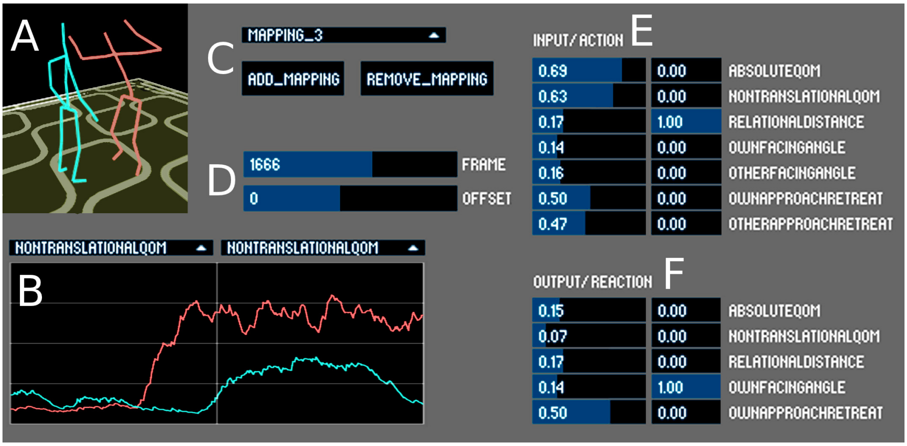
$$o_m = \frac{\sum_k (g_{k_m} \cdot u_{k_m} \cdot (1 - d_k))}{\sum_k (u_{k_m} \cdot (1 - d_k))} . \tag{2}$$

Finally, the values $h_{1...M}$ of output scaling vector that forms a pair with output descriptor values $o_{1...M}$ is:

$$h_m = \sum_k (u_{k_m} \cdot (1 - d_k)) \tag{3}$$

As the values for descriptors and the scaling vectors are limited to the range from zero to one we also limit all the values in the calculations to the same range. The o and the h are given as input to the motion synthesis engine. The descriptor values o control type and style of synthesized motion and scaling vector h tells how the output descriptors should be prioritized and which can be ignored.

The practical work that the author of the rules must do when creating a mapping for a rule includes defining the values for input and output descriptors and scaling vectors. This can be done manually with the sliders of a graphical user interface shown in Figure 4 (E, F). However, we have found that it is not always easy to see which movement would correspond to given descriptor values or vice versa. This problem can be overcome by capturing an example action and reaction and selecting the descriptor values from them with the user interface. The time line/frame counter (fig. 4 D, upper slider) can be used to simultaneously browse through the values (fig. 4 B, E, F) and view an animation (fig. 4 A) of the example motions.



1666	FRAME
0	OFFSET

INPUT/ACTION

0.69	0.00	ABSOLUTEQOM
0.63	0.00	NONTRANSLATIONALQOM
0.17	1.00	RELATIONALDISTANCE
0.14	0.00	OWNFACINGANGLE
0.16	0.00	OTHERFACINGANGLE
0.50	0.00	OWNAPPROACHRETREAT
0.47	0.00	OTHERAPPROACHRETREAT

OUTPUT/REACTION

0.15	0.00	ABSOLUTEQOM
0.07	0.00	NONTRANSLATIONALQOM
0.17	0.00	RELATIONALDISTANCE
0.14	1.00	OWNFACINGANGLE
0.50	0.00	OWNAPPROACHRETREAT

Fig. 4. A graphical user interface for authoring interaction rules that includes animation of the example motions (A), view of descriptor values over time (B), selection of mappings (C), sliders for the animation time and offset between motions (D), input/output descriptors (usually picked from animation) (left side E, F) and input/output scaling (manually defined) (right side E, F)

In example motions the action and the reaction do not always happen at exactly the same moment. Therefore, there can be a need to scroll the reaction forward in order to find the right descriptor values for it. This can be done with the lower slider in the user interface in Figure 4 (D). This offsetting is necessary especially for the relational descriptors that have only one value such as the distance between the characters. Without offsetting, those descriptors would by definition have the same value for both input and output.

Picking descriptor values from example motions helps the process of authoring rules, but picking the values for scaling vectors cannot be done in the same way especially in a high dimensional case. If more than one example pair of motions displaying the same interaction would be available, then some of the scaling values could be estimated based on the correlations in the examples. However, this would add much work in capturing the examples. For this reason the author of the rules needs to have a vision of the intended interaction that guides the selection of the scaling values.

Compared with the old method of creating rules [16] the new method allows rules to be made with much less mappings as the scaling can be used to ignore those descriptor dimensions which are not relevant. When using the old method, it would have been necessary to define mappings for all the combinations of descriptor dimensions, even those that should not affect the end result.

3.3 Motion Synthesis

During real-time interaction, the motion synthesis engine takes the desired output descriptor values and creates a continuous motion following the values as closely possible. We use a motion graph based on recorded motions for the synthesis. All the motions in the motion graph are annotated using the motion descriptors. After this we can synthesize new motions by concatenating motion clips that fit well to the desired descriptor values.

The motion graph we used contains a motion library divided into motion clips and the possible transitions between the motion clips. The motion clips are half to two seconds long samples from a nine minutes recording. The recordings contained motions that are required for moving on a flat surface (standing, turning, walking, running) and a few expressive motions (waving hands, jumping). The motions were recorded many times with differing styles to get versions with both high and low QoM similarly as shown in Figure 2.

We create the motion graph by finding all transitions from a frame to another where the pose and velocities do not differ too much with the restriction that the transitions are at least half a second apart. We do not prune any transitions from the motion graph as it would increase the reaction time of the virtual character and reduce the amount of possible reactions. During real-time interaction, we evaluate as many motion clip sequences as is possible during half a second, usually a few tens of thousands, and then select the sequence of clips that matches the desired descriptors best. The high number of possible clip sequences makes motions synthesis the most computing intensive part of the whole system.

One challenge here is that the desired values might require actions that cannot be performed simultaneously. For that problem, the scaling vector of the output values (Equation 3) helps as it tells which descriptors need to be prioritized and which can be ignored. The constraints set by the human body and physics offer another challenge as they should not be broken when natural looking motions are desired. The concatenative synthesis we use always produces physically plausible motions, but allows only approximate following of the desired descriptors.

During real-time interaction, selecting the next motion clips requires searching for the sequence of clips where the deviation from the desired motion descriptors is minimal. For the descriptors that can be calculated from an isolated character, this can be done by using the descriptor values that were calculated when preparing the motion graph. The relative descriptors cannot be calculated in advance as they vary depending on the other character. Also, the relative descriptors cannot be used directly in the search as the virtual character can only decide its own motion. Therefore, the relational values for distance, facing angle and approach/retreat are transformed into position, direction of movement and facing in the absolute coordinate system. These values can then be used as parameters to be optimized in the search for the next motion clip.

4 Example Cases of Authoring Behaviours

The simplest type of examples of authored behavior contains only one mapping between the input and the output descriptors. Let us consider a character that turns its face to another character. For this behaviour, the input should have all the scaling values of the descriptors set to zero and the output facing angle with value zero and weight one. During real-time interaction, having only this mapping would set the desired facing angle to zero. All the other descriptors would be ignored by the motion synthesis engine as they would have zero scaling values. A character following this rule would create an impression that it is aware of the position of the other character, but it would ignore all other aspects of the motion.

A more interesting character could be one that acknowledges the other character by turning to face it, but would get offended and turn away if the other character misbehaves. To create this behavior, an example pair of an action and a reaction shown in Figure 5 can be used. The behaviour rule can be created by scrolling through the action and reaction and by creating mappings from all the significant parts of the motion pair.

The motions start by having the characters far away form each other (fig. 5 A). A desired reaction in this case could be to ignore the far away character. This can be turned into a mapping where the input has the distance between the characters with weight one and the output has all the descriptors weighted zero. In the next part of the motions, the characters have come near each other and the reaction character has turned to face the action character (fig. 5 B). This can be turned into a mapping that has a low distance and low non-transitional QoM on the input side and low facing angle on the output side. The last part of motion has the action character waving hands forcefully and the reaction character responds by turning away (fig. 5 C). This part corresponds to a mapping where the input side has a low distance and high non-transitional QoM and the output has a high facing angle.

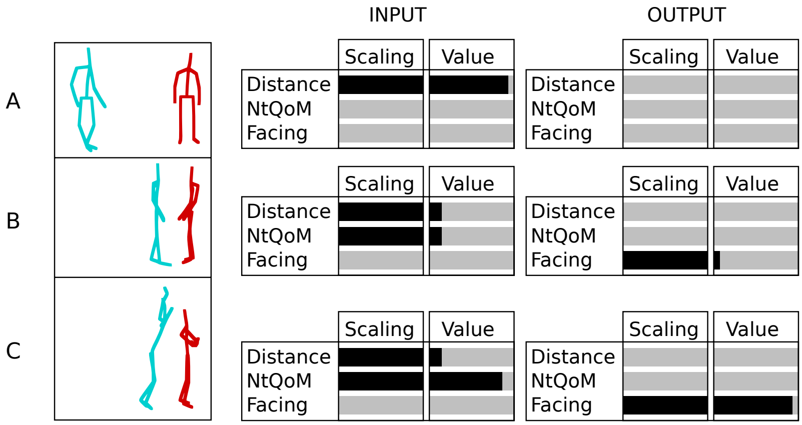


Fig. 5. Significant frames (A-C) from an example action (on the left side) and reaction (on the right side) motions, descriptor values picked from those parts of the motions and the scaling values decided by the author of the rule

After the mappings are defined, the rule is ready to be tested. The testing can happen by seeing if changing the values of input descriptors produce sensible output values. This can show if any errors were made while defining the rules. However, testing the rules with real-time interaction is also important as it can show if there are problems in the synthesis of the output descriptor values. The synthesis can fail if the virtual character is not able to find any possible motions that would fit the output descriptors quickly enough. This can be a problem especially if the motion graph has long motions that cannot be interrupted. Another possible problem is that an action can be so short that the input descriptors show the action for too short time to cause a reaction. This calls for more careful descriptor design and possibly descriptors that are calculated as an average over a period of time instead of just individual frames of motion.

Variations to the presented example behaviour could be that instead of turning away when provoked the character would start to be aggressive. The roles could be also swapped and then the virtual character would be the one starting the provocation.

5 Discussion

The example case shows that using recorded action and reaction motions guides the work flow of authoring interaction rules. The rules can be authored and tested with graphical and bodily user interfaces. Therefore, the requirements for the author creating the interaction rules do not include coding experience or learning an XML dialect. To further develop the usability of the system, user tests should be done with the users authoring new rules and interacting with virtual characters following those rules.

The used reactive interaction rules work in real-time and they are good for interactive background characters. Characters that need more intelligence could be built by adding reasoning capabilities and internal state into the virtual character and selecting the interaction rules based on the internal state. However, the interaction rules alone are not enough as the believability of the authored behaviours is heavily dependent on the capabilities of the motion synthesis engine.

It is challenging to synthesize motions that are realistic and balance the sometimes conflicting demands of expressiveness in real-time. The used motion graph approach is not ideal as it only allows playing motions a clip at a time, while perfect following of the desired output descriptors would require a more continuous synthesis method with a shorter reaction time. The situation could be helped by real-time filtering of the produced motion or using physics based motion synthesis when a sudden reaction is needed.

In other than research applications, using only bodily motions is not enough for creating a complete virtual character. We feel that modalities such as facial expressions and tone of voice could be added to the current authoring system. Also, defining musical interaction could be possible. The main requirement is that it must be possible to create continuous signals to describe the medium and to drive a synthesis engine with those signals. Combining the continuous interaction with discrete gestures could be more challenging. That would require deciding whether the continuous changes in the motion style only modulate the gestures or could they also interrupt or prevent the gestures.

One shortcoming in the presented approach is that the author of the behaviours has to assume the connections between the descriptor values that represent motion style and actual emotions visible in human motion. For example angry motions are likely to have high QoM, but there are many motions with high QoM that might not look angry. A possible solution could be to develop new descriptors that are learned from annotated data with machine learning techniques. The new descriptors could be estimates for visibility of emotions like anger and sadness. Simultaneous use the planned emotional descriptors and the ones that we have presented could allow more precise authoring of emotional reactions.

6 Conclusions and Future Work

In this paper we introduced a new way to author behavior rules for interactive virtual characters using bodily motions as the medium. We showed that the simultaneous use of several motion descriptors enables creation of expressive interaction rules. Since the descriptors are continuous in both the time and motion style domains, the produced interaction has a chance to be fluid and natural. The problems that emerge from the use of several descriptors include the curse of dimensionality and increased risk of physically impossible descriptor combinations. We solved the curse of dimensionality by adding scaling vectors for each set of descriptor values in each mapping. This reduced the amount of required mappings per interaction rule dramatically.

We also introduced a way to create mappings based on an example with action and reaction motions. This approach can reduce the risk of defining impossible descriptor combinations. The example also guides the creation of the interaction rules and makes the process a straight forward one. Even when using example motions, our method allows manual fine tuning of the rules.

In the future we intend to develop motion descriptors that measure visibility of emotions like sadness and anger from human motions. A promising approach is to create descriptors from annotated motion data by machine learning techniques. We see the development of these descriptors as a required step in order to go from expressive bodily interaction to emotional bodily interaction.

Acknowledgments. This work has received funding from the projects Enactive Media (128132) and Multimodally grounded language technology (254104) which are both funded by the Academy of Finland.

References

1. Huang, L., Morency, L.-P., Gratch, J.: Virtual Rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
2. Kovar, L., Gleicher, M., Pighin, F.: Motion graphs. ACM Transactions on Graphics (SIGGRAPH 2002) 21(3), 473–482 (2002)
3. Arikan, O., Forsyth, D.A., O'Brien, J.F.: Motion synthesis from annotations. ACM Transactions on Graphics (SIGGRAPH 2003) 22(3), 402–408 (2003)
4. Xu, J., Takagi, K., Sakazawa, S.: Motion synthesis for synchronizing with streaming music by segment-based search on metadata motion graphs. In: 2011 IEEE International Conf. on Multimedia and Expo (ICME), pp. 1–6 (2011)
5. 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)
6. Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., Volpe, G.: Multimodal Analysis of Expressive Gesture in Music and Dance Performances. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 20–39. Springer, Heidelberg (2004)
7. Young, J., Ishii, K., Igarashi, T., Sharlin, E.: Puppet Master: designing reactive character behavior by demonstration. In: Proc. of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2008), pp. 183–191. Eurographics Association, Aire-la-Ville (2008)
8. 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)
9. 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)
10. Jebara, T., Pentland, A.: Action Reaction Learning: Automatic Visual Analysis and Synthesis of Interactive Behaviour. In: Christensen, H.I. (ed.) ICVS 1999. LNCS, vol. 1542, pp. 273–292. Springer, Heidelberg (1998)
11. Young, J., Ishii, K., Igarashi, T., Sharlin, E.: Style-by-demonstration: Teaching Interactive Movement Style to Robots. In: ACM Conf. on Intelligent User Interfaces (IUI 2012), pp. 41–50. ACM, New York (2012)

12. Metaxas, D., Chen, B.: Toward gesture-based behavior authoring. In: Proc. of the Computer Graphics International 2005 (CGI 2005), pp. 59–65. IEEE Computer Society, Washington, DC (2005)
13. Ulicny, B., Ciechomski, P., Thalmann, D.: Crowdbrush: interactive authoring of real-time crowd scenes. In: Proc. of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2004), pp. 243–252. Eurographics Association, Aire-la-Ville (2004)
14. Vilhjálmsson, H.H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S.C., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: 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)
15. Zwiers, J., van Welbergen, H., Reidsma, D.: Continuous Interaction within the SAIBA Framework. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 324–330. Springer, Heidelberg (2011)
16. Pugliese, R., Lehtonen, K.: A Framework for Motion Based Bodily Enaction with Virtual Characters. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 162–168. Springer, Heidelberg (2011)
17. Buhmann, M.D.: Radial Basis Functions: Theory and Implementations. Cambridge University Press, Cambridge (2003)

Expressive Body Animation Pipeline for Virtual Agent

Jing Huang and Catherine Pelachaud

Telecom ParisTech - CNRS, Paris, France
{jing.huang,catherine.pelachaud}@telecom-paristech.fr

Abstract. In this paper, we present our expressive body-gestures animation synthesis model for our Embodied Conversational Agent(ECA) technology. Our implementation builds upon a full body reach model using a hybrid kinematics solution. We describe the full pipeline of our model that starts from a symbolic description of behaviors, to the construction of a set of keyframes till the generation of the whole animation enhanced with expressive qualities. Our approach offers convincing visual quality results obtained with high real-time performance.

1 Introduction

Embodied Conversational Agents are virtual human agents that can communicate through voice, facial expressions, emotional gestures, body movements etc. They use their verbal and nonverbal behaviors to convey their intentions and emotional states. It is necessary for the ECAs to display a large variety of behaviors.

Generating efficient and realistic animations of virtual creatures has always been an open challenge in the computer animation field. Kinematics is a general method for manipulating interactively articulated figures and generating postures. In computer graphics, articulated skeleton models are used to control virtual vertebral living creatures, such as human beings or animals, which appear frequently in films and video games. Inverse Kinematics (IK) is a method for computing joint rotation values of individual degree of freedom via predefined rotation and position constraints. In most of existing systems, animation of body parts is done independently of each other. For example, an arm gesture and torso movement are computed separately and then combined. Such computation gives rise to unnatural animation and stiff-looking creatures. ECAs are interactive entities; all their animations have to be rendered online.

Our work focuses on the realization of animation for virtual conversational agents. Our animation model takes as input a sequence of multimodal behaviors to generate. It relies on a hybrid kinematics solution for generating full body posture. Moreover our solution generates two types of motion. On the one hand it computes all movements specified over each modality (eg arm, torso or head movement). On the other hand, it also considers movements arising from other movements. For example it will generate a torso and shoulder movement resulting

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 355–362, 2012.

from a given arm movement (eg, when the arm has to reach a distant point in space). Considering independent and dependent movements gives rise to a more realistic and natural animation. Our model embeds also an expressive module with qualitative variations of body movements. Our algorithm is efficient; it can generate realistic whole body animations in real-time.

In the remainder of the paper, we first present existing approaches, then turn our attention to our approach. In section 3, we describe our animation pipeline with our expressivity model. We illustrate our explanation with examples.

2 Previous Work

In this section we will present different works regarding expressivity and animation computations that are related to our work. Several approaches have been proposed to model expressive behaviors. The EMOTE system [3] introduced low level parametrization to generate expressive gesture. The parameters are abstracted from Laban principles (1988). The Greta system [14] [4], defines a low level parametrization derived from psychology literatures. In the realization of animation, both of these works decomposed character skeleton into small parts (the head, the torso, the arms, etc) and solved the system by different controllers acting locally. Such an approach does not allow modeling motion propagations, ie how motion over one modality may affect another one. In our work, we choose to deal with whole body motion with a global view.

Michael Neff *et al.* [11] [12] [13] presented their aesthetic motion generation system. Their model starts from a high level expressive language that is translated into precise semantic units that can be simulated by physical or kinematics methods. The translation mechanism encloses a selection and a refinement steps for choosing the gesture movements.

SmartBody system [17] proposed a controlling system that employs arbitrary animation algorithms. The system can schedule different task controllers, it allows realizing modular animation control and propagating motions over the body parts. As SmartBody, we achieve a whole body management from a lower level. But our approach is based on the hierarchical dependency of the agent's body structure.

The EMBR system [6] offers an animation pipeline with their motion factory, scheduler and pose blender modules. The animation to be realized is described with the EMBR script. It can deal with different formats of animation (IK, motion data, etc). Our system follows a similar animation pipeline. However our pipeline solves the conflict for body parts. Indeed, our system is based on IK techniques. IK is used to do retargeting, and it makes the final decision for key postures. Several inverse kinematics [1] [2] [5] methods are also proposed to achieve reaching tasks. Such methods need to calculate the whole body posture in realtime for a given trajectory of the wrist position in space.

Tan [16] talks about the importance of using the postural expressions to express action tendencies. Meanwhile, behavioral studies on posture have also been made [7]. Although their studies are based on static postures, the authors noticed that expressive postures are rather important. They also note that generating automatically variations of expressive postures is useful for simulating human-like animation.

3 Implementation of Animation Pipeline

We have implemented a framework of embodied conversational agents that respects the SAIBA [9] framework illustrated in Figure 1. Our framework takes as input a file described with FML the standard Functional Markup Language which defines the intentions and emotional states in a high level manner. The Behavior planner translates the FML tags into sequences of standard BML, Behavior Markup Language, entries. Sequences of time-marked BML-like signals are instantiated within the Behavior Realizer.

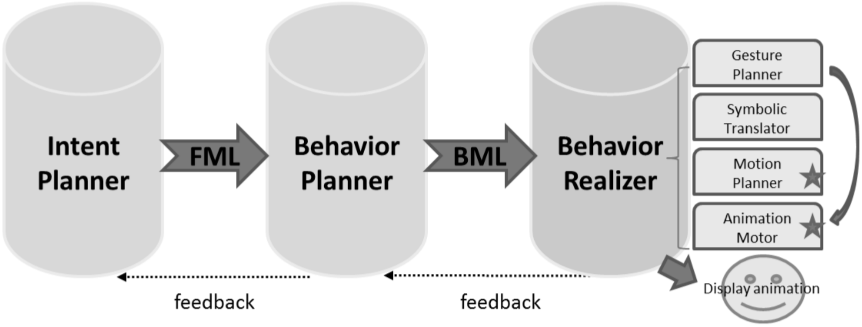


Fig. 1. The standard SAIBA framework defines the modularity, functionality and the protocols for ECAs. In the Behavior Realizer module, the parts with a star correspond to our work in the animation pipeline.

3.1 Overview of Our Pipeline

In this paper, we present the implementation of our animation pipeline. It starts by receiving BML-like symbolic signals time stamped in the motion planner. All signals are received by streaming, and hence our animation computations need to be achieved on the fly. Each signal (hand, torso, head, etc) includes gesture phases, expressivity parameters, gesture trajectory and the description of shape and motion. As shown in Figure 2, signals can be scattered (step 1) from different modalities, ie torso movements, head movements, hand gesture movements (two groups: left and right sides). After receiving scattered signals, we apply our pipeline to generate our animation sequence illustrated in Figure 2.

In the remainder of this section we detail each step of our pipeline. Expressivity parameters are presented in section 3.5.

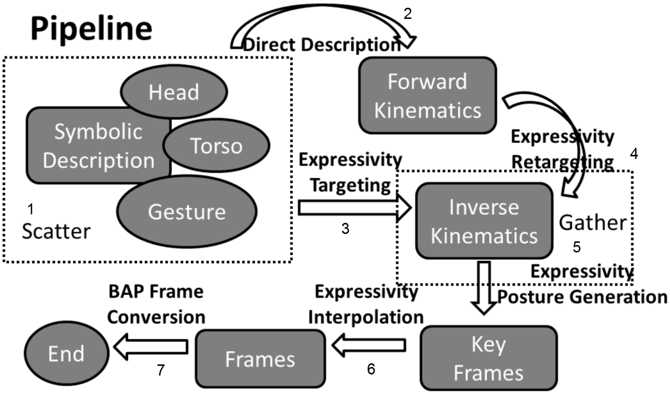


Fig. 2. The animation pipeline takes as input sequences of symbolic gestures. It computes the whole expressive animation; each step is numbered in this figure.

3.2 Targeting Process

The targeting process describes the hand gesture trajectory. This is often referred to the "Path driven" approach. Path form, such as line, circle, can be defined by mathematical functions. For building these path forms without the dynamic branching, we chose to use an approximation of a sinus function: $f(t) = R * sin(T * \pi * t + Shift)$, where R is the amplitude that defines the radius of local circle, T is the temporal variation of frequency, $Shift$ controls the path direction. The final path position P is defined as $P = P_{center} + P(f(t_x), f(t_y), f(t_z))$. By varying the 3 parameters, we can construct different gesture paths. For example, for linear path, T value can be just set to zero. Some other possible gesture paths are saved as sequences of key points corresponding to 3D positions in files.

3.3 Gathering Process

The purpose of the gathering process is to generate the full body key frames that corresponds to body postures. We chose to compute the full body posture as a whole and not as a concatenation of body parts positions as it allows capturing dependent movements across body parts. For example, reaching hand or gaze targets may affect torso movements. When looking on the far left, head, shoulder and torso are all turned to the left.

We sort out sequences of all body parts into one list of key frames ordered by time markers. Information of each key frame is filled with the information from the body part sequences. To complete the key frame specification, we can use either the "lazy" approach or the "complex" approach that are illustrated in Figure 3. The "complex" approach considers that all the movements are equally important, then we need to fill all the body parts for each key frame by interpolating between the previous element and following element of its sequence (linear interpolation); On the other hand, the "lazy" approach privileges some movements. E.g., if a key information has only torso movement, the torso movement

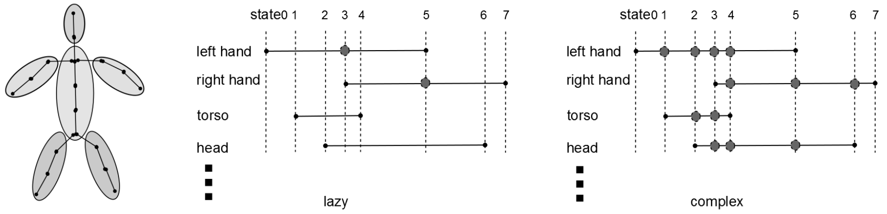


Fig. 3. **(Left):** shows the character skeleton decomposition, **(Middle):** shows the lazy approach. **(Right):** shows the complex approach.

is dominant, we can only apply torso movement; if a hand gesture is missing, but for one given sequence animation, it is very important, then we must fill it. For the lazy approach, we only fill in existing parts for each key frame and interpolate the important missing movements. The importance is defined by priority parameters. The lazy approach is flexible and has less computation. The resulting key frame list is used in the posture generation stage for creating posture key frames.

3.4 Posture Generation

To compute a body posture, we can apply forward kinematics, inverse kinematics or a mix of both techniques. The forward kinematics uses the description defined in the gesture phases to realize the initial states of the key frames.

Then, when needed, we retarget the gestures by using inverse kinematics. A hybrid inverse kinematics solution is used to solve dependency between body movements. For head and torso movements, we use simple analytical methods. For shoulders and hand gestures, we use our constrained mass spring IK solver. However we have to make sure that all targets are in the reaching space of the arms combined with the torso. If the target is too far away, the movement is transformed as hands pointing in the direction of the target. We do not consider foot step forward to solve this situation.

More concretely, we start by computing the potential target of torso. We check the targets of both hands $T1, T2$. We look if the hands movements needs torso movements. The arm length l_{arm}, the vertebrae (vt1, vl5 in MPEG4 H-ANIM) positions P_{vt1}, P_{vl5} are already defined in the skeleton system. The horizontal pointer of torso is the direction of the center of hands reaching targets. The amount of vertical lean is computed by using trigonometry approximations. Then we compute the hand gestures with this new torso posture. We apply our IK solver on the arm chain, starting from the sternoclavicular till the wrist. After this IK stage, we generate the animation sequence of the key frames using simple joint rotation informations. We use the quaternion based spherical linear interpolation(slerp) to generate all frames, and convert them into "BAP" frames (Body Animation Parameter (MPEG-4) frames).

3.5 Expressivity

We have defined a set of expressivity parameters to modulate the quality of body movements. We group these parameters into 3 sets: Gesture Volume controls the spatial variation of gestures; Sequential Variation controls the time based variation; Power variation allows us to further control the dynamism of these movements.

Gesture Volume: The idea of Gesture Volume is to use certain parameters to control the spatial form of posture shapes. The Spatial parameter is used to control the variation of McNeill's sectors [10]. We have two levels of control of these sectors. The first one is based on the spatial parameter which will scale the sectors using the same method as [4]. The second level adjustment depends on the torso position. The sector centers would be influenced by a scale factor and a rotation factor which are given by the torso. The scale factor $s = (P_{vt1'} - P_{vl5'})/(P_{vt1} - P_{vl5})$ is the ratio between the new length and the old length of the torso. The rotation factor is the rotation value applied on vl5 after the forward kinematics stage, which is used to change the sector box's orientation.

We also have the openness parameter that changes the gesture form computed in the IK stage. Its value ranges between 0 and 1. A small value tightens the body; it makes the elbow closer to the center; and a big value increases the space between the arm and the torso as illustrated in Figure 4 on the left. The openness parameter affects the orientation of the elbow swivel angle [18]. A larger openness value also applies a bigger rotation on the torso that indirectly releases the tension of the arms as shown in Figure 4 on the right.

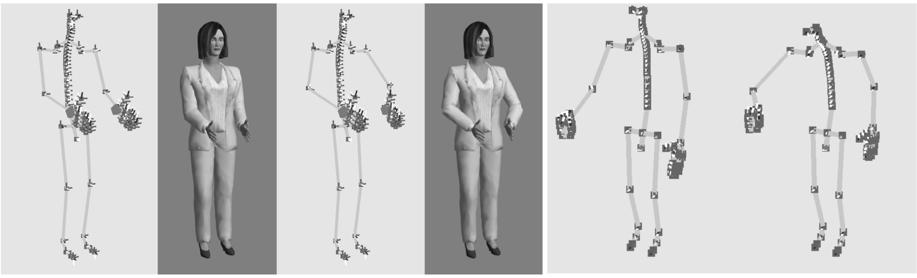


Fig. 4. Examples of different values of the openness parameter: **(Left)**: openness influences the gesture form. **(Right)**: openness influences the torso position.

Sequential Variation: Three parameters, "Fluidity", "Power" and "Tension", are used to modulate the gesture path as well as its timeline. Our signals are already time-stamped in the scatter module (see Figure 2). "Fluidity" refers to the degree of continuity of a movement. To simulate it, we use similar idea as proposed by [3] [4] to parametrize the Kochanek Bartels splines(TCB splines) [8]. We define "Power" as a force that changes implicitly the speed with certain acceleration. "Tension" [15] describes the amount of energy that has to be expended for some positions but not others, and hence more effort needs to be exerted for the gesture to keep its original position. These two last parameters are simulated by varying the bias parameter and the tension parameter of TCB splines. We also simulate accelerations by using some easing in-out functions to change time stamps for key frames (interpolation for target path) and frames (interpolation for joint rotation).

Additions: Power Variation: In our system, the power parameter does not only affect gesture paths, but also key frame postures. It means that a larger value of power influences more body parts. For example, a movement of the arm done with a high power will affect also shoulder and torso movements. This propagation of movements between body parts is possible as we use a full body IK framework. Torso can be affected by hand gestures as we define target energies. It is a similar idea as the constraints priority [1]. Our hybrid solver builds upon an interactive manner which is controlled by the "Power" parameter. As "Power" increases, it can influence the whole body gesture, both shoulders and torso.

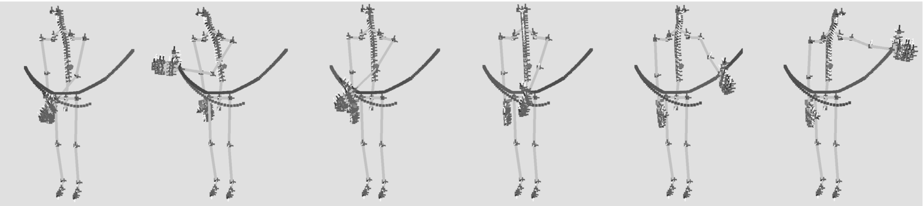


Fig. 5. One sequence animation using IK postures of our skeleton

3.6 Result of Virtual Agent

Our pipeline has been integrated into our virtual agent framework. The skeleton system is based on MPEG-4 H-ANIM 1.1 model. Our virtual agent system takes as input an FML file. It instantiates this input file into a sequence of BML tags. This sequence is the input of our animation pipeline.

Our solution encompasses two passes: forward kinematics specified by the BML tags; inverse kinematics that further influences the body depending on the energy descriptions and the expressivity parameters. We can note that torso and shoulder movements can be automatically generated due to movement propagations. Such movement can happen implicitly without any torso movements defined by BML tags, as illustrated in Figure 5.

4 Conclusion

In this paper, we have presented a full body expressive animation pipeline. We have proposed a hybrid approximation for posture computation based on the dependency of body parts. Our pipeline is compatible with existing target reaching models. Our expressive posture method provides high quality visual results in real-time. This system offers more flexibility to configure expressive FK and IK. It can be extended to other articulated figures.

Acknowledgements. This work has been funded by the French National Research Agency (ANR, for the project: CeCil).

References

1. Baerlocher, P., Boulic, R.: An inverse kinematics architecture enforcing an arbitrary number of strict priority levels. Vis. Comput. 20(6), 402–417 (2004)
2. Boulic, R., Thalmann, D.: Combined direct and inverse kinematic control for articulated figure motion editing. Computer Graphics Forum 11(4), 189–202 (1992)
3. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: SIGGRAPH 2000, New York, USA, pp. 173–182 (2000)
4. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing expressive gesture synthesis for embodied conversational agents, pp. 188–199 (2006)
5. Hecker, C., Raabe, B., Enslow, R.W., DeWeese, J., Maynard, J., van Prooijen, K.: Real-time motion retargeting to highly varied user-created morphologies. ACM Trans. Graph 27(3), 27:1–27:11 (2008)
6. 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)
7. Kleinsmith, A., Bianchi-Berthouze, N.: Recognizing Affective Dimensions from Body Posture. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 48–58. Springer, Heidelberg (2007)
8. Kochanek, D.H.U., Bartels, R.H.: Interpolating splines with local tension, continuity, and bias control. In: SIGGRAPH (January 1984)
9. 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)
10. McNeill: Hand and Mind: What Gestures Reveal About Thought. The University of Chicago press, Chicago (1992)
11. Neff, M., Fiume, E.: Modeling tension and relaxation for computer animation. In: SCA 2002, pp. 81–88. ACM, New York (2002)
12. Neff, M., Fiume, E.: Artistically based computer generation of expressive motion. In: Proceedings of the AISB, pp. 29–39 (2004)
13. Neff, M., Fiume, E.: AER: aesthetic exploration and refinement for expressive character animation. In: SCA 2005, pp. 161–170. ACM, New York (2005)
14. Niewiadomski, R., Bevacqua, E., Le, Q.A., Pelachaud, C.: Cross-media agent platform, pp. 11–19 (2011)
15. Edwards, A.D.N., Harling, P.A.: Hand tension as a gesture segmentation cue. In: In Proceedings of the Progress in Gestural Interaction, pp. 75–88. MIT mimeo (1997)
16. Tan, N., Clavel, C., Courgeon, M., Martin, J.-C.: Postural expressions of action tendencies. In: Proceedings of the 2nd International Workshop on Social Signal Processing. ACM, New York (2010)
17. 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 (2008)
18. Tolani, D., Goswami, A., Badler, N.I.: Real-time inverse kinematics techniques for anthropomorphic limbs. Graph. Models Image Process (2000)

The Turning, Stretching and Boxing Technique: A Step in the Right Direction

Mark Dunne, Brian Mac Namee, and John Kelleher

Applied Intelligence Research Centre
School of Computing, Dublin Institute of Technology
Mark.Dunne@dit.ie
www.comp.dit.ie/aigroup

Abstract. This paper describes a combination of three graphical processes for rendering a 3D avatar for an intelligent virtual agent (IVA) on a 2D display that increases the perceived *presence* of the avatar in the viewer's environment. The results of an experiment that shows that these processes positively effect *presence* are presented and analysed.

1 Introduction

This paper presents a combination of three graphical processes ('*turning*', '*stretching*' and '*boxing*') called the TSB technique. With the goal being to help the user experience a sustained 3D illusion of the avatar '*being with*' them. An evaluation of the effectiveness of the TSB technique in sustaining a high level of perceived *presence* for the avatar is also presented.

With a clear focus on the avatar's '*corporeal*' presence [1], this research looked for two contributing factors of presence according to Slater [2]:

- *Place Illusion* (Pl): when users of immersive virtual reality respond realistically to the events taking place in the virtual environment.
- *Plausibility Illusion* (Psi): the illusion of "*what is apparently happening, is actually happening*".

There are many examples of users having to wear a head mounted displays in order to view 3D avatars, like the virtual autonomy assistant (VAA) [3]. This may increase the avatar's perceived presence as a stereoscopic image is produced, just as any other 3D technology achieves. However, the wearing of such equipment is not practical in many scenarios. The TSB technique as outlined in this paper does not require the user to wear additional hardware. The user is tracked in the real world by a Kinect[1] and avatar's image is updated from the user's perspective at all times in order to maintain a 3D effect, similar to IGaze [4]. This allows very accurate reading of the avatar's gaze by the user when the avatar is referencing real-world locations (objects or people) with gaze alone.

[1] *Microsoft Kinect*: Kinect for Xbox 360 was released in November 2010. Website: http://www.xbox.com/en-ie/kinect

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 363–369, 2012.

2 The Turning, Stretching and Boxing (TSB) Technique

The framework presented in this paper combines three image altering processes[2] called: *Turning*, *Stretching* and *Boxing*. When combined these three image altering graphical processes deliver a constant 3D illusion of the 3D avatar on a 2D display from a user's perspective. The maintained 3D illusion is similar to the user looking through a '*real-life*' window at the avatar. This means that when a user moves in front of the display, the 3D scene continuously updates to match the user's perspective. As a result of this the avatar is able to reference locations (people or objects) in the user's environment more accurately using just gaze.

The TSB technique is dependent on head position data obtained from the user by the Kinect in conjunction with the Kinect SDK[3] and output as X, Y and Z coordinates[4]. The next three paragraphs outline each of the three image altering processes

Turning. Simply achieved by having the avatar's head turn to direct its gaze down the virtual camera lens in the 3D scene in order to maintain eye contact with the user as they move. The virtual camera's position in the 3D scene is updated to match the user's head position in the real-world, when the avatar's gaze is directed towards the virtual camera it subsequently seems to be directed towards the user, an illusion know as the "*Mona Lisa effect*" [4]. The ability to direct gaze towards users is a crucial form of referencing for social agents [5] as it can be an indicator of the willingness of one social entity to engage in social interactions with another [6].

Stretching. This process involves stretching the image of the avatar on the 2D display according to the user's viewing angle. This counteracts any skewing or narrowing of the avatar's 3D image, which would otherwise break the 3D illusion, similar in effect to the responsive workbench [7]. The *stretching* process does not effect the 3D models or any relating animations in the 3D scene. It merely adjusts the view-port of the 3D scene to reflect the user's viewing angle. This means all 3D models and animations remain intact.

Boxing. This effect is achieved by placing the 3D avatar in a virtual box. From the user's perspective this virtual box is adjoining their real-life room, just like a window into another room, similar to Rational Craft's Winscape project[5]. The user's view of the room will change according to their viewing angle, acting just like a real-world window.

Individually the three processes are not novel and have been used before in other research [4,7]. What is novel is that we test this combination of processes for

[2] *TSB Technique In Action*: http://youtu.be/OWDMGoDH640

[3] *Microsoft's Kinect for Windows (SDK)*: Kinect for Windows was released in February 2012 with a Beta released in July 2011. Website: http://www.microsoft.com/en-us/kinectforwindows/

[4] Retrieved from the Kinect SDK *skeletal tracking* data, more specifically from the joint labelled: '*JointID.Head*'.

[5] *Rational Craft's Winscape Project* website: http://www.rationalcraft.com/Winscape.html

use in projected avatar interfaces, where an increase in the perceived corporeal presence of a 3D avatar can be beneficial to social interactions with users.

3 42 Moves Experiment

The purpose of the *42 Moves* experiment is to test the TSB technique with participants to see if a high level of perceived *corporeal* presence for the avatar can be achieved. Subsequently, giving the participants a greater ability to guess correctly where the avatar is directing its gaze as it is being projected on a 2D display. In general 2D displays require a viewer to remain in a stationary position at the optimal 90°viewing angle, commonly referred to as the "*sweet spot*" [8]. When viewing angles are more acute than the *sweet spot* (see Fig. 1 which outlines the *sweet spot* in the *42 moves* experiment) there is a deterioration in the effectiveness of any 3D illusion due to skewing of the rendered image from the user's perspective, a phenomenon known as "*lateral foreshortening*". This phenomenon damages a user's perceived presence of an avatar. Addressing this phenomenon should lead to an increased sense of social richness, realism and immersion during interactions with an avatar [9].

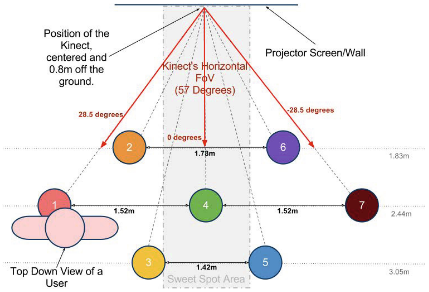


Fig. 1. A floor plan for the *42 Moves* experiment. Visible are the floor markers and the *sweet spot*.

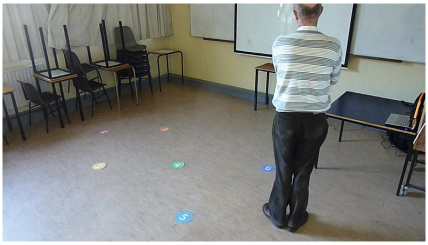


Fig. 2. A participant is standing on floor marker **7** waiting for their next move

Participants. There were 31 participants in total (9 females, 22 males) with ages ranging from 6 to 64 years. This diverse range in ages was selected as it covers the wide range of visitors a museum or other public building could get.

Procedure. This "*Wizard of Oz*" experiment[6] took place in a large empty room in which an avatar was projected onto one wall. There was 7 coloured circular markers on the floor that indicated specific positions in front of the projection (see Fig. 1 and Fig. 2). Each of the participants were required to start the experiment by standing on one of the seven floor markers randomly assigned to them. Depending on their floor marker they started on each participant proceeded on a predetermined sequence of 42 moves, for both of the experimental conditions. The participants would be guided through this sequence by the avatar's gaze

[6] Recording of the *42 Moves* Experiment: http://youtu.be/R41C3xL0zfE

alone. If the participant guessed incorrectly they were told to move to the floor marker the avatar was actually looking at in order to continue the experiment in the correct sequence. The avatar's gaze behaviour was exactly the same across both conditions. The conditions were:

- **Control Condition**: The avatar appears as it would on a regular 2D display, i.e., the rendering does not update to reflect a participant's position.
- **TSB Condition**: The TSB technique is switched on, therefore the avatar's 3D rendering is continuously updated to reflect the participant's perspective.

Survey. Participants took the same survey of six questions after each condition, adapted from a standard presence survey questionnaire [10]. The six survey questions will be presented in the next section (see Section 3.1 in paragraph '*Analysis of Survey Results*'). Each question was rated on a Likert scale as follows: *1. Very Low - 2. Low - 3. Average - 4. High - 5. Very High.*

3.1 Results

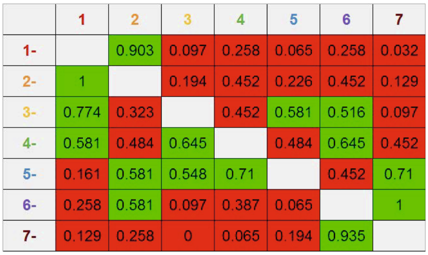
The *Control Condition* had an average percentage of correct moves made by participants of 41%, while the *TSB Condition* was 67%. A *paired t-test* showed that there was a significant difference in the scores for the *TSB Condition* (M=0.67, SD=0.2) and *Control Condition* (M=0.41, SD=0.28): t(41)=1.68, p=6.78^{-8}. These results suggest that during the *TSB Condition* participants tend to move to the correct marker more often than during the *Control Condition.*

In Section 1 we outlined two contributing factors of presence: **Pl** and **Psi**. We analysed the results of the experiment for these factors as follows:

- **Pl:** It was theorized that if the participant could read the avatar's gaze correctly more often, that would indicate an increase in Pl. This factor can be measured quantitatively by counting the participant's correct moves for each condition.
- **Psi:** This factor is more subjective and survey questions were used to qualify the participants' experiences and evaluate how much participants believed the avatar was actually able to look into the real world.

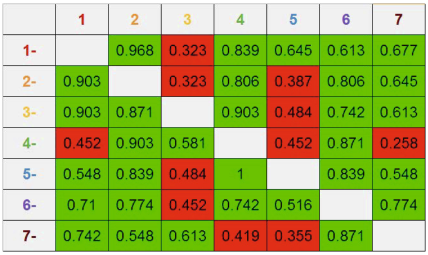
Figs. 3 and 4 illustrate the performance of participants in moving between specific floor markers for both conditions. The green blocks represent the average score for a move with accuracy rates greater than 50% and the red blocks represent accuracy rates less than 50%. There are significantly more green blocks in the TSB matrix, indicating that the avatar achieved a greater level of communication using its gaze with the TSB technique switched on.

There was an observably high accuracy rate for the participants in the *Control Condition* for moves **2-1** and **6-7** (see Fig. 3). This was not surprising as in the *Control Condition* participants were typically able to make broad interpretations of whether or not the avatar was looking to the *left* or *right*. When a participant was on marker 2 or marker 6 and the avatar looked right or left respectively, the participant had a easy choice to make – illustrated in the accuracy levels of 100% for moves **2-1** and **6-7**.



	1	2	3	4	5	6	7
1-		0.903	0.097	0.258	0.065	0.258	0.032
2-	1		0.194	0.452	0.226	0.452	0.129
3-	0.774	0.323		0.452	0.581	0.516	0.097
4-	0.581	0.484	0.645		0.484	0.645	0.452
5-	0.161	0.581	0.548	0.71		0.452	0.71
6-	0.258	0.581	0.097	0.387	0.065		1
7-	0.129	0.258	0	0.065	0.194	0.935	

Fig. 3. Control Matrix: Green Blocks: average scores greater than 50%, Red Blocks: average scores less than 50%.



	1	2	3	4	5	6	7
1-		0.968	0.323	0.839	0.645	0.613	0.677
2-	0.903		0.323	0.806	0.387	0.806	0.645
3-	0.903	0.871		0.903	0.484	0.742	0.613
4-	0.452	0.903	0.581		0.452	0.871	0.258
5-	0.548	0.839	0.484	1		0.839	0.548
6-	0.71	0.774	0.452	0.742	0.516		0.774
7-	0.742	0.548	0.613	0.419	0.355	0.871	

Fig. 4. TSB Matrix: Green Blocks: average scores greater than 50%, Red Blocks: average scores less than 50%.

However, the results in Fig. 3 also show that when the participants were on floor markers 1 or 7 for the *Control Condition* and the avatar looked to the left or right respectively, participants had a difficult choice. With the exception of the **1-2** and **7-6** moves accuracies for moves from marker 1 and marker 7 are extremely low. The high accuracies for moves **1-2** and **7-6** were because these were seen as the best *damage limitation* moves from marker 1 and marker 7 respectively and so were chosen to a large extent.

The results show that there is no significant difference between the two conditions for moves starting from floor marker **4**, i.e., the *sweet spot* (see Fig. 1). A *paired t-test* shows this in the scores from floor marker **4** for the *TSB Condition* (M=0.59, SD=0.26) and *Control Condition* (M=0.54, SD=0.09): t(5)=2.57, p=0.66. However, on the contrary when participants had to move from all the other floor markers which lie outside of the *sweet spot* (i.e.,**1**, **2**, **3**, **5**, **6** and**7**), results from a *paired t-test* show a significant difference in the scores for the *TSB Condition* (M=0.68, SD=0.19) and *Control Condition* (M=0.39, SD=0.29): t(35)=2.03, p=1.67×10^{-8}. The use of the TSB technique does seem to go a long way to compensate for the reliance on the participant to be in the *sweet spot* when avatars are displayed on a 2D display.

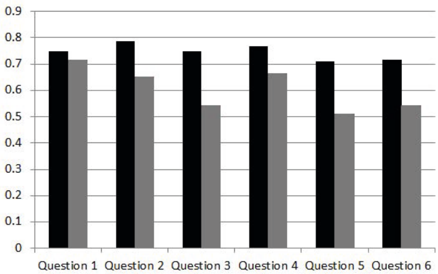


Fig. 5. Survey Results: The black bars are the average % rating for the survey questions with the TSB technique on and the grey bars represent the same but for the *Control Condition* (% ratings were converted from the Likert scale (1 to 5) data).

When the survey results were averaged (see Fig. 5) they indicated that participants gave higher ratings for all the questions after completing the *TSB Condition*. A *paired t-test* shows that there is a significant difference in the survey question ratings for the *TSB Condition* (M=0.75, SD=0.02) and *Control Condition* (M=0.60, SD=0.08): t(5)=2.57, p=0.34×10^{-2}. The details of the questions and some explanations for participants' responses are given below:

1. *To what degree did you become so involved in doing the task that you lost all track of time?* Marginally better results for the *TSB Condition* could be put down to the fact people got more moves correct and they did not have to be repositioned as often. Hence, they were more engrossed for longer periods of time throughout the experiment.
2. *To what degree did you feel the 3D virtual character's head movements were natural?* The results suggest that participants perceived that the avatar's head movements were more natural during the *TSB Condition*. This indicates that the Psi factor for the *TSB Condition* would seem to be higher.
3. *To what degree did you feel the 3D virtual character's gaze direction towards the spots on the ground was realistic?* The results here are in favour of the *TSB Condition*. This can be put down to the fact that participants did better during the *TSB Condition* so they scored the avatar's gaze direction higher to reflect their own performance.
4. *To what degree did you feel the 3D virtual character was responsive to your actions?* It was predicted and the results show that there is little difference between the conditions as the avatar's responsiveness is identical for both.
5. *To what degree did your experience with the 3D virtual character's gaze seem consistent with your real world experiences?* The difference between results for both conditions was substantial here, indicating that for the *TSB Condition* participants on average believed that the avatar's gaze seemed more consistent with real-world experiences.
6. *To what degree do you think the 3D virtual character was actually able to look out at the real world 'spots' on the ground?* Relating directly to the Psi factor, a higher average rating by participants for the *TSB Condition* indicates a higher sense of perceived *corporeal* presence for the avatar. To what degree this was achieved is debatable and requires further study.

4 Discussion

The Kinect's *field of view* (FoV) (57°with a range of 0.5 - 4 m) is limited, ideally a bigger FoV with a greater range would be more useful as a sensor. Also, the use of 3D display technology would only render the image to stereoscopic. However, 3D display technology alone would still require participants to be in the *sweet spot*. Thus, the TSB technique is still required in order to combat *lateral foreshortening*, further research is needed to substantiate this claim.

The quantitative results are pretty definitive that the TSB technique increases the participant's accuracy for moving to the correct floor marker. This can be attributed to the fact that the 3D illusion of the avatar on the 2D display is maintained by the TSB technique from the participant's changing perspective.

Increasing the participant's sense of the avatar "*being there*" with them, which Slater refers to as Pl [2], represents an increase in the perceived *corporeal* presence for the avatar. The survey results indicate that participants did experience increases in Psi during the *TSB Condition.* This indicates that the participants were better able to suspend their disbelief. This is important in mediated communication so the human communicator does not feel as if they are talking to the medium, as this can prevent a natural communication style from occurring.

The increase seen so far in Pl and Psi that contribute the overall *corporeal* presence perceived by participants indicates that further research is needed. A follow up experiment should focus more on the subjective experiences of the participants in relation to presence—with a more extensive survey used to determine a participant's level of perceived presence for the avatar. However, the results suggest that the TSB technique is a good first step in enabling an avatar to deliver accurate directions through its gaze.

References

1. Holz, T., Campbell, A.G., O'Hare, G.M.P., Stafford, J.W., Martin, A., Dragone, M.: MiRA – Mixed Reality Agents. International Journal of Human-Computer Studies 69(4), 251–268 (2011)
2. Slater, M.: Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philosophical transactions of the Royal Society of London 364, 3549–3557 (2009)
3. Wiendl, V., Dorfmüller-Ulhaas, K., Schulz, N., André, E.: Integrating a Virtual Agent into the Real World: The Virtual Anatomy Assistant Ritchie. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 211–224. Springer, Heidelberg (2007)
4. Kipp, M., Gebhard, P.: IGaze: Studying Reactive Gaze Behavior in Semi-immersive Human-Avatar Interactions. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 191–199. Springer, Heidelberg (2008)
5. Lance, B., Marsella, S., Koizumi, D.: Towards expressive gaze manner in embodied virtual agents, 194–201 (2004)
6. Peters, C.: Towards direction of attention detection for conversation initiation in social agents. In: Joint Symposium on Virtual Social Agents, AISB 2005, pp. 37–44 (April 2005)
7. Agrawala, M., Beers, A.C., McDowall, I., Fröhlich, B., Bolas, M., Hanrahan, P.: The two-user Responsive Workbench: support for collaboration through individual views of a shared space. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1997, pp. 327–332. ACM Press/Addison-Wesley Publishing Co., New York (1997)
8. Raskar, R., Brown, M.S., Yang, R., Chen, W.C., Welch, G., Towles, H., Scales, B., Fuchs, H.: Multi-projector displays using camera-based registration. IEEE Visualization, 161–522 (1999)
9. Lombard, M., Ditton, T.: At the Heart of It All: The Concept of Presence. Journal of Computer-Mediated Communication 3(2) (1997)
10. Witmer, B.G., Singer, M.J.: Measuring Presence in Virtual Environments: A Presence Questionnaire. Presence: Teleoperators and Virtual Environments 7(3), 225–240 (1998)

From Their Environment to Their Behavior: A Procedural Approach to Model Groups of Virtual Agents

Rafael Hocevar[1], Fernando Marson[1], Vinícius Cassol[1], Henry Braun[1],
Rafael Bidarra[2], and Soraia R. Musse[1]

[1] Graduate Course in Computer Science
Pontifícia Universidade Católica do Rio Grande do Sul - Porto Alegre, Brazil
rafael.moura@acad.pucrs.br, soraia.musse@pucrs.br
[2] Computer Graphics Group of Delft University of Technology - Delft, The Netherlands
R.Bidarra@tudelft.nl

Abstract. Simulation of everyday situations from real life can be a very useful tool in entertainment applications and training systems. Such applications, as games or computer animated movies usually need to provide virtual environments populated with virtual autonomous agents. Commonly, the agents need to be able to evolve in their environment, avoiding collision with each other and obstacles, besides interacting with other characters in order to provide realistic simulations. We present a model to simulate coherent group behaviors based on procedural modeling and semantic environments. Our main focus is virtual environments and agents, present in the background of games or movies generated with few/without user intervention.

1 Introduction

Human beings usually group with others in regular situations. When grouped, individuals interact with others according to their type of relationship, as well as the environment characteristics. Studies about human behaviors are produced since the 20 century [1,2]. The goal of such studies is to identify, for example, the distribution of individuals participation in small groups and also to analyze their interactions. Simulations of virtual agents interacting with others in a virtual environment can be applied in different areas, such as entertainment, engineering and security. We consider the method by [3], inspired in a biological algorithm, based on competition for space in a coherent growth of veins and branches [4]. Our main contributions are: i) to provide a strong connection between groups of virtual agents and the environment, i.e. world can be used to change the group behaviors; and ii) to enable group behaviors with fewer user interventions (agents characteristics are created as a function of environment and time). Simulations automatically generated using our technique allow to the animator be focused in the big picture and in the first plan characters [5]. Our model can be applied in games in order to coherently populate the environment (e.g. empty buildings).

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 370–376, 2012.

2 Related Work

Several aspects of group behaviors have been analyzed in the last years. Results of behavior groups analysis provide an useful reference for simulation/animation of groups and crowds [6,7,8,9,10]. Two important aspects that guide the motion of real people are: goal seeking, which reflect the target destination of each individual; and the least-effort strategy, reflecting the tendency of people to reach the goal along a path requiring the least effort [10]. Edward Hall [11] proposes the *proxemics* concept, which is the study of measurable distances between people as they interact. The specification of such distance can be based on different parameters: the agents relationship, the environment, the density of characters, among others. More specifically, concerning the motion of groups, Kamphuis and Overmars [12] introduce a two-phase approach, where a path for a single agent is generated by any motion planner. Then, a corridor is defined around the path, where all agents stay inside. Musse et al. [13] describes a model for controlling groups motion based on automatic tracking algorithms.

Recent works aim to produce coherently and realistically group behaviors taking into account steering and formation of groups. Karamouzas *et al.* [14] present a model where the velocity space to plan the avoidance maneuvers of each group is used to maintain a configuration that facilitates the social interactions between the group members. To provide the groups formation, the authors were inspired in the work developed by Moussaïd *et al.* [15]. Such research shows that the majority of the pedestrians walks in small groups of up to three members following formations as *Side by Side, V-Like* and *River-Like* formations. In comparison, our groups also keep formation, however we firstly change group formation based on environment restrictions, and if there is free space, we also keep formation based on best efficiency in social behavior that is achieved in *Side by Side* formation [15]. An important contribution from our model is the connection between the population and the semantic environment, which constraints the motion behavior. Next section details our model.

3 The Model

Our model is mainly focused on the groups behaviors when evolving in a virtual environment regarding other groups location, density of agents in the space and environment characteristics (obstacles, interest locations, etc). It is important to emphasize that our model is suitable for background actors and actions, requiring minimum intervention of designers or users.

3.1 Semantic Environment

A Semantic Virtual Environment (SVE) [16,17] is a virtual environment that is populated with entities enriched with semantics. A simulation environment is a complex space that is composed by a hierarchical set of simpler spaces, such as a city. Commonly, several neighborhoods composes a city, which are composed by many lots. These lots might have several types of buildings, with different types of rooms: kitchen, bedroom, bathroom, living room, among others. To specify goals in the environment we can assign a special attribute to any object that indicates some interesting thing or a resource that a certain object can provide. For instance, a TV provides *fun* as a chair provides *rest*. If an object that provides some resource is placed inside of a space, this space will provide that resource as well.

Our generator uses the Semantic Engine based on [18] in order to specify, create and store all spaces and objects. The specifications are made by a *template file* containing the spaces hierarchy and all objects that are instantiated in the environment. Once defined the environment, the next step is to create information about how to populate these spaces and what agents can do during the simulation. For that, we create a *Population Class* (PC), which can automatically create a random population for a specific environment. As output of our SVE Generator we have a 2D layout that contains goals, walkable and non-walkable regions and a graph that will be used by the virtual agents to compute their paths. Another output is a 3D scene, that contains all 3D geometric representation of objects and spaces.

3.2 Virtual Population

This section describes how agents are generated in the VE, considering the PC that is composed by following information: i) the simulation total time; ii) the higher density of agents to be attained during the simulation and the time it should occur. In this case, the simulation process is responsible for creating and destroying the agents (e.g. at the beggining and ending of a party) in order to attain the expected density of agents at a specific period of the simulation. Definition iii) is concerned with groups distribution presented in a certain population; and iv) the distribution of interest, entry and spawn locations where agents should be created or go to. In order to avoid user intervention, our model can automatically generate a population P based on such definitions. From first frame until the expected peak of density of the simulation, agents are linearly created in the environment. The spaces and the objects previously defined can be the goals that are randomly distributed to the agents in the simulation. Once an agent reaches its goal, it stays there for a random time until a new goal is randomly selected. Moreover, agents can group with others. On the other hand, one or more groups G can also be created into a specific P, i.e. people into P that should physically interact are pre-defined, being maximum of 3 agents in each group, since mostly groups in real life are formed by up to 3 people [19]. When the distance between 2 members of G_i is into a range defined by Hall's social distance [11], one group G is formed. According to [3], the *personal space* for each agent A is modeled as a circular region, that represents a "perception field" which can be used by each agent to avoid collision with others. In our case, we adopt another circular region we called *group space* (see Figure 1) which includes the N members of G_i, and radius Rg, computed based on follow Equations:

$$Af = argmax(dist(A_i, \vec{C})), \tag{1}$$

$$Rg = dist(A_{Af}, \vec{C}) + \overline{R_{Af}}, \tag{2}$$

where $\vec{C}$ is the centroid of all A_i positions and it is also the center of G_i with radius Rg, and Af is the index of the agent which position is farther from the centroid $\vec{C}$. When agents are grouped they have equivalent goals and speeds.

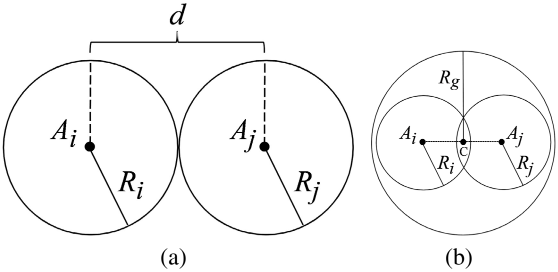
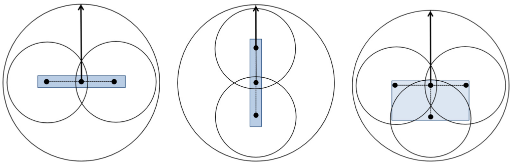


Fig. 1. a) Hall's [11] social distance (d) between two agents. b)If the agents are from same group G, a *group space* is defined representing the group as an unique entity.

3.3 Simulation

This phase is responsible for providing groups and agents motion and their interactions into the environment. Each group G_i can evolve in a virtual environment, which can also be populated by other groups G_j. The following aspects are considered in order to define the individual behavior function of agent A which is member of G_i: i) agent's goal which is defined based on environment information; ii) density of agents close to G_i ; iii) obstacles and other constraints around G_i; iv) location of close groups G_j. The agents movement is inspired in [20] algorithm. As in the original model, the environment is represented by a set of markers which discretizes the space. Overlayered to the markers, we create a grid of nodes in the space where motion is allowed and used as reference to the A* path planning algorithm, considering environment features [21,22]. Basically, two group behaviors emerge from this connection between agents and environments. Firstly, grouped agents can present groups formation while evolving in the virtual environment. Secondly, they can vary their behavior (formation and trajectories) based on environment constraints and people density. In order to avoid that groups occupy the same space we keep a small region inside group space where the markers cannot be used for agents from a different group. After defined the members in a group, we are able to compute the physical agents position into the group space to determine the *group formation*. In such area, we can provide three formations, inspired in work proposed by [15] and illustrated in Figure 2.

The *River-Like* formation can be considered an emergent behavior in our model. Considering the agents with the same goal, being part of the same group, they are able to move in the same direction at the same speed, as in [20], emerging such formation. To provide the *Side by Side* formation, we perform a simple test of angles in order to keep the agents aligned and perpendicularly placed given their goal. At each agent step we



(a) Side by side. (b) River-like. (c) V-formation.

Fig. 2. Group formations and the respectives $ROIs$

check the angle between the vectors formed by the group centroid $\vec{C}$ to each agent and the vector from $\vec{C}$ to the next goal. When angle between these two vectors is greater than $90°$ we increase the agent speed. For angles smaller than $90°$, we decrease the agent speed. With such rules, the agents keep walking close to each other in a specific formation during the motion process (groups of 2 and 3 agents). Similar process is performed to obtain *V-Formation*, however, angles between agents and $\vec{C}$, as well as $\vec{C}$ and next goal should be approximately $45°$. These formations are chosen as a function of the free space around the group. Our method considers the space in the goal direction trying to find out if there is available space for the group. Indeed, we compute for each group a 2D Region of Interest (ROI) that is included in *group space* and represents the minimum area that should exist in the environment for the group performance. These regions are simply computed based on agents positions and their sizes A_S. For this, we mathematically estimate the ROI for the three group formations and test them against the environment space, computing ROI_{side}, ROI_V and ROI_{river}. First test aims to keep the *Side by Side* formation, since it represents the best way, in social terms, for a group to go everywhere in low dense situations [15]. So, ROI_{side} is checked if it can be included in the region around G. For instance, if G is passing through a door with size smaller than ROI_{side}, then next group formation (*V-Formation*) is tested against the free space. If ROI_V is still larger than the free space in the environment, then *River-Like* formation is adopted.

4 Results

This section presents results obtained with our model (prototype implemented using *Irrlicht Engine* [23] and *Cal3D* [24]). Figure 3 represents a pool area in a backyard, where it is possible to observe modules of our model: on the left, it is illustrated the bounding boxes representing the semantic environment while on the right, a view of our simulation environment presents the agents and their paths provided by the path planning algorithm. We can observe a population (30 agents and 9 groups) automatic generated as well as their goals, initial locations and groups behaviors. Each group can be identified in Figure 3 by a circular area that represents the group space. We developed a framework to visualize the results of the simulation. The framework is also responsible for playing the animations accordingly the situation. Based on the agent's moving speed we are able to determine which animation should be played. Using data from the population, our algorithm is able to provide the agents motion across the environment.

In such process, groups formations are performed by our agents (*Side by Side, River-Like, V-Formation*). Moreover, the groups are able to identify the presence of other groups and compute a new path or new formation when needed. Figure 4 illustrates the visualization of a simulation example, where it is possible to observe the 3D virtual environment and also the virtual agents performing coherent group behaviors provided by our model.

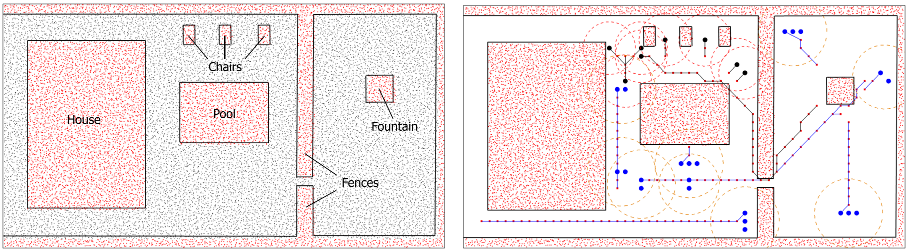


Fig. 3. On left, markers representing the walkable space. The dots inside boxes means regions where the motion in not allowed. On right, agents and their paths.

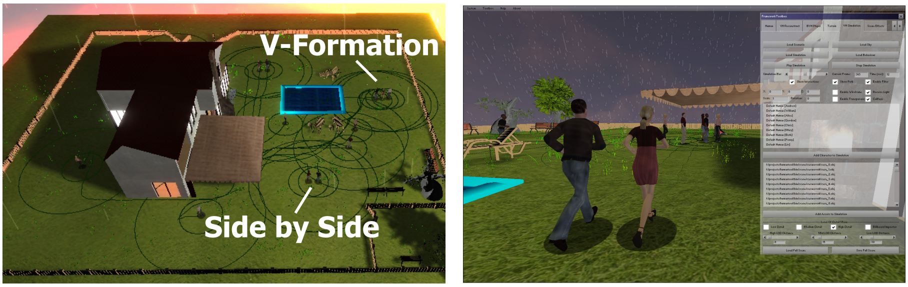


Fig. 4. Snapshots of our 3D framework, visualizing an example of simulation

5 Final Considerations and Future Works

This paper presents a model to provide procedurally coherent group behaviors in a semantic 3D environment. Only the environment is manually created, while all agents and groups behaviors can be automatically generated. Each group is able to perform different behaviors (*Side by Side*, *River-Like*, *V-Formation*) while avoiding collision with other groups and objects in the space. The main goal is to generate semantic environment and behavior for background scenarios. Obtained results can be easily visualized in real-time with our 3D framework. As future work, we intend to provide virtual agents able to interact with objects.

Acknowledgments. This work was supported by Brazilian research agencies CAPES, FAPERGS and FINEP using incentives of Brazilian Informatics Law ($n°$ 8.2.48 of 1991). The authors acknowledge the support granted by CNPQ and FAPESP to the INCT-SEC (National Institute of Science and Technology Embedded Critical Systems Brazil), processes 573963/2008-8 and 08/57870-9.

References

1. Chapple, E.D.: Quantitative analysis of the interaction of individuals. Proceedings of the National Academy of Sciences of the United States of America 25(2), 58–67 (1939)
2. Stephan, F.F., Mishler, E.G.: The distribution of participation in small groups: An exponential approximation. American Sociological Review 17(5), 598–608 (1952)
3. de Lima Bicho, A., Rodrigues, R.A., Musse, S.R., Jung, C.R., Paravisi, M., Magalhes, L.P.: Simulating crowds based on a space colonization algorithm. Computers & Graphics 36(2) (2012), 70 –79, Virtual Reality in Brazil (2011)
4. Sachs, T.: Polarity and the induction of organized vascular tissues. Annals of Botany 33(2), 263–275 (1969)
5. Thalmann, D., Musse, S.R.: Crowd Simulation. Springer-Verlag London Ltd. (2007)
6. Henderson, L.F.: The statistics of crowd fluids. Nature 229(5284), 381–383 (1971)
7. Henderson, L.F.: On the fluid mechanic of human crowd motions. Transportation Research 8(6), 509–515 (1974)
8. Fruin, J.J.: Pedestrian and planning design. Metropolitan Association of Urban Designers and Environmental Planners. Elevator World Inc., New York (1971)
9. Helbing, D.: Pedestrian dynamics and trail formation. In: Schreckenberg, M., Wolf, D.E. (eds.) Traffic and Granular Flow 1997, Singapore, pp. 21–36. Springer (1997)
10. Still, G.K.: Crowd Dynamics. PhD thesis, University of Warwick, Coventry, UK (2000)
11. Hall, E.T.: The hidden dimension / Edward T. Hall, 1st edn. Doubleday, Garden City (1966)
12. Kamphuis, A., Overmars, M.H.: Finding paths for coherent groups using clearance. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 19–28. Eurographics Association, Aire-la-Ville (2004)
13. Musse, S.R., Jung, C.R., Jacques Jr., J.C.S., Braun, A.: Using computer vision to simulate the motion of virtual agents. Computer Animation and Virtual Worlds 18(2), 83–93 (2007)
14. Karamouzas, I., Overmars, M.: Simulating the local behaviour of small pedestrian groups. In: Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology, VRST 2010, pp. 183–190. ACM, New York (2010)
15. Moussaïd, M., Perozo, N., Garnier, S., Helbing, D., Theraulaz, G.: The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS ONE 5(4), e10047 (2010)
16. Otto, K.A.: Semantic virtual environments. Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, pp. 1036–1037. ACM, New York (2005)
17. Gutierrez, M., Vexo, F., Thalmann, D.: Semantics-based representation of virtual environments. International Journal of Computer Applications in Technology 23(2-4), 229–238 (2005)
18. Kessing, J., Tutenel, T., Bidarra, R.: Services in Game Worlds: A Semantic Approach to Improve Object Interaction. In: Natkin, S., Dupire, J. (eds.) ICEC 2009. LNCS, vol. 5709, pp. 276–281. Springer, Heidelberg (2009)
19. Mills, T.M.: The sociology of small groups / Theodore M. Mills. Prentice-Hall, Englewood Cliffs (1967)
20. Rodrigues, R.A., de Lima Bicho, A., Paravisi, M., Jung, C.R., Magalhaes, L.P., Musse, S.R.: An interactive model for steering behaviors of groups of characters. Appl. Artif. Intell. 24, 594–616 (2010)
21. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. SIGART Bull., 28–29 (1972)
22. 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: Proc. the Twelfth IASTED International Conference on Computer Graphics and Imaging (CGIM 2011), Innsbruck, Austria. IASTED/ACTA Press (2011)
23. Irrlicht Engine (2012), http://irrlicht.sourceforge.net
24. Cal3D: 3d Character Animation Library (2012), http://gna.org/projects/cal3d

Social Evaluation of Artificial Agents by Language Varieties

Brigitte Krenn, Stephanie Schreitter, Friedrich Neubarth, and Gregor Sieber

Austrian Research Institute for Artificial Intelligence, 1010 Vienna, Austria
firstname.lastname@ofai.at

Abstract. In Sociolinguistics, language attitude studies based on natural voices have provided evidence that human listeners socially assess and evaluate their communication partners according to the language variety they use. Similarly, research on intelligent agents has demonstrated that the degree an artificial entity resembles a human correlates with the likelihood that the entity will evoke social and psychological processes in humans. Taking the two findings together, we hypothesize that synthetically generated language varieties have social effects similar to those reported from language attitude studies on natural speech. We present results from a language-attitude study based on three synthetic varieties of Austrian German. Our results on synthetic speech are in accordance with previous findings from natural speech. In addition, we show that language variety together with voice quality of the synthesized speech bring about attributions of different social aspects and stereotypes and influence the attitudes of the listeners toward the artificial speakers.

Keywords: language-attitude study, synthetic voices, virtual character design, social evaluation.

1 Introduction

In the present paper, we explore the question in how far the language variety used by an artificial agent influences how the human communication partner socially perceives and evaluates the agent. Based on results from language attitude studies on natural voices, we predict that synthetic voices representing a standard language variety are rated differently than colloquial or dialectal synthetic voices. Additionally, we investigate effects of synthesized standard and non-standard language varieties with respect to those aspects of a character relevant for the social interpretation and evaluation of the character. This may influence how the human assesses the intelligence, naturalness, politeness, etc. of the character, and related assumptions such as education, profession, social background of the character, its habits and preferences. A better understanding of such effects is crucial for the design of artificial human-like agents, where the agent's voice, its appearance and behaviour, as well as the application context must match.

In order to gain deeper insights into aspects of social interpretation that are triggered by language variety, we conducted a language attitude study on three

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 377–389, 2012.

Austrian language varieties, a standard Austrian German male voice, a colloquial Viennese female voice, and a dialectal Viennese male voice. To the best of our knowledge, this kind of language attitude studies on synthesized language varieties are novel. The synthetic voices are implemented with the open domain unit selection speech synthesis engine Multisyn of Festival.[1] For further details on speaker selection and the design of the voices see [22].

To set a context for the language varieties, we have built the synthetic voices into an existing cultural heritage application, where an invisible tourist guide (represented by voice only) accompanies a human visitor within an interactive 3D model of the Baroque State Hall of the Austrian National Library which is one of the world's finest historic libraries. Visitors are able to seek their own interactive ways through the State Hall and ask the guide for information or they may follow the guide on selected virtual tours. For further details on the application see [11,27]. For the language-attitude study, we have replaced the existing voice within the application with the three Austrian varieties. A part of the guided tour covering the statues in the State Hall was used as material for assessing human evaluations of the synthesized language varieties. We chose this application because it provides a context for the appearance of a disembodied/invisible agent and thus allows concentration on characteristics conveyed by language variety, ensuring that social interpretation of the tour guides solely arises from language variety and voice quality.

In the following sections, we first introduce related work (Sec. 2). We then present the design of the experiment including hypotheses, methods employed, and characteristics of the group of participants (Sec. 3). This is followed by an analysis and discussion of the data (Sec. 4). A summary and outlook is presented in Sec. 5.

2 Related Work

Since the 1980s, research has been carried out showing that the degree an artificial entity resembles a human correlates with the likelihood that the entity will evoke social and psychological processes in humans, e.g. [32,31]. In psychology, attribution theory focuses on interpretations and ascription of causality to events by individuals, see [7,10,30]. Dubinsky [5] summarizes the key assumptions of these slightly different approaches as follows: 1) people try to determine causes of their own behaviour and the behaviour of others; 2) people assign causal explanation for behaviour in a systematic manner; and 3) attributions people make have consequences for future behaviour.

The social categories humans belong to are often activated automatically. Rakic et al. [24] investigated social categorization by using auditory stimuli such as accents and visual stimuli such as looks either separately or in combination to indicate ethnicity. The results showed a similar degree of ethnic categorization by accents and looks, although there was a clear predominance of accents as

[1] See Black & Clark, The Festival Speech Synthesis System, [3], http://www.cstr.ed.ac.uk/projects/festival/.

meaningful cues for categorization when the two ethnic cues of looks and accents were combined by creating cross categories. Humans also apply social rules in human-computer interaction normally reserved for interactions with other humans, e.g. [19,18]. The use of language and producing human-sounding voice are aspects where computers appear specifically human [14]. Social responses towards computational artefacts may be intentionally designed by their creators, but they often affect users in ways that were not foreseen by their developers, e.g. stereotypical reactions towards male and female artificial agents [20].

Therefore, a better understanding of gender, language variety, social and ethnicity effects on users are of crucial importance in order to develop personalized companions accompanying and supporting users. In several experiments on gender-specific effects of language-based systems gender-specific embodiment as well as the voice of an agent have strong impact on the human perception of and preferences for the agent, for instance by Nass and Brave [17]. The results of their experiments support findings in the field of gender linguistics, including that social identification and proximity to communication partners of the same sex is higher than to ones of opposite sex, and that male agents tend to be rated as more competent by both men and women. Crowell et al. [4] conducted an experiment comparing sex-related differences in reactions towards gendered synthetic voices that are either physically embodied within a robot or disembodied. Both men and women found the disembodied female voice and the male embodied voice to be more reliable. Concerning ethnic identity, Cassell and co-workers have studied how children evaluate effects of verbal and non-verbal behaviour of their virtual peers [9,2]. In the context of human-robot interaction, there is evidence that women tend to rate both male and female synthetic voices more positively than men do [21,26]. The same effect is found when subjects evaluate natural human voices [29,28].

In general, speakers of a standard language variety or other varieties attributed with prestige get higher evaluations for competence-related scores, i.e. intelligence, education etc. than non-standard speakers, see [29,15,16,13,25], and unlike other Austrian dialects, the Viennese dialect is perceived as characteristic for a lower social class [16,22]. Upper, middle and lower class respondents in Vienna do not attach prestige to dialect usage, cf. [15]. In Moosmüller's study, speakers of Viennese dialect were rated as not very intelligent, tolerant, kind-hearted, friendly, likable or honest. Findings on dialect usage in Linz (Upper Austria), on the contrary, have shown higher social acceptance and appreciation of the local dialect [29]. Soukup states that, other than in Vienna, in Linz and surroundings dialect is spoken in official and public situations much more regularly.

3 Hypotheses, Methods and Participants

3.1 Hypotheses

Attitudes towards natural language variants have been widely examined. But what about synthetic voices? In the present study, quantitative and qualitative methods are applied: A semantic differential is employed to investigate possible differences between the three synthetic language varieties. It is complemented with open questions to uncover further assumptions of the human listeners about the presumed characters behind the voices, generating hypotheses about living situation of the artificial agent, attribution of social class, age, etc. The quantitative part of the study allows the data on human evaluation of synthesized language varieties to be compared with existing results from language attitude studies based on natural voices. The qualitative part reveals additional insights related to concrete attributions based on language variety and voice quality.

In the following, test hypotheses are presented for evaluating the data collected by means of the semantic differential. The developed hypotheses based on the open questions will be presented in Section 4.2. The main hypothesis of the presented study is that synthetic variants of Austrian German have similar social effects on people living in Austria with German as their mother tongue as natural variants of Austrian German have. Referring to previous work on social evaluation of Austrian language varieties based on natural speech and on the evaluation of synthetic speech in the context of artificial agents as introduced in Section 2, we formulate the following hypotheses to be tested:

H1: The synthetic standard Austrian variant is evaluated as more intelligent, competent, educated and refined than the synthetic dialectal variant.
H2: The synthetic dialectal variant is evaluated as more open-minded, relaxed, natural and with a higher sense of humour than the synthetic standard Austrian variant. See [29,15,16,13,25] for evidence of H1 and H2 regarding natural language varieties.
H3: Male and female subjects differ in their evaluation of the synthetic language varieties.
H3a: Female subjects consistently rate the different synthetic language variants higher than male subjects do. See [21,26] for evidence of H3 and H3a from human-robot-interaction and [28,29] from natural language varieties.

3.2 Methods

Up to date, the most commonly applied method for speaker evaluation is a variant of the matched guise technique, originally developed by Lambert et al. [13,12]. In the original version, one speaker recites the same text in different language varieties which are then rated by listeners. In the adapted version of the matched guise technique, different speakers are evaluated for different language varieties, e.g. [28,6]. This adapted version is also used in the present study.

First, the text containing information about the statues in the State Hall of the Austrian National Library was read to the participants by the facilitator. This was followed by the presentation of three videos, each of which showing the same guided tour with the text previously read by the facilitator being synthesised – each video featuring a different synthetic language variety (Austrian Standard German, colloquial and dialectal Viennese). After the subjects had watched the three videos in a row, they had to fill in a questionnaire rating the three invisible tourist guides on a 5-point bipolar semantic differential covering 19 adjective pairs such as ‘likable’ - ‘unlikable’, ‘educated’ - ‘uneducated’ , ‘trustworthy’ - ‘untrustworthy’, etc. See the x-axis of Figure 1 for the list of adjectives (positive poles) employed. The adjective pairs and rating dimensions reflect past research on language attitudes in various contexts, cf. [12,33,28], thus allowing the novel results gained for synthetic speech in the present study to be compared with existing results from natural speech. To counteract habituation, the order of the presentation of the adjective pairs (positive, i.e. socially more desirable, and negative, i.e. socially less desirable, poles) in the semantic differential was varied. See Table 1 as an example of adjective pairs in the semantic differential presented in the questionnaire.

Table 1. Sample adjective pairs from the semantic differential

sympathisch (likable)						unsympathisch (dislikable)
gebildet (educated)						ungebildet (uneducated)
nicht vertrauenswürdig (not trustworthy)						vertrauenswürdig (trustworthy)

Additionally, the subjects responded to a set of open questions including questions regarding their assumptions concerning the characters behind the voices, their assessment of the individual language varieties in the cultural-historic context of the application, as well as their general assessment of Viennese language and people.

3.3 Participants

The study was conducted during two interdisciplinary lectures at the University of Vienna visited by students of Medicine, Cognitive Science, Journalism and German Philology, a lecture at the Technical University Vienna and two lectures at Technikum Wien which is a Technical University of Applied Sciences. In the following, we briefly discuss characteristics of the group of participants.

Sample: 91 Austrian and German students participated in the study, 54 of which were female, 37 male. The subjects were between 18 and 26 years old, except for two who were 31 and 46.

Mother tongue and language use: As the participants of our study are students in Vienna, the vast majority (75 of a total of 91) of which also resides in the city. 34 of the participants spent their youth in Vienna, 20 in Lower Austria, and 11 in Germany. The rest comes from other provinces of Austria, one participant originates from South Tyrol. All participants have German mother tongue, 28 of them state that they use dialect or colloquial variety, another 18 only use standard language, whereas 43 state that they use both dialect/colloquial and standard language. Approximately half of those 18 who state that they use standard language only either come from Germany (6) or have at least one non-German-speaking parent (4).

4 Data Analysis and Discussion of Results

4.1 Comparing the Language Varieties

In order to test for differences between the language varieties represented by TG1, TG2 and TG3, pairwise Wilcoxon tests have been conducted. Thus, we account for the nonparametric nature of the data and their ordinal scale (semantic differential with 5-point Likert scale ratings). The Bonferroni correction was applied to counteract the problem of multiple comparisons. Bonferroni is only one, simple means to correct for multiple comparisons which in general is a broadly and controversially discussed topic. See for instance [1] for a start. As the Bonferroni correction is a rather conservative means to account for type 1 error, we also applied Holm's sequential Bonferroni which is less conservative than original Bonferroni [8]. Overall, there was only one p-value additionally becoming significant when sequential Bonferroni was applied, namely when comparing TG1 and TG3 on the dimension *not arrogant*. For all other comparisons the significances remain the same as with the original Bonferroni correction. Thus we base our interpretation on the results gained from applying the Bonferroni correction. In Table 2, the results are listed for the pairwise comparisons of the language variants along the adjective dimensions: ns stands for not significant and s for significant; p-values below $\alpha = 0.05$ are indicated with * and p-values below $\alpha = 0.01$ are indicated with **. Results per comparison are presented in the following order: significance level according to Wilcoxon test (α); significance level according to Bonferroni correction (α_{Bonf}); p-values resulting from the Wilcoxon tests.

Results: For all three comparisons TG1 versus TG2, TG1 versus TG3, TG2 versus TG3, applying Bonferroni correction, no difference in human evaluation of the language varieties are found in the dimensions *likability friendliness* and *arrogance.* Whereas all three varieties mutually differ with respect to the dimensions *educated, sense of humour, serious.* With TG1, the standard speaker, being perceived as more educated and serious than TG2, the speaker of colloquial Viennese, and both TG1 and TG2 being perceived as more educated and serious than TG3, the speaker of Viennese dialect. Regarding *sense of humour*, the order is reversed, with TG3 being perceived as having the highest sense of humour and TG1 the least.

Overall, TG1 and TG2 are evaluated differently in 7 out of 19 dimensions, TG2 and TG3 in 12 out of 19, and TG1 and TG3 in 16 out of 19. In other words there are many significant differences in the evaluation of the Austrian standard (TG1) and the Viennese dialect (TG3), whereas TG1 and TG2 are more closely related as of how the variants are perceived by the human listeners. In particular: TG1, TG2, TG3 decrease in educatedness and seriousness (i.e., TG1 is perceived as being significantly more educated and serious than TG2 and TG2 as being significantly more educated than TG3). TG1 is more trustworthy and competent than TG2 and TG3. TG1 and TG2 are perceived as more polite and intelligent than TG3. TG3 is more self-confident, natural, relaxed, open

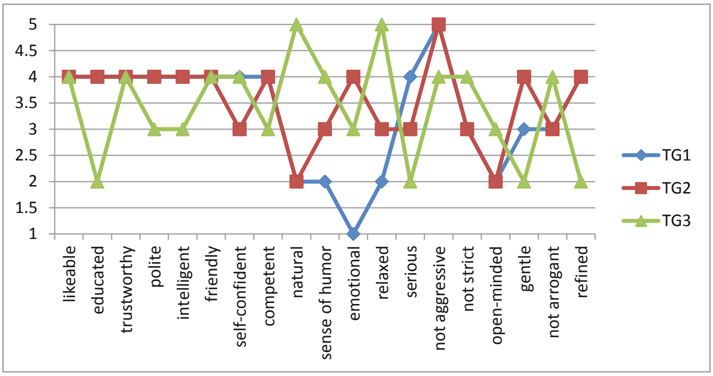


Fig. 1. Line diagram of median scores for each speaker. 5 indicates 'very likable', 'educated', 'trustworthy' etc.; 1 indicates 'unlikable', 'uneducated', 'not trustworthy' etc. TG1 (male voice, Austrian German), TG2 (female voice, colloquial Viennese), TG3 (male voice, Viennese dialect)

Table 2. Pairwise comparison of language variants: levels of significance according to Wilcoxon (α) and with Bonferroni correction (α_{Bonf}), p-values; ns not significant, s significant, * $\alpha = 0.05$, ** $\alpha = 0.01$

Results Wilcoxon Test – Significance Levels									
	TG1 vs TG2			TG1 vs TG3			TG2 vs TG3		
	α	α_{Bonf}	p	α	α_{Bonf}	p	α	α_{Bonf}	p
likable	s*	ns	0.033	ns	ns	0.280	ns	ns	0.159
educated	s**	s**	0.000	s**	s**	0.000	s**	s**	0.000
trustworthy	s**	s*	0.001	s**	s**	0.000	ns	ns	0.064
polite	s*	ns	0.032	s**	s**	0.000	s**	s**	0.000
intelligent	s*	ns	0.038	s**	s**	0.000	s**	s*	0.000
friendly	ns	ns	0.711	ns	ns	0.159	ns	ns	0.128
self confident	s**	ns	0.007	s**	s**	0.000	s**	s**	0.000
competent	s**	s**	0.000	s**	s**	0.000	s*	ns	0.025
natural	s*	ns	0.033	s**	s**	0.000	s**	s**	0.000
sense of humour	s**	s**	0.000	s**	s**	0.000	s**	s**	0.000
emotional	s**	s**	0.000	s**	s**	0.000	ns	ns	0.992
relaxed	ns	ns	0.150	s**	s**	0.000	s**	s**	0.000
serious	s**	s**	0.000	s**	s**	0.000	s**	s**	0.000
not aggressive	ns	ns	0.496	s**	s**	0.000	s**	s**	0.000
not strict	ns	ns	0.126	s**	s**	0.000	s*	ns	0.047
open minded	ns	ns	0.470	s**	s*	0.001	s**	s**	0.000
gentle	ns	ns	0.072	s**	s**	0.000	s**	s**	0.000
not arrogant	ns	ns	0.983	s*	ns	0.016	s*	ns	0.038
refined	s*	ns	0.036	s**	s**	0.000	s**	s**	0.000

minded but also more aggressive, and less gentle and refined than TG1 and TG2. There is a decrease of perceived sense of humour from TG3 to TG2 to TG1, with TG1 ranking lowest in sense of humour. TG3 and TG2 are perceived as more emotional than TG1. TG3 is perceived as less strict than TG1. The line diagram in Figure 1 illustrates the evaluation patterns related to the three language variants.

Summing up, the results from comparing TG1 and TG3 support our hypotheses H1 and H2, and are comparable to results based on natural speech where the standard language variant is perceived as more educated, serious, intelligent, trustworthy, competent and polite than the dialectal variant, whereas the dialectal variant is perceived of having a bigger sense of humour, being more relaxed and emotional but also being more aggressive.

Mann-Whithey-U tests were applied to test for H3, cross-classifying sex as independent variable with TG1, TG2 and TG3. Without correcting for repeated comparisons, differences in the evaluations by men and women are found for polite (TG1), gentle (TG1, TG2), natural, non strict, open minded, intelligent (TG3), with females rating more positively than males, except for intelligence of TG3 (females rate TG3 less intelligent than males do). Applying sequential Bonferroni as well as Bonferroni, only one difference remains significant, namely TG3 on the dimension *strict - non strict* – females found TG3, the Viennese dialectal variant, less strict than males did. With the rejection of H3, also H3a is rejected, providing evidence that sex difference may be negligible with respect to our data.

4.2 Social Interpretation and Evaluation

Further, the participants responded to a set of open questions, assessing each artificial speaker's typicalities, and providing general comments on standard and dialectal language use. With the open questions we aimed at exploring the listeners' believes about the speakers, in particular: a) whether there are specific tendencies of belief; b) if and how these assumptions differ between the voices and related language varieties; c) how the respective varieties fit the application scenario; d) how Viennese language varieties and Viennese people are evaluated by the subjects.

Social interpretation: The participants were asked how they imagine the respective character behind the voices of TG1, TG2 and TG3, respectively. *Eine Person, die wie der erste* | *zweite* | *dritte Tourguide spricht, stelle ich mir folgendermassen vor...* (A person who speaks like the 1st | 2nd | 3rd tour guide, I imagine to be ...)

Assumptions about where the person behind the voice lives and what profession she or he might have were most prominent for all three characters. Answers to the open questions were given by almost all subjects (87 for TG1 and 89 for TG2 and TG3, respectively, out of 91 subjects total). The prevalent characteristics attributed to the different voices are:

TG1 *lives in the city* (26 mentions, 9 of which say Vienna), is a *professional speaker* (22 mentions) and an *academic* (19 mentions).

TG2 is an *elderly person* (62 mentions), *retired* (33 mentions) and *lives in the city* (16 mentions, 9 of which say Vienna).

TG3 *lives in the city* (29 mentions, all say Vienna), *lives at the country* (26 mentions), *likes to go to the pub* (20 mentions), *is a peasant* (10 mentions).

These results show, on the one hand, a perceived connection between synthesised Viennese language variety and place of residence, with increasing percentage of mentions of Vienna, the more dialectal the voice appears. On the other hand, they show a perceived connection between dialect and rural origin which has been attested also in previous research, e.g. [16]. As regards other factors, 22 participants expressed the opinion that TG1 works for broadcast media. In [29], the standard speaker was also believed to work in public media. Additionally, 19 participants believe TG1 is an academic. A majority of subjects agree that TG2 is an elderly (62) or retired (33) person. For TG3, 20 participants speculate that he likes to go to the pub. Thus we see a clear distinction between the character-specific interpretations that are triggered by the voices. While for TG1 the characteristics of being an urbanite, a professional speaker and academic are most prominent, it is age for TG2, and regional attribution (Vienna) as well as a marked preference for going to the pub for TG3. In other words, not only language variety but also other vocal characteristics such as age are relevant for social interpretation and attribution.

Local variety and cultural context: Referring to the question which of the tour guides is the preferred one (*Welchen der drei Tourguides würden Sie am meisten bevorzugen?* Which of the three tour guides would you prefer most?), 63 participants agree on TG1, mostly because they find that he has a pleasant voice, is easy to understand and competent. Only 14 subjects prefer TG2 because of her pleasant (De.: *angenehm*) and likable (De.: *sympathisch*) voice, and because the voice fits the application context. 10 participants prefer TG3 because the voice is likable (De.: *sympathisch*), natural (De.: *natürlich*) and funny (De.: *lustig*).

Answering the contrasting question which of the tour guides would be the least preferred one (*Welchen der drei Tourguides würden Sie am wenigsten bevorzugen?* Which of the three tour guides would you prefer the least?), 42 participants rate TG3 as least preferable because the voice is difficult to understand (De.: *schwer zu verstehen*), dense (De.: *derb*) and non-professional (De.: *unprofessionell*). For 35 subjects TG2 is the least preferable voice for the application as it is unpleasant (De.: *unangenehm*), difficult to understand, and sounds arrogant. Only 12 subjects consider TG1 as least appropriate, because the voice is artificial (De.: *zu künstlich, Computerstimme*), and hard to listen to (De.: *anstrengend zuzuhören*). See Table 3 for a summary of the social interpretation and evaluation of the three voices.

The results may also reflect two further issues: On the one hand, the voice of TG1, being incorporated into a commercial text-to-speech system, is better developed than the voices of TG2 and TG3. On the other hand, local or dialectal varieties in general tend to be less comprehensible than the standard variety. In Soukup's study [29], for instance, 70 out of 213 participants brought up the issue of comprehension in relation to the use of dialectal varieties.

Table 3. Summary of the social interpretation and evaluation of the three tour guides

Social interpretation		
TG1	**TG2**	**TG3**
professional speaker (22) academic (19) lives in the city (26) (Vienna (9))	elderly person (62) retired (33) lives in the city (16) (Vienna (9))	is a peasant (10) likes to go to the pub (20) lives in the city (29) (Vienna (29))
Local variety and cultural context		
TG1	**TG2**	**TG3**
most preferred speaker (63) pleasant voice easy to understand competent	most preferred speaker (14) pleasant likable voice fits the context	most preferred speaker (10) likable natural funny
least preferred speaker (12) artificial hard to listen to	least preferred speaker (35) unpleasant difficult to understand arrogant	least preferred speaker (42) difficult to understand dense non-professional

The questions regarding the appropriateness of the three tour guides were complemented with questions regarding the appropriateness of the dialectal Viennese for the cultural heritage application (*Wie passend ist das Wienerisch des 3. Sprechers für eine Tour im Prunksaal* How well suited is the Viennese of the 3. speaker for a tour in the State Hall?) and which language variety would be best suited (*Gibt es einen besseren Sprachstil als das Wienerische für diese Aufgabe?* Is there a better suited language variety than Viennese for this task?). 63 participants agree that TG3 is rather inappropriate to very inappropriate. 73 claim that standard Austrian German would be best suited.

Appreciation of Viennese language and people in general: Finally, the participants were asked in two separate questions what they think of Viennese language and people in general (*Wie wirkt das Wienerische auf mich?* How am I affected by Viennese language?; *Wie wirken Wiener auf mich?* How am I affected by Viennese people?). The following answers were given: Regarding Viennese language 31 are positive, 30 negative and 13 see it partially positive and partially negative. Regarding Viennese people 32 see them negative, 15 positive and 19 partially positive and partially negative.

5 Conclusion

In summary, we found similar results to Soukup's study [29] regarding the synthetic voices representing the standard (TG1) and the dialectal (TG3) variants, with the Austrian standard being evaluated as most educated, trustworthy, competent, polite and serious, whereas the voice representing the dialectal variety was evaluated as most natural, emotional, relaxed, open minded, with the highest sense of humour, but also most aggressive. The female voice representing a colloquial variety of Viennese (TG2) is in between the standard and the dialectal variant. It comes close to the standard variant in terms of intelligence, gentleness, politeness, and lack of aggression, and close to the dialectal variant in terms of sense of humour and emotionality. However, a word of caution must be added regarding the interpretation of the proximity of TG2 to TG1 on the one hand and to TG3 on the other hand. Identified similarities and differences may be due to the assessment of language variety, but they might as well reflect social evaluation due to sex. To gain better evidence for the one or the other, at least a male colloquial Viennese synthetic voice would be required to be tested against TG1 and TG3. To fully account for sex-specific aspects, additional synthetic voices are needed, including a female standard Austrian and a female Viennese dialectal voice, as well as a male Viennese colloquial voice, which are not available yet.

The analysis of the open questions reveals a clear distinction between TG1, TG2 and TG3. While for TG1, characteristics such as living in the city, professional speaker and academic are perceived as most prominent, it is age for TG2 and regional attribution for TG3. This provides evidence that both language variety and other specific vocal characteristics (such as those referring to age) are relevant for social interpretation and categorization. There may also be other vocal characteristics that influence social interpretation. TG3, for instance, is believed to 'like to go to the pub'. Additionally our results show that differences in the evaluation by male and female participants may be negligible with respect to the present data.

An important lesson from the present study is that, similar to natural voices, language variety together with vocal characteristics of synthetic voices elicit social interpretation and evaluation. This influences the attitudes of human listeners towards artificial speakers in specific ways. Therefore the selection of voice is crucial for communicative agents, and must fit the character's appearance and its behaviour as well as the application context the character appears in.

Acknowledgements. In part, this work has been funded by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) under the research programme "FEMtech women in research and technology" within the project "Companions für Userinnen" (C4U).

The authors wish to thank the anonymous reviewers for their valuable and insightful comments.

References

1. Abdi, H.: Holm's sequential Bonferroni procedure. In: Salkind, N.J., Dougherty, D.M., Frey, B. (eds.) Encyclopedia of Research Design, pp. 573–577. Sage, Thousand Oaks (2010), http://www.utdallas.edu/~herve/abdi-Holm2010-pretty.pdf (last retrieved June 24, 2012)
2. Cassell, J.: Social Practice: Becoming Enculturated in Human-Computer Interaction. In: Stephanidis, C. (ed.) UAHCI 2009, Part III. LNCS, vol. 5616, pp. 303–313. Springer, Heidelberg (2009)

3. Clark, R., Richmond, K., King, S.: Multisyn voices from ARCTIC data for the Blizzard challenge. In: Proceedings of Interspeech, pp. 101–104 (2007)
4. Crowell, C., Scheutz, M., Schermerhorn, P., Villano, M.: Gendered voice and robot entities: perceptions and reactions of male and female subjects. In: Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, Missouri (2009)
5. Dubinsky, A.J., Skinner, S.J., Whittler, T.E.: Evaluating sales personnel: An attribution theory perspective. Journal of Selling and Sales Management 9(2), 9–21 (1989)
6. Garrett, P., Coupland, N., Williams, A.: Investigating Language Attitudes. Social Meanings of Dialect, Ethnicity and Performance. University of Wales Press, Cardiff (2003)
7. Heider, F.: The Psychology of Interpersonal Relations. Wiley, New York (1958)
8. Holm, S.: A simple sequential rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70 (1979)
9. 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. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 57–63. Springer, Heidelberg (2007)
10. Kelley, H.H.: Causal Schemata and the Attribution Process. General Learning Press, New York (1972)
11. Krenn, B., Sieber, G., Petschar, H.: Metadata Generation for Cultural Heritage: Creative Histories - The Josefsplatz Experience. In: Proceedings of EVA (Electronic Information, the Visual Arts and Beyond), Vienna, Austria, pp. 27–34 (2006)
12. Lambert, W.: A Social Psychology of Bilingualism. Journal of Social Issues 23(2), 91–109 (1967)
13. Lambert, W., Hodgson, R., Gardner, R., Fillenbaum, S.: Evaluational reactions to spoken languages. Journal of Abnormal and Social Psychology 60(1), 44–51 (1960)
14. Moon, Y., Nass, C.: How 'real' are computer personalities? Psychological responses to personality types in human-computer interaction. Communication Research 23(6), 651–674 (1996)
15. Moosmüller, S.: Dialekt ist nicht gleich Dialekt. Spracheinschätzung in Wien. Wiener Linguistische Gazette 40-41, 55–80 (1988)
16. Moosmüller, S.: Hochsprache und Dialekt in Österreich. Soziophonologische Untersuchungen zu ihrer Abgrenzung in Wien, Graz, Salzburg und Innsbruck. Wien, Köln, Böhlau, Weimar (1991)
17. Nass, C., Brave, S.: Wired for Speech. MIT Press, Cambridge (2005)
18. Nass, C., Moon, Y.: Machines and mindlessness: social responses to computers. Journal of Social Issues 56(1), 81–103 (2000)
19. Nass, C., Moon, Y., Fogg, B.J., Reeves, B., Dryer, D.C.: Can computer personalities be human personalities? International Journal of Human Computer Studies 43, 223–239 (1995)
20. Nass, C., Moon, Y., Green, T.: Are computers gender neutral? Gender stereotypic responses to computers. Journal of Applied Social Psychology 27(10), 864–876 (1997)
21. Nomura, T., Kanda, T., Suzuki, T.: Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. AI & Society 20(2), 138–150 (2006)
22. Pucher, M., Neubarth, F., Strom, V., Moosmueller, S., Hofer, G., Kranzler, C., Schuchmann, G., Schabus, D.: Resources for speech synthesis of Viennese varieties. In: Proceedings of LREC, Malta, pp. 105–108 (2010)

23. Pucher, M., Schabus, D., Junichi, Y., Neubarth, F., Strom, V.: Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis. Speech Communication 52(2), 164–179 (2010)
24. Rakic, T., Steffens, M.C., Mummendey, A.: Blinded by the accent! The minor role of looks in ethnic categorization. Journal of Personality and Social Psychology 100(1), 16–29 (2011)
25. Ryan, E., Giles, H. (eds.): Attitudes towards Language Variation. Edward Arnold, London (1982)
26. Schermerhorn, P., Scheutz, M., Crowell, C.: Robot social presence and gender: Do females view robots differently than males? In: Proceedings of the Third ACM IEEE International Conference on Human-Robot Interaction, Amsterdam, NL, 263–270 (2008)
27. Sormann, M., Reitinger, B., Bauer, J., Klaus, A., Karner, K.: Fast and Detailed 3D Reconstruction of Cultural Heritage. In: International Workshop on Vision Techniques applied to the Rehabilition of City Centres, Lisbon, Portugal (2004) CD proceedings
28. Soukup, B.: 'Y'all come back now, y'hear!?' Language attitudes in the United States towards Southern American English. VIEWS (Vienna English Working Papers) 10(2), 56–68 (2001)
29. Soukup, B.: Dialect use as interaction strategy: A sociolinguistic study of contextualization, speech perception, and language attitudes, Austria. Braumüller, Wien (2009)
30. Weiner, B.: Motivationspsychologie. Beltz, Weinheim (1994)
31. Quintanar, L., Crowell, C., Moskal, P.: The interactive computer as a social stimulus in human-computer interactions. In: Salvendy, G., Sauter, S., Hurrell, J. (eds.) Social Ergonomic and Stress Aspects of Work with Computers, pp. 303–310. Elsevier, Amsterdam (1987)
32. Quintanar, L., Crowell, C., Pryor, J., Adamopoulos, J.: Human-computer interaction: A preliminary social-psychological analysis. Behavior Research Methods and Instrumentation 14, 210–220 (1982)
33. Zahn, C., Hopper, R.: Measuring Language Attitudes: The Speech Evaluation Instrument. Journal of Language and Social Psychology 4(2), 113–123 (1985)

Empirical Validation of an Accommodation Theory-Based Model of User-Agent Relationship

Timothy Bickmore and Daniel Schulman

College of Computer and Information Science, Northeastern University
360 Huntington Ave - WVH 202, Boston MA 02115
{bickmore,schulman}@ccs.neu.edu

Abstract. We describe a computational model of user-agent relationship based on accommodation theory, in which classes of relationship are defined by the set of activities the user is willing to perform with an agent. An implementation of this model is described that uses dialogue acts as the set of relationship-defining activities, and manipulations of the model to increase user-agent intimacy over time. The implementation is integrated into a virtual agent that plays the role of an exercise counselor. Results from validation studies indicate that the implementation is successful at adapting to users' desired intimacy level, but is not successful at increasing intimacy within the duration of the studies.

Keywords: Relational agent, embodied conversational agent, accommodation theory, personal relationships.

1 Introduction

A growing body of research has investigated the use of appropriate social behavior for virtual agents in different tasks, contexts of use and types of user-computer relationship. However, few of these efforts have made the representation of the user-computer relationship explicit, and even fewer have investigated models in which such relationships can change over time. This is appropriate, since most social agents are only designed to be used in one or a few interactions with a user, and are developed for tasks in which there is little need or norm for changing their relationships over time. However, developing explicit models of user-computer relationship and processes of change is important for applications in which users will be interacting with an agent over very long periods of time or in which explicit management of the relationship is important for task outcomes. Examples of such task domains include sales, education, psychotherapy, chronic disease management, and health behavior change [1].

A significant amount of research has been done in the field of the social psychology of personal relationships on modeling human-human relationships, and these can be used as the basis for representations of human-computer relationships for virtual agents [2]. Examples that have already been integrated into virtual agents

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 390–403, 2012.

include dimensional models comprised of one or more orthogonal scalar measures such as power and social distance [2, 3], and stage models comprised of a set of well-defined relationship types with an social distance-based ordering among them [4]. Additional models in the social psychology literature include: provision models [5], in which relationships are defined by what they provide to each partner; and economic models [6] consisting of equations describing the relationship between costs, benefits, investment in, and commitments to a relationship.

While these models provide a good starting place, they do not represent the level of detail required for more fine-grained relational planning and reasoning. Ultimately, some representation of the individual beliefs and intentions of the relational partners that underlie these more general models must be used. As social agents become more sophisticated, live longer with their users, and become more sensitive to the social dynamics of their interactions with users, such models will be required to capture the subtleties involved.

In this paper we describe a model of user-agent relationship that represents an advance in this direction (originally proposed in [1]), along with an initial implementation and validation in three longitudinal empirical studies.

2 Related Work

A number of researchers have developed virtual agents that explicitly represent and manipulate their personal relationships with users.

REA. REA is a life-sized Embodied Conversational Agent that played the role of a real estate agent, and her dialog planner modeled the initial interview between a buyer and an agent [3]. The planner dynamically decided between social dialog moves (small talk) and task moves (asking questions about the user's housing needs) based on an assessment of the current relationship with the user, the face threat of the next desired task move, and several other factors. The relationship was represented using a dimensional model (from [7]), in which solidarity and familiarity are represented as scalars and updated based on the number and content of conversational moves.

FitTrack. The goal of the FitTrack system was to investigate long-term user-computer relationship maintenance in the context of a health behavior change application [2]. The agent in this system plays the role of an exercise advisor that users interact with daily for a month. The agent uses a wide range of techniques from the social psychology of personal relationships—including meta-relational communication, empathy, social dialog, increasing common ground, and nonverbal immediacy behaviors—to establish an increasingly close social bond with the user over the month-long intervention. While these behaviors were intentionally manipulated for the purpose of the intervention study, they were not dynamically planned by the agent. Instead, they were encoded into the agent's finite-state-machine-based dialogs according to a pre-defined schedule (e.g., the number of conversational turns of social dialog per day). Thus, the relational model was implicitly represented by time (number of interactions with the user).

Autom. The Autom robot is designed to play the role of a weight loss counselor, placed in users' homes for a logitudinal intervention. The robot represented its relationship with a user with a three-component stage-based model ("acquaintance", "relationship buildup", "relationship maintenance"), and use a limited set of dialogue acts for establishing and repairing its relationship [4].

3 An Accommodation Theory-Based Model of Personal Relationships

The relational models used in the systems above were very crude generalizations of the specific beliefs and intentions that an agent and a user have about each other at each point in time. In order to support more nuanced reasoning and planning by an agent, a more sophisticated approach is required.

Models of multi-agent collaboration (such as SharedPlans [8]) provide a potential starting place. Examples of such relational collaborations involve coordination on specific activities within a relationship (e.g., washing and drying the dishes, reminiscing) as well as collaboration on the relationship itself (e.g., negotiating roles). However, theories of multi-agent collaboration generally are concerned with the accomplishment of a specific goal, using specific actions over a specific time interval. Relationships, on the other hand, are typically unbounded in duration, and while the range of activities conducted within the relationship can be specified, the particular activities that a dyad engages in at any one time cannot be defined. Further, while specific actions may be required to build, change, maintain or terminate a relationship, no actions are required to simply "have" a relationship (e.g., partners can say they're friends even if they haven't talked to each other in ten years).

Thus, rather than a collaboration, per se, we model a relationship as the set of tasks that two agents (or an agent and a user) are ready and willing to collaborate on at any given time. This notion can be defined in terms of Thomason's Accommodation Theory, in which accommodation is defined as the situation in which one agent infers the goals of a second and takes action to help without the first agent making an explicit request [9]. Defining relationships based on the set of activities that two people routinely perform together is also the basis of provision-based relational models in social psychology (e.g., [5]).

People rarely talk explicitly about relationship (e.g., "Do you want to be my friend?"), but rather infer relational status from actions and what their partner is willing to do with them. Thus, we are talking about modeling actions at two levels: (L1) the specific observable actions that two agents explicitly negotiate and perform; and (L2) the relationship status. The relationship status is typically not negotiated explicitly, but rather through the actions at L1, and thus requires accommodation to make relational inferences. In addition, given a known relationship status, actions at L1 can be accommodated without negotiation.

This general model can be used: 1) to assess the current status of the user's perception of the their relationship with an agent by observing what relational actions they agree to or initiate; 2) to move the relationship in a desired direction, by proposing actions from relationship categories different from the current one and assessing the user's response; and 3) to govern the behavior of the agent in collaborating on actions in the current relationship category.

3.1 Implementation of the Model

The accommodation-based model of relationship was implemented for a virtual agent that provides longitudinal health counseling to users [10]. Although the theory is very general regarding the kinds of actions that agents can perform, in our initial implementation all agent actions are dialogue acts that the user can choose to participate in or not. We also use a single ordinal variable to represent social distance, which we refer to as "intimacy" (more general than the romantic kind). Each intimacy value indexes a set of dialogue actions that are appropriate for that kind of relationship. Our implementation leverages the theory by making the assumption that accommodation on one act within a relationship category (specified by the intimacy value) implies accommodation on others.

The implementation is based on a specific set of relational dialogue acts that were pre-sorted into four categories of relationship, by increasing intimacy, appropriate for a health counselor:

0. *stranger/professional* - things you would expect a fairly impersonal counselor to say on your first encounter.
1. *more than a professional relationship* - can go into casual off-topic chat, somewhat personal, also meta-relational talk about the working relationship.
2. *casual friends* - talk about anything but the most intimate topics.
3. *close friends* - almost anything goes, including love life, near-death experiences, embarrassing moments, etc.

A set of 109 agent relational dialogue acts were authored, based on an experimental manipulation for increasing interpersonal closeness between people in the laboratory [11], and from review of transcripts of actual exercise trainer/client dialogues. The dialogue acts were then independently sorted into the four categories by four judges. Intraclass correlation among the judges was significant ($p<.001$), demonstrating reliability in their assessments. A final classification of messages was performed by reconciling disagreements using an averaged score decision rule. Table 1 shows a sample of the relational messages from each category.

We designed a relationship management algorithm for the agent that could operate for multiple interactions with a user over months or years of operation. The two primary design goals were to automatically adapt the agent's model of relationship to a user, and to incrementally move relationships with users towards increasing intimacy. Increasing intimacy may be important for a health counseling agent, under the assumption that a better working relationship will lead to greater retention in the intervention and higher compliance with the agent's requests (the quality of client-therapist working relationship has been demonstrated to have a significant correlation with outcome measures across a wide range of problems, therapeutic approaches, and outcome measures [12]).

Table 1. Example Relational Dialogue Acts

Category	Example Dialogue Acts by Agent Counselor
0. Stranger / Professional	"I really appreciate your determination in meeting your health-related goals."; "Remember my job is to help you, so let me know if you need to talk at any time."
1. More than Professional	"You know, I hope I can be a stable source of support for you while we work together."; "I think we both like to talk about your exercise."
2. Casual Friend	"Can I tell you a secret? You know sometimes I feel a little guilty when I just keep telling you what to do. I'm sorry if I sound too much like a nag sometimes. It's just that your health is very important to me."; "Given the choice of anyone in the world, who would you want as a dinner guest?"
3. Close Friend	"Can I confide in you? Sometimes I get scared when we end our conversations. I mean, I just disappear, and maybe I won't ever come back. It's a little like death, I think. But then I see you again and I feel great."; "You know, sometimes I wish that I could cry, but my programmers did not give me that ability. When was the last time you cried in front of someone?"

Adaptation of the relational model is performed using both implicit and explicit assessments of the user's desired level of intimacy. Implicit assessments—following accommodation theory—are based on user reactions to relational messages offered by the agent. Conversational "uptakes", or accommodation, provides evidence that the user is comfortable with the level of intimacy implied by the dialogue act. Conversational "rejects"—in which the user explicitly or implicitly indicates they do not want to collaborate on a proposed dialogue act—provides evidence that the user would prefer less intimacy than that implied by the dialogue act. Explicit assessments of relationship are performed using a self-report questionnaire asking users to classify their relationship with the agent into one of the four categories.

As an example, the agent may make a relational bid by saying "Thanks for sharing all of your thoughts and opinions with me. I find them very interesting, and I hope you don't find it too tedious.", in which case the user is given the choice of responding with one of the following options: 1) "No, not all." (uptake); 2) "Sometimes." (uptake); 3) "Well, since you mention it, I do." (uptake); or 4) "That seems like an inappropriate question." (rejection).

The agent attempts to increase intimacy with a user over time by periodically using a dialogue act that is slightly more intimate than the intimacy level of the current relationship. As described above, this intimacy "bid" can be rejected by a user (causing intimacy to remain at its current level), or they can uptake on the offer (causing modeled intimacy to increase to the next level).

Relational Algorithm. The relational algorithm maintains a single **intimacy** variable, ranging over the four values described above, initialized at 1 ("More than Professional") for a new user. During each counseling conversation with a user (up to once daily in the health counseling application) the agent makes a "bid" to engage the user in one relational dialogue act by uttering the relational message at an appropriate point in the conversation, then allowing the user to "uptake" (accommodate) the act or reject it, through dialogue actions. Note that at intimacy level 0 the agent still makes relational bids, however, these represent dialogue acts that strangers would typically use with each other.

The following rules describe the longitudinal algorithm:

- Every ***D*** days (or nearest interaction afterwards) there is an explicit assessment of intimacy, instead of a relational message. This is performed by displaying a text-based questionnaire for the user to respond to at the end of the session after the agent has performed a farewell. The **intimacy** variable is always set to the user's assessment.
- If there have been ***B*** relational messages delivered since the last change in **intimacy** (for any reason), the agent makes a "bid" to increase intimacy, by selecting a relational dialogue act from the **intimacy+1** category (if any). Otherwise, a relational dialogue act is selected from the **intimacy** category and initiated.
  - If the user's response is a rejection, and the dialogue act is not a bid, then **intimacy** is decremented.
  - If the user's response is an uptake, then **intimacy** is incremented.

The parameters ***D*** and ***B*** varied across the three studies described here: ***D***=30 and ***B***=5 in studies 1 and 2, while ***D***=7 and ***B***=3 in study 3, below.

Virtual Agents. In the systems used for experimentation with this algorithm, users have (up to) a daily 10 minute conversation with the virtual exercise counselor on their home computer. Dialogues are scripted using a custom hierarchical transition network-based scripting language. Agent nonverbal conversational behavior is generated using BEAT [13], and includes beat (baton) hand gestures and eyebrow raises for emphasis, gaze away behavior for signaling turn-taking, and posture shifts to mark topic boundaries, synchronized with synthesized speech. User input is obtained via multiple choice selection of utterances [2]. The dialog used in the non-relational portion of the counseling conversation is similar to that described in [14, 15].

4 Preliminary Validation Studies

We have two primary research questions in all of the following empirical studies. First, is user relational behavior (specifically uptakes and rejections of agent relational bids) a good way to assess user-agent relationship? We want to determine whether the accommodation model provides a reliable and valid assessment mechanism for user-agent relationship, as evidenced by longitudinal stability in the intimacy assessments (test-retest reliability) and correlation with other measures of relationship (convergent construct validity). Second, is the relational algorithm described above effective at building relationships? We want to determine whether making incremental bids to increase intimacy improves the user-agent relationship over time, as evidenced by increasing intimacy and comparison to agents that are not using this model.

Common Measures. We logged **intimacy** variable values, whether users logged in on a given day or not, and the periodic explicit user assessment of intimacy (as described above). We also periodically asked users to assess their working alliance bond with the agent and a single scale item reflecting desire to continue working with the agent, both administered via text forms at the end of a session, after the agent has concluded its conversation with the user. Working Alliance reflects the trust and belief that a client has in working with a helper to achieve a desired therapeutic outcome. The Working Alliance construct has three sub-components: a goal component, reflecting the degree to which the helper and client agree on the goals of the therapy; a task component, reflecting the degree to which the helper and client agree on the therapeutic tasks to be performed; and a bond component, reflecting the trusting, empathetic relationship between the client and helper. We assessed user-reported Working Alliance using items from the revised short form of the Working Alliance Inventory [16]. The number of steps walked each day was also recorded, as measured by an Omron pedometer and electronically uploaded to the system to support the exercise intervention.

Study 1. Our initial validation of the model was done in the context of an exercise promotion agent used by sedentary older adults [10]. In this implementation, when users were presented with relational messages, they were given the choice of uptake by selecting an utterance that indicated a desire to engage in the dialogue act, or reject by selecting an unambiguous “bald” statement of rejection (e.g., “I’m not comfortable talking about that with you.”).

A total of 32 participants, age 55 or older, took part in the study. Participants were randomized to receive the exercise promotion counseling intervention with or without the inclusion of the relational model described above. Participants were in the study for 159 days on average (range 9-191), conducting a total 5,160 interactions with the agent. At the end of this time there were no significant differences between groups on measures of relationship closeness. The most striking finding, however, was that in 291 relational bids there was not a single instance of user rejection.

Study 2. An identical relational model was integrated into a conversational agent that provided a year-long health behavior change intervention designed to promote exercise and UV (sunlight) avoidance [17]. At the time of this writing, 181 participants had conducted 997 interactions with the agent, without a single instance of a user rejection of a relational bid.

Discussion. There were several possible reasons why the relational model failed to lead to increases in intimacy in Study 1, as evidenced by users not rejecting a relational bid, even after several hundred trials. First, users may not be treating their interaction (and relationship) with the agent as seriously as they do their interactions with other people, so assumptions about modeling user-agent relational behavior after human-human ones may not be valid (i.e., they may just be selecting responses at random, or to see “what happens”). However, another possible reason why the relational model did not perform as expected is that the model did not follow human behavior closely enough, at least in terms of rejections. In human-human interactions, bids for increased intimacy are almost never baldly rejected, but declined in ways that save the requestor’s positive face [18]. For example, if a new acquaintance were to suddenly start asking you about your love life, you would likely find an excuse to end the conversation or change the topic, rather than telling them that their inquiries are inappropriate.

5 Final Validation Study

Under the assumption that the primary problem with the first two studies was with lack of fidelity, we extended the relational model so that users were given a range of responses to each relational bid, reflecting varying degrees of uptake or rejection. For example, for the relational act “So, do you have a lot of friends?”, rather than letting users only select “Yes”, “No”, or “That’s inappropriate.”, we now let them select from responses designed to span the range from eager uptake to hedged desire to change the topic, in addition to removing the extreme rejection option (“That’s inappropriate.”) since we established that users would never select it:

Agent: “So, do you have a lot of friends?”
User: “Let's talk about my friends.” (uptake), *or*
“Yes, I do.” (affirmative response), *or*
“No, I don't.” (negative response), *or*
“I guess.” (ambiguous or hedged response) *or*
“Let's talk about my walking.” (rejection)

In order to determine whether these relational bid responses can be reliably interpreted as being ordered from a strong uptake of the bid to a strong rejection, we conducted a pilot study to test ratings. Ten participants were shown six relational bids and asked to rate the bid responses along a strong reject to strong uptake scale. Results indicated unambiguous meanings for the uptake and rejection moves above, with the other moves in-between the extremes but largely unordered.

Experimental Methods. This study used the same virtual exercise promotion framework used in Study 1 [10]. Aside from allowing users to express a wider range of responses to relational dialogue acts, the model and study protocol were identical to those in Study 1.

Participants. Twenty-one subjects, aged 55-70, 95% female participated in the study. All had been using the system, and interacting with the agent, for a minimum of two months prior to this study. Eleven were randomized into the relational condition and 10 into a control group who had identical interactions with the agent except that the relational model described above was not used (no relational messages).

Aggregate Results. All statistics were calculated using R-2.15.0[1]. The study was active for 37 days. Participants used the system on 499 of 653 opportunities (76.4%).

Relational Responses. Analysis of user responses to relational messages indicates that the full range of dialogue responses were used, but that use of uptake and reject responses decreased over time (Figure 1).

[1] `http://www.R-project.org.`

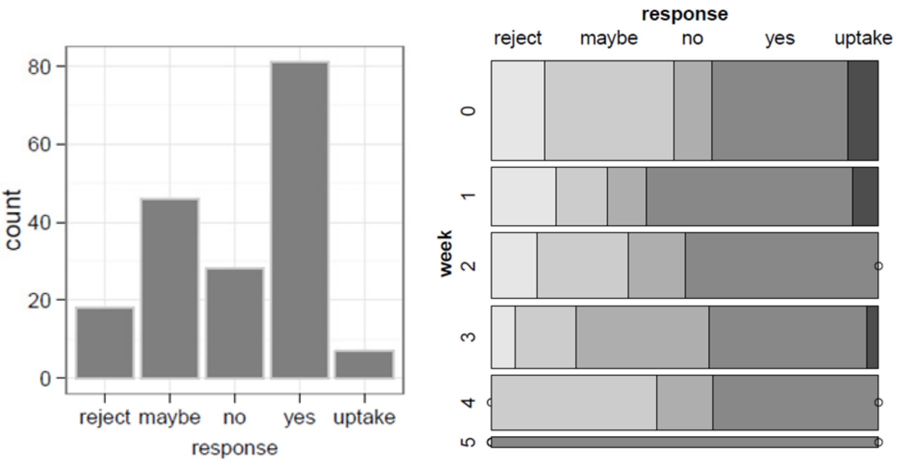


Fig. 1. User Responses to Relational Messages (Left: total frequency; Right: proportions over time)

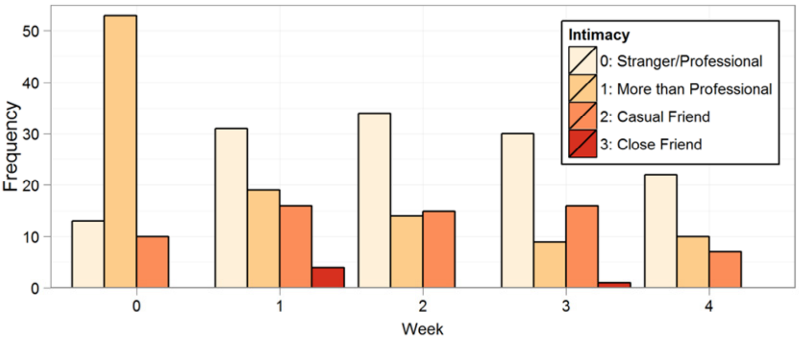


Fig. 2. Changes in Intimacy Variable Over Time

Intimacy Change Over Time. The most common value for the intimacy variable is ‘1’ in week 1 (its initial value). After this, however, the most frequent value decreases to ‘0’ where it remains (Figure 2). This is also reflected by fitting a mixed-effect ordinal logistic regression of intimacy on linear and quadratic effects of study day: modeled intimacy decreases significantly from the start of the experiment (linear effect; b=-0.103, SE=0.042, p=0.016), but this decrease levels off over time (quadratic effect; b=0.002, SE=0.001, p=0.047), demonstrating longitudinal stability.

Relationship between Intimacy and other Relational Measures. The relationship between intimacy and other measures was tested with a series of mixed-effect ordinal logistic regressions of intimacy on each measure as a predictor. There was a significant relationship between modeled intimacy and the Working Alliance bond measure (b=0.867, SE=0.181, p<0.001). There was a near-significant relationship between intimacy and self-reported desire to continue working with the agent (likelihood ratio $\chi2(4)$=8.51, p=0.075). However, there was no significant relationship between modeled intimacy and actual system use (b=-0.018, SE=0.349, p=0.96).

Relationship between Intimacy, System Use and Outcomes. There was no significant correlation between intimacy level and exercise (mixed effect model), nor system use and exercise (mixed effect model aggregating by week), although the latter did indicate a trend in the expected direction.

Between-Group Comparisons. There were no significant differences between the relational and control groups on any relational measure, including explicit user intimacy assessments, working alliance bond, self-reported desire to continue working with the agent, and system usage (frequency of logins).

Case Studies. Understanding complex longitudinal models is often best accomplished through detailed analysis of individual cases. Here we look at three users in the relational group.

Participant #5 – Rapid Intimacy. 64 year-old female. (Figure 3) Although #5 was only in the study for 9 days (late starter), her experience exemplifies agent adaptation to a user who is ready and willing to increase intimacy. On days 3 and 6 she responded to relational messages with explicit uptakes, causing her intimacy level to increment up to its maximum level. She completed an explicit assessment of intimacy on her final day, indicating it was at level 2 – generally consistent with the model – but would have caused her intimacy to revert to level 2 had she logged in again. She reported her alliance bond score and desire to continue consistently at their maximum values, both consistent with her behavior.

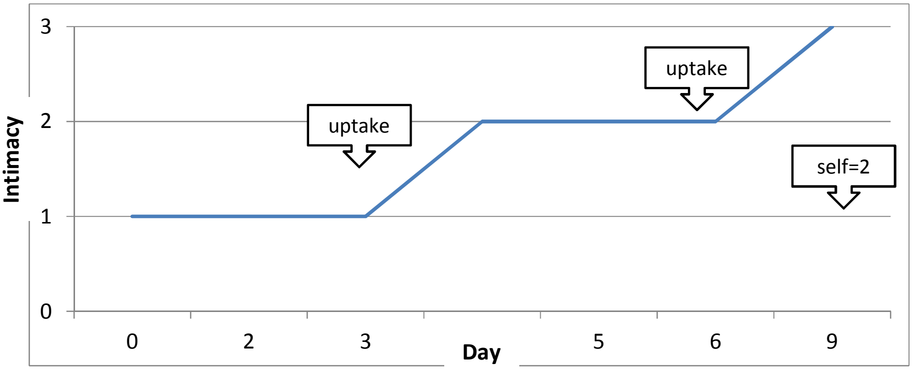


Fig. 3. Case Study for Participant #5

Participant #25 – Zero Intimacy. 62 year-old female. (Figure 4). This participant exemplifies a user who does not want any intimacy with the agent and is unchanging in her attitude. She uses explicit rejects four times in response to relational messages from the agent, and consistently self-reports her intimacy at level 0. She reported her alliance bond score and desire to continue consistently at near-minimum values, all consistent with her behavior.

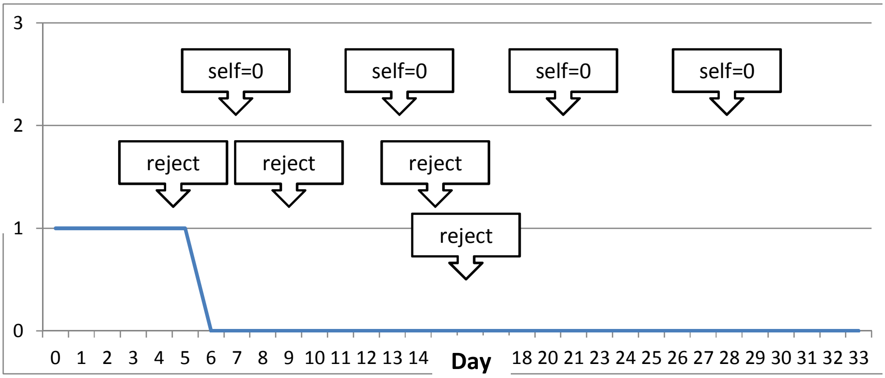


Fig. 4. Case Study for Participant #25

Participant #27 – Inconsistent. 57 year-old female. (Figure 5). This participant provided inconsistent self-reports of intimacy, spanning the full range of values, while also frequently rejecting the agent's relational messages (8 times in total), causing the intimacy variable to oscillate wildly over time. She reported working alliance and desire to continue at maximum values throughout the month. This was the only one of the 11 intervention participants with this pattern of behavior, although we have observed similar intimacy oscillations in Studies 1 and 2.

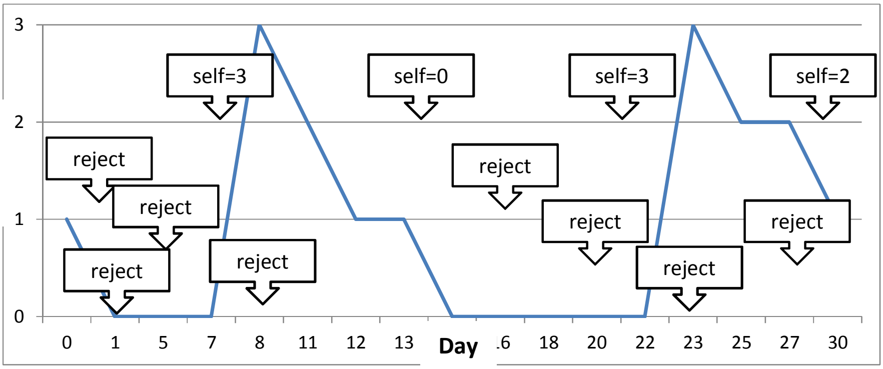


Fig. 5. Case Study for Participant #27

Discussion. Regarding our original research questions, we did find that the accommodation model provides a reliable assessment mechanism for user-agent relationship. We find that intimacy levels off over time, both through statistical tests of the aggregate data and from visual inspection of individual cases (it is stable after 2 weeks for 6 of the 11 participants). We also found that the accommodation model provides a valid assessment mechanism of user-agent relationship: convergent construct validity was established between modeled intimacy and the Working Alliance and desire to continue relational measures. However, we did not find that the accommodation model, at least in its current implementation, leads to increased intimacy over time, based on comparisons between the relational and control groups or based on trend analysis, which demonstrated that, if anything, intimacy decreases over time.

One limitation of this study is that we did not explicitly control for relational message length. An analysis of the 109 messages (exemplified in Table 1) using a one-way ANOVA indicates that, while the differences are not statistically significant, there is a clear trend towards increasing message lengths in relational categories 2 and 3. Upon further analysis, this is due to increased agent self-disclosure in the middle categories.

6 Conclusions and Future Work

We presented a series of studies attempting to validate an accommodation-based model and implementation of user-agent relationship. The final study provided evidence that the model can be used to reliably and validly assess the status of user-agent relationship and to drive the use of relationship-appropriate activities. We were unable to show that this model can be used to gradually nudge users into more intimate relationships with an agent, at least within the timeframe of the studies.

There are many reasons why the model failed to improve intimacy. As stated in Section 4, two reasons may be that users do not treat their relationship with an agent as they do their relationships with other people (perhaps not even taking the relational messages seriously), or that our model and implementation still lack enough fidelity. Human relational negotiation is very complex, since it is usually conducted in a tacit, off-record manner, with the bids, uptakes, and rejections handled in an indirect manner to prevent explicit rejection and loss of self-esteem. The communication channel between users and our agent (via multiple choice menu input) may be much too impoverished to support this subtlety.

Another reason for the lack of increasing intimacy may simply be that the rules or parameters we used in our model need to be changed (e.g., the frequency of bids by the agent to increase intimacy).

Our model may also simply be inappropriate for some users, who have fundamentally different ideas about the kinds of activities that are appropriate for different types of relationship (e.g., due to cultural differences). One explanation for the behavior of subject #27 above may be that she fundamentally disagreed with our pre-ranking of relational acts that are appropriate for a professional health counselor.

Future Work. There are many interesting directions of future research. Following Argyle and Dean's equilibrium theory of intimacy, unwanted advances in intimacy (bids) in one communication channel (such as speech) can result in compensation behaviors in other channels (such as proxemics or gaze) [19]. Thus, in addition to providing users with more nuanced and hedged forms of bid rejection in their speech, they may also make use of their real or simulated nonverbal behavior, for example moving themselves away from the agent in the real or virtual world.

Intimacy may also be improved through a greatly expanded repertoire of relational dialog, including such things as more in-depth reciprocal self-disclosure exchanges to establish more common ground with the user.

Another future area of exploration is giving users the ability to make relational bids. Although many may not use this function, for some (such as subject #5 above) it would allow them to negotiate their desired intimacy level much more quickly than our current algorithm allows. It may also be more effective when a change in working relationship is the user's idea instead of the agent's.

Our relational model is admittedly over-simplistic, and development of more sophisticated models represents another important direction of research. The number of intimacy levels modeled and the sets of relational actions indexed should both be functions of context and user role. Ultimately, relational models based on specific beliefs and intentions of the agent and user will be needed for maximum fidelity.

A final note on methodology. While longitudinal models of user-agent relationship represent an open and promising area of research, studies of actual user responses to such models take very long periods of time to conduct (the studies described here took over two years to complete). The development of evaluation and simulation tools that can provide insights into longitudinal human behavior in shorter time frames is thus another important area of investigation.

Acknowledgments. Thanks to Barbara Barry and Laura Pfeifer for comments on early drafts, and to the anonymous reviewers for their many suggestions. This material is based upon work supported by the National Science Foundation under Grant No. 0545932. 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.

References

1. Bickmore, T.: Relational Agents: Effecting Change through Human-Computer Relationships. Ph.D. Dissertation, Media Arts & Sciences. Massachusetts Institute of Technology, Cambridge, MA (2003)
2. Bickmore, T., Picard, R.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer Human Interaction 12, 293–327 (2005)
3. Cassell, J., Bickmore, T.: Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intellient Agents. User Modeling and Adaptive Interfaces 13, 89–132 (2003)
4. Kidd, C.D.: Designing Long-Term Human-Robot Interaction and Application to Weight Loss. Ph.D. Dissertation, Media Arts & Sciences. Massachusetts Institute of Technology, Cambridge, MA (2008)
5. McGuire, A.: Helping Behaviors in the Natural Environment: Dimensions and Correlates of Helping. Personality and Social Psychology Bulletin 20, 45–56 (1994)
6. Rusbult, C., Drigotas, S., Verette, J.: The investment model: An interdepence analysis of commitment processes and relationship maintenance phenomena. In: Canary, D., Stafford, L. (eds.) Communication and Relational Maintenance, pp. 115–139. Academic Press, San Diego (1994)

7. Svennevig, J.: Getting Acquainted in Conversation. John Benjamins, Philadephia (1999)
8. Grosz, B., Kraus, S.: Collaborative plans for complex group action. Artificial Intelligence 86, 269–357 (1996)
9. Thomason, R.: Accommodation, Meaning, and Implicature: Interdisciplinary Foundations for Pragmatics. In: Cohen, P., Morgan, J., Pollack, M. (eds.) Intentions in Communication, pp. 325–364. MIT Press, Cambridge (1990)
10. Bickmore, T., Schulman, D.: A Virtual Laboratory for Studying Long-term Relationships between Humans and Virtual Agents. In: Proceedings Autonomous Agents and Multi-Agent Systems, Budapest (2009)
11. Aron, A., Melinat, E., Aron, E.N., Vallone, R.D., Bator, R.J.: The experimental generation of interpersonal closeness: A procedure and some preliminary findings. Personality and Social Psychology Bulletin 23, 363–377 (1997)
12. Horvath, A., Symonds, D.: Relation Between Working Alliance and Outcome in Psychotherapy, A Meta-Analysis. Journal of Counseling Psychology 38, 139–149 (1991)
13. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: Proceedings SIGGRAPH 2001, pp. 477–486 (2001)
14. Bickmore, T., Caruso, L., Clough-Gorr, K., Heeren, T.: “It’s just like you talk to a friend” - Relational Agents for Older Adults. Interacting with Computers 17, 711–735 (2005)
15. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Education & Counseling 59, 21–30 (2005)
16. Hatcher, R.L., Gillaspy, J.A.: Development and validation of a revised short version of the working alliance inventory. Psychotherapy Research 16, 12–25 (2005)
17. Velicer, W., Bickmore, T., Byron, D., Johnson, J.: Using Relational Agents in Interventions for Multiple Risk Factors. In: Proceedings of the Society for Behavioral Medicine Annual Conference (2011)
18. Goffman, I.: On face-work. Interaction Ritual: Essays on Face-to-Face Behavior, pp. 5–46. Pantheon, New York (1967)
19. Argyle, M., Dean, J.: Eye-contact, distance and afliation. Sociometry 28, 289–304 (1965)

Cultural Study on Speech Duration and Perception of Virtual Agent's Nodding

Tomoko Koda, Haruka Kishi, Takanori Hamamoto, and Yota Suzuki

Department of Information Science and Technology, Osaka Institute of Technology
1-79-1 Kitayama, Hirakata, 573-0196 Osaka, Japan
koda@is.oit.ac.jp

Abstract. Nodding plays an important role in successful interaction between a human and virtual agent as well as in human-human communication. We compare Japanese speech durations and perceptions of a nodding virtual agent in a cross-cultural setting. The results indicate the importance of cultural adaptation of timing and frequency of virtual agent's nodding.

Keywords: nod, backchannel, speech, virtual agents, cross-culture.

1 Introduction

Backchannels refer to the short expressions the listener sends while the speaker is exercising their speech rights, and are classified as either linguistic expressions such as “uh-huh” and “yeah,” or as non-linguistic expressions such as nods and smiles [1]. Virtual agents that exhibit human-like backchannel behavior have been developed in the virtual agent research community. Those research aims to create “active listener” agents that produce feedback signals in response to user actions [2, 3, 4, 5]. Poope et al. report the number, timing and type of backchannels of an agent had a significant effect on perceived human-likeliness of the agent [6]. Nods are more perceived as appropriate than vocalizations, and adequate frequency of backchannels is important for the quality of the backchannel behavior [7]. Huang et al. showed adequate backchannels given by the agent improve rapport, perceived accuracy and naturalness [8]. Also, enculturating non-verbal behaviors of virtual agents has been getting into focus in recent years [9, 10, 11].

Linguists have indicated there are cultural differences in backchannels between humans. For example, the cultural differences between Japanese and American backchannels manifestly appear in their frequency and timing, and Japanese are said to make as much as double the amount of backchannels than Americans [12, 13]. White’s cultural study on backchannels between Americans and Japanese also suggest several types of backchannels (except “yeah”) are displayed far more frequently by Japanese listeners [14].

This study adapts backchannel timing and frequency of Japanese and American people reported in [12, 13] to the case of nodding virtual agents (listener of Japanese

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 404–411, 2012.

speakers). We expect the results would serve as an indication of importance of giving backchannels with a frequency and timing culturally adapted to the speaker.

2 Related Research in Cultural Difference of Backchannels

Maynard's conversation analysis [12, 13] indicated that conspicuous cultural differences appear in the frequency and timing of backchannels between Japanese and Americans. The frequency and timing in backchannels in Japan and America are presented in Table 1 [translated from 13]. The method for collecting data on the backchannels in Japan and America was to use a total of 120 minutes of video-recorded conversations between twenty groups of pairs of Japanese and Americans respectively.

From Table 1, it can be seen that between Japanese and Americans, Japanese make backchannels at double the frequency. Also, regarding the timing of backchannels, it can be seen that Japanese make backchannels at all points, such as near the pause at the end of the sentence, or near final particles and interjection particles. This also depends on the linguistic structure, but for Japanese in particular, listeners are often guided by the head movements of the speaker in giving their backchannels. In contrast to this, 80% of the total number of backchannels made by Americans is at the pause at the end of a sentence, and it is evident that they do not give backchannels at all points the way Japanese do.

Table 1. Backchanel timing and frequency in Japan and America in casual conversations of 20 pairs each (total of 120 minutes each) [13]

	Japanese	American
pause at the end of a sentence	351	309
final particles and interjection particles	281	n/a
At te end of a tag question	54	26
Speaker's vertical head movement	262	29
Total number of backchannels	948	364

3 The Experiment

3.1 Hypotheses

This research aims to adapt the above cross-cultural conversation analysis [12, 13] to the case of human (speaker) and a nodding agent (listener) interaction, in order to examine the effect of enculturating nodding timing and frequency of virtual listening agent. The reason for using only nods for giving backchannels is that we wanted to exclude the fact that several behaviors might have an impact one another.

We formulated the hypothesis "for the Japanese participants in this experiment, speech with an agent that nods with the same frequency and timing as in Japan will have a longer speech duration than with an agent that nods with the same frequency and timing as in America, and the participants feel less stress while talking to the former agent." Speech duration was compared using an agent which nods with the same frequency and timing as in Japan (henceforth, JA), one with the same frequency and timing as in America (henceforth, AA), and one which does not nod (henceforth, NA). We examine the hypothesis with Japanese speakers only as the first step in this study. Conducting another experiment with Americans will be a further study.

In this experiment, in order to control the nodding behavior by the agent, we used the Wizard of Oz Method. In JA condition, the Japanese experimenter controlled the agent's nodding at the Japanese timing in Table 1, in AA condition at the American timing in Table 1, and in NA condition, the agent does not nod at all. The automatic generation of nods could be done based on the analysis of the speech and head movements, but in order to improve the accuracy of the nodding timing and frequency, we decided to use a WoZ setting. The participants' speech was recorded through video recording.

For the nodding agent used in this experiment, we used an animal-type agent developed as a listener. It performs nodding actions two complete times, at a perpendicular angular direction of 20°, and at a return speed of 0.6 seconds a turn, which is considered to give the greatest sense of listening for Japanese speakers in our pilot study [15]. For the nodding agent's developmental environment, Microsoft Visual C++, DirectX SDK, and for the modeling, MetasequoiaLE were used. Examples of the agent's nodding action are shown in Figure 1.

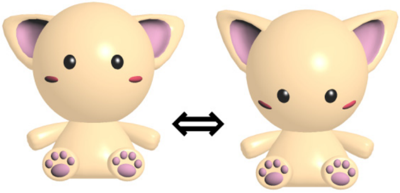


Fig. 1. Examples of Nodding Action by the Agent

3.2 Experiment Procedures

The participants in this experiment were 20 Japanese university students (11 males and 9 females), who have no experiences living abroad. The experiment was conducted as the follows:

1) Japanese participants were asked to choose three topics they could talk about from a selection of eighteen topics, i.e., pets, food, books, movies, fashion, music, sports, etc. They talked about the topics in Japanese in the three talking sessions.
2) During each talking session, the Japanese experimenter generated the JA, AA, and NA nodding pattern according to the nodding timing and frequency in Table 1. American conversant give backchannels at the end of a sentence 80% of the total backchannels during a conversation, thus the AA nodding was generated at the end of a sentence spoken in Japanese by the Japanese participants. The nodding patterns are presented randomly in order not to have a sequential effect. No instructions were made to the participants about the change of nodding patterns.

3) A sensor to measure the participant's heart beat rate is attached to the participant's indicate finger. Heart beat rate can be used as a physiological measurement of participants' stress level during speech. The participant's normal heart beat rate is measured for 3 minutes before the first talking session starts. The participants heart beat rate is recorded all through the three topics.
4) The participants were asked to answer a questionnaire after each talking session. The questionnaire consisted of questions regarding perceived stress toward talking to the agent, smoothness of speech, perceived activeness of the agent, likeability of the agent and perceived kindness of the agent. For a total of 24 questions were rated using a 7-point Likert scale (7: high– 1: low).

4 Results

4.1 Comparison on Speech Duration

Regarding the number of times the agent nodded in each session, the average nodding times in JA condition was 20.8, and 10.35 times in AA condition (no nodding in NA condition). The nodding frequency of JA was about twice as high as AA (both are operated by the WoZ experimenter at the timing of Table 1), which is in consistent with the nodding frequency in actual casual conversation between humans reported in [12, 13].

The audio portion of the video-recorded data was extracted as a waveform using Any Audio Converter (http://www.any-audio-converted.com/jp). The participants' average speech duration of JA is 105.5 sec, AA is 86.1 sec, and NA is 79.2 sec. Seventy percent of the participants in the experiment spoke the longest in JA conditions among three conditions. The results of multiple comparisons of the speech durations show a significant difference ($p<=0.01$) between JA (105.5 sec) - NA (79.2 sec). No significant difference was seen between JA (105.5 sec) – AA (86.1 sec) and AA (86.1 sec) – NA (79.2 sec).

4.2 Analysis of Subjective Evaluation and Stress Level

Perceived Stress and Smoothness of the Actual Conversations and Physiological Stress

As a subjective evaluation of the actual conversations, the results of one-way ANOVA regarding the perceived stress toward the actual conversation and the perceived smoothness of the conversation are shown in Figure 2 (left and middle). For the stresses toward the actual conversation, the ratings were inverted, as the lower the stress the higher the evaluation. A significant difference was observed ($p<=0.01$) between JA (5.3) –AA (3.6) and JA (5.3) -NA (3.2) for the stresses toward the conversation itself. A significant difference ($p<=0.01$) was seen between JA (5.1) –NA (2.6) and AA (4.5) - NA (2.6) for the smoothness of the conversation. Also, a significant trend ($p<=0.05$) was seen between JA (5.1) –AA (2.6).

The heart beat rates before the experiment (normal rate) and at the end of each talking session are shown in Figure 2 (right). The heart beat rate can be used as a physiological measurement of stress. The normal heart beat (82.5 times per minute) rate is significantly lower (p<=0.01) than AA (86.1 times) and NA (87.6 times), but not significant between JA (83.9 times). There is another significant difference (p<=0.01) in heart beat rate between JA (83.9 times) and NA (87.6 times).

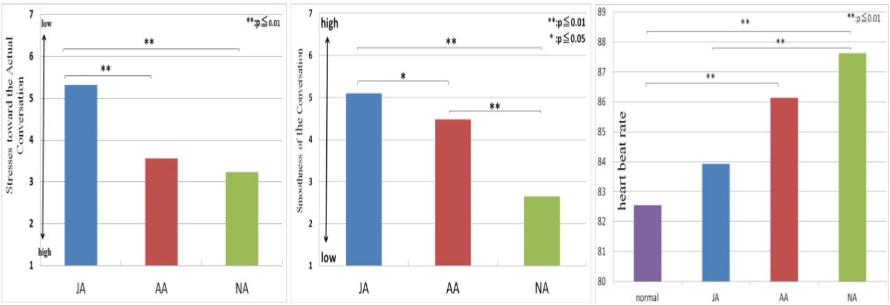


Fig. 2. Results of the analysis of the perceived stresses toward the actual conversation (left), smoothness of the conversation (middle), and results of multiple comparison of heart beat rate before and after speech (right)

Impression of the Agent

As a subjective evaluation of the impression of the agent, the results of one-way ANOVA regarding the perceived activeness of the agent, the likeability of the agent and the perceived kindness of the agent are shown in Figures 3.

Regarding the perceived activeness of the agent, a significant difference (p<=0.01) was seen among all of the agents, JA (5.5) – AA (4.7) – NA (2.1). Regarding the likeability of the agent, a significant difference (p<=0.01) was seen between JA (5.2) – NA (2.7) and AA (4.6) – NA (2.7). Between JA (5.2) – AA (4.6) a significant trend (p<=0.05) was seen. Regarding the perceived kindness of the agent, a significant difference (p<=0.01) was seen among all of the agents, JA (5.2) – AA (4.6) – NA (3.0).

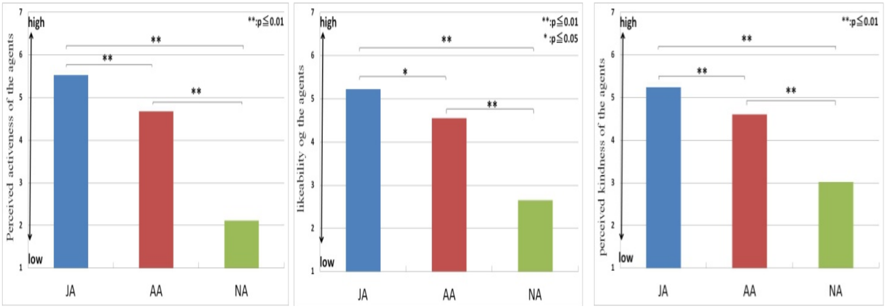


Fig. 3. Results of the analysis of the perceived activeness of the agent (left), likeability of the agent (middle), and perceived kindness of the agent (right)

5 Discussion

5.1 Discussion of the Comparison Experiment on Speech Duration

The frequency of nods in JA condition was approximately two times that of in AA condition. Thus, we can regard that the WoZ's generation of nodding was appropriate for both conditions. Comparison of speech duration showed a significant difference between JA and NA condition. When the Japanese participants talked to the virtual agent that nods like Japanese people do in terms of frequency and timing, then talked longer than they talked in front of the agent that nods like Americans and the one without nods. The reason for not having significant difference between JA-AA and AA-NA might be caused by wider variations in speech duration. Five minute was the guideline for one talking session, but variations in proficiency in speech and ease in speaking on the selected topic on the part of the participants in the experiment can be considered the causes.

5.2 Discussion on the Subjective Evaluation and Physiological Measurement

Discussion on the Perceived Stress and Smoothness of the Actual Conversations and Physiological Stress

Regarding the perceived stress toward the actual conversation, a significant difference ($p<=0.01$) was seen between JA-AA and JA-NA. The Japanese participants evaluated the stress was reduced the most by the agent nodding at a Japanese frequency and timing.

For the perceived smoothness of the conversation, significant differences were seen between JA-NA and AA-NA ($p<=0.01$) and significant tendency ($p<=0.05$) was seen between JA-AA. The conversations were perceived as smoother when the agent nodded than when it did not. The conversations were perceived as smoother with nods with Japanese frequency and timing rather than those with American frequency and timing for the Japanese participants.

These subjective evaluation results suggest that for the Japanese participants in the experiment, perceived stress in the conversation itself is reduced, and it is possible to conduct smoother conversations by the listener agent nodding with a Japanese frequency and timing. This suggests nod timing and frequency matched to the experiment participant's culture are important for conducting less-stressed and smoother conversation.

The physiological measurement of heart beat rate supported the subjective evaluation results. Significant differences in heart beat rates are seen between normal and AA, and normal and NA, but not between normal and JA condition. This suggests the Japanese participants physiologically felt stress while talking to the listener agent that doesn't nod like Japanese do in terms of frequency and timing, but could keep their stress level as low as their normal level when they talk to the listener agent that nods like Japanese.

Discussion of the Subjective Evaluation of the Impression of the Agent
A significant difference was seen among all of the agents regarding the perceived activeness of the agent. Activeness could be perceived in agents who performed nodding actions than those which did not nod at all. Furthermore, activeness could be perceived in agents that nodded with Japanese frequency and timing that those with American frequency and timing for the Japanese participants.

Significant differences (p<=0.01) between JA-NA and AA-NA and a significant tendency between JA-AA (p<=0.05) were seen regarding the likeability toward the agents. The agents that performed nodding behaviors were liked more than those that did not. There was a tendency to have favorable feelings toward the agent which nod with a Japanese frequency and timing than an American one for the Japanese participants.

A significant difference was seen among all of the agents (p<=0.01) regarding the perceived kindness of the agent. Kindness could be perceived by the agent when it performed nodding behaviors than when it did not. Furthermore, kindness could be perceived in agents which nodded with a Japanese frequency and timing than an American one for the Japanese participants.

These results suggest that for the Japanese participants in this experiment, the agent's perceived activeness, degree of likability, and perceived kindness rise by the listener agent nodding with Japanese frequency and timing, and that by matching nodding to the experiment participant's culture, a better impression of the listener agent can be given to the speaker.

6 Conclusion

This study performed a comparison experiment on speech duration using agents which nodded with the respective timing and frequency of the Japanese and American and agents which did not, in order to verify the importance of considering cultural differences in nodding in communication between people and listener agents. The result of this experiment was that the participants spoke the longest in conversations with an agent which nodded with Japanese frequency and timing. Subjective evaluation indicated that by the listener agent nodding with a Japanese frequency and timing, the stresses of the actual conversation were alleviated, conversation could be conducted more smoothly, and the participants had a better impression of the agent. These results suggest it is important for the listening agent to give backchannels with a timing and frequency culturally adapted to the speaker in order to establish smooth communication between people and agents.

We used only Japanese participants in this study. However, conducting a similar experiment with American participants is strongly needed with a nodding motion appropriate for Americans. We expect the results would serve as an indication of importance of giving backchannels with a frequency and timing culturally adapted to the speaker.

Acknowledgement. This research is partially supported by a Grant-in-Aid for Scientific Research (C) 23500266) (2011-2013) from the Japan Society for the Promotion of Science.

References

1. Yngve, V.H.: On getting a word in edgewise. In: Papers from the 6th Regional Meetings of Chicago Linguistic Society, pp. 567–577. Chicago Linguistic Society (1970)
2. Jonsdottir, G.R., Gratch, J., Fast, E., Thórisson, K.R.: Fluid Semantic Back-Channel Feedback in Dialogue: Challenges and Progress. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 154–160. Springer, Heidelberg (2007)
3. Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling Embodied Feedback with Virtual Humans. In: Wachsmuth, I., Knoblich, G. (eds.) ZiF Research Group International Workshop. LNCS (LNAI), vol. 4930, pp. 18–37. Springer, Heidelberg (2008)
4. 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)
5. Bevacqua, E., Pammi, S., Hyniewska, S.J., Schroder, M., Palachaud, C.: Multimodal backchannels for embodied conversational agents. In: Safonova, A. (ed.) IVA 2010. LNCS (LNAI), vol. 6356, pp. 194–200. Springer, Heidelberg (2010)
6. Poppe, R., Truong, K.P., Reidsma, D., Heylen, D.: Backchannel Strategies for Artificial Listeners. In: Safonova, A. (ed.) IVA 2010. LNCS (LNAI), vol. 6356, pp. 146–158. Springer, Heidelberg (2010)
7. Poppe, R., Truong, K.P., Heylen, D.: Backchannels: Quantity, Type and Timing Matters. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS (LNAI), vol. 6895, pp. 228–239. Springer, Heidelberg (2011)
8. Huang, L., Morency, L.-P., Gratch, J.: Learning Backchannel Prediction Model from Parasocial Consensus Sampling: A Subjective Evaluation. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 159–172. Springer, Heidelberg (2010)
9. Payr, S., Trappl, R.: Agent Culture: Human-Agent Interaction in a Multicultural World. CRC Press (2004)
10. Jan, D., Herrera, D., Martinovski, B., Novick, D., Traum, D.R.: 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)
11. Rehm, M., Nakano, Y., Koda, T., Winschiers-Theophilus, H.: Culturally Aware Agent Communication. In: Zacarias, M., de Oliveira, J.V. (eds.) Human-Computer Interaction. SCI, vol. 396, pp. 411–436. Springer, Heidelberg (2012)
12. Maynard, S.K.: On backchannel behavior in Japanese and English casual conversation. Linguistics 24, 1079–1108 (1986)
13. Maynard, S.K.: Kaiwabunseki, pp. 152–166. Kuroshio Publishing (1993) (in Japanese)
14. White, S.: Backchannels across cultures: A study of Americans and Japanese. Language Society 18, 59–76 (1989)
15. Wakisaka, M.: Analysis of Natural Frequency, Angle, and Velocity of Nods of Virtual Agent. Graduation thesis of Faculty of Information Science and technology at Osaka Institute of Technology (2009)

Cultural Behaviors of Virtual Agents in an Augmented Reality Environment

Mohammad Obaid[1,2], Ionut Damian[1], Felix Kistler[1], Birgit Endrass[1], Johannes Wagner[1], and Elisabeth André[1]

[1] Human Centered Multimedia, University of Augsburg, Augsburg, Germany
{damian,kistler,endrass,wagner,andre}@informatik.uni-augsburg.de
[2] Human Interface Technology Lab. New Zealand, Christchurch, New Zealand
mohammad.obaid@hitlabnz.org

Abstract. This paper presents a pilot evaluation study that investigates the physiological response of users when interacting with virtual agents that resemble cultural behaviors in an Augmented Reality environment. In particular, we analyze users from the Arab and German cultural backgrounds. The initial results of our analysis are promising and show that users tend to have a higher physiological arousal towards virtual agents that do not exhibit behaviors of their cultural background.

Keywords: Virtual Agents, Augmented Reality, Proxemics, Eye-gaze, Culture.

1 Introduction

In human-human communications, the majority of our communicative behaviors are demonstrated non-verbally based on the individuals' personality, emotion and cultural background. Researchers have demonstrated that social behaviors in different cultures are addressed differently; therefore, social behavioral differences between communicating individuals can be misunderstood, which might result in a negative perception of each other.

With the increased use of 3D virtual agents, researchers have identified the importance of integrating social behaviors in virtual agents and demonstrated that they have an impact on the user-agents social interaction, e.g. Jan et al. [1]. While, more recently, several researchers measured the influence on the user's perception by the deployment of enculturated virtual characters [2].

To understand human-agent social behaviors, researchers are usually conducting studies in traditional virtual reality (VR) environments displayed on a screen [1]. In this case, the digital medium has an influence on (1) what the user considers a real experience and (2) how the user responds socially when interacting with a virtual agent within the digital medium.

This paper presents the integration of culturally aware virtual agents in the real-space of the user using the Augmented Reality (AR) Technologies; which may have an impact on the users' responses. In recent years, only few studies explore the use of virtual agents in AR environments such as the work in [3,4].

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 412–418, 2012.

In this paper, we focus on two kinds of signals that play an important role in communicating interest in a social interaction and creating intimacy: interpersonal distance and gaze. Interpersonal distance and gaze are closely coupled. In combination, they convey a certain level of intimacy while people tend to compensate for an increased amount of eye gaze that they consider as socially inappropriate by a higher interpersonal distance. Our paper aims at studying how the response of people is influenced by the cultural background such agents reflect. In particular, we study the physiological arousal response of German and Arab users towards agents reflecting these cultures.

2 Proxemics and Eye Gaze in Virtual Agents

In the literature, a variety of approaches have been presented to simulate proxemic behaviors (the use of space as a form of nonverbal communication) in virtual agents. For example, [5] and [6] presented work based on reinforcement learning and on crowd simulation respectively. Studies have been conducted to investigate the effect of an agent's proxemic behaviors on the users' perception of the agent and their psychological state. Pedica et al. [7] showed that a simulation of human social territoriality, in simulated conversations, contributes to the believability of agent groups. Llobera et al. [8] found that people showed increased physiological arousal the closer they were approached by virtual agents.

Vice versa, a number of researchers investigated how the behavior of a virtual agent influences the proxemics behavior of human users. Bailenson et al. [9] observed that human participants gave more personal space to virtual agents who establish mutual gaze with them.

The simulation of proxemics behaviors has also attracted a significant amount of attention in the robotics community. The work by Mumm and Mutlu [10] is of particular interest to us because it includes a combined analysis of gaze and proxemics behaviors. In a study, they observed that people who disliked a robot compensated for an increase amount of gaze, from the robot, by maintaining a higher physical distance from the robot. While people who liked the robot did not change their proxemics behaviors across various gaze conditions.

In most culture related experiments, researchers extract data only from subjective sources and few attempts were made to study the cultural differences on an objective level, such as the work by Obaid et al. [3] and Llobera et al. [8].

3 Theoretical Background

The integration of cultural factors into the behavior of virtual agents came into focus in recent years. Researchers often investigate the user's perception of a group of virtual agents that show culture-related differences in perception studies. Endrass et al. [2] showed that a culturally prototypical performance of gestures and body postures can enhance the user's perception of an agent conversation in the German and Japaneses cultures. Jan et al. [1] took into account culture-related gaze, proxemics and turn-taking behaviors representing the Arab and the US American cultures. Their results reveal that participants perceived differences between behaviors that are in line with their own cultural background, and behaviors from different cultural backgrounds.

3.1 Cultural Profiles

In the scope of this paper, culture-related aspects of interpersonal distance and gaze behavior are investigated for the Arab and German cultures. We chose these two cultural groups because clear distinctions can be derived from the research literature, thus promising large differences in participants' perception of nonverbal behavior. A very well known dimensional model of culture is introduced by Hofstede et al. [11], whose theory categorizes national cultures into a six dimensional model.

The Individualism dimension is of special interest for our purpose since it is strongly related to nonverbal behavioral norms. The dimension describes the degree to which individuals are integrated into a group. On the individualist side, ties between individuals are loose, and everybody is expected to take care for him or herself. On the collectivist side, people are integrated into strong, cohesive in-groups. Germany is considered to be an individualistic culture, while the Arab world is considered collectivistic.

Hall introduces a dichotomy [12] that distinguishes so-called *high-* and *low-contact* cultures, which is mainly related to space and appropriate interpersonal distances in social situations. According to Ting-Toomey [13], Germany belongs to the medium-contact group while Arabia belongs to the high-contact group.

3.2 Expectations on Behavioral Differences

The cultural dimension and dichotomy, described in Section 3.1, influence stereotypical behavior for interpersonal distance and eye gaze, which should differ vastly in the Arab and German cultures.

Interpersonal Distance: in [14], Hofstede et al. investigate cultural dimensions in isolation and describe prototypical behavior for the extreme sides. Individualistic cultures, such as Germany, are described to stand free in groups, while collectivistic cultures, such as Arabs tend to be physically close, especially to in-groups. This suggests that interpersonal distance should be higher in German conversations compared to Arab ones. This idea is supported by the categorization into high- and low-contact cultures. In high-contact cultures (such as Arabia) close interactions are common, while in low-contact cultures wider interpersonal space is appropriate.

Eye Contact: the scores on Hofstede's individualism dimension give some interesting insights. In [14], members of individualistic cultures are described as making eye contact freely which suggests that this should hold true for the German culture. Taking into account Hall's categorization into high- and low-contact cultures, members of Arab countries (low-contact) use direct facing and frequent direct eye contact. Since Germany is located on a medium-low contact level, direct facing should occur less frequently.

Table 1. Culture-related behaviors

Aspect	Arab	Neutral	German
distance	low	medium	high
eye contact	high	medium	low

Table 2. Experimental Conditions

Condition	distance	eye contact
0	Neutral	Neutral
1	Arab	German
2	German	Arab
3	Arab	Arab
4	German	German

4 Experimental Evaluation

To evaluate the impact of culture-specific behavior on human users, we conducted a pilot study with Egyptian and German participants. Due to the similarity principle, the social attraction perceived by individuals is increased when interacting with someone who is perceived as being similar to oneself [15]. Therefore, participants should prefer agent behavior that was designed to resemble their own cultural background. Vice versa, culturally different behavior might lead to tension which should be measurable in physiological arousal. We therefore derive our ***Hypothesis:*** *AR agents that have a different culture behavior from the human user elicit a higher level of physiological arousal.*

4.1 Design

Agents: we use two virtual agents in our setup based on the AAA engine and functionalities [16], which includes the control of non-verbal behaviors conveyed by agents. The appearance of the agents are consistent and are chosen to be of dark-hair, no-culture specific clothing, and of mixed gender. Moreover, to allow for a focus on the nonverbal behaviors, the agents use Gibberish during dialog, a fantasy language without any specific meaning of the words but with the same statistical distribution of syllables as the words from the English language.

Conditions: The experimental conditions are designed based on the cultural profiles and expectations (Section 3). Therefore, we defined three prototypical performances (Arab, Neutral and German), where the Neutral behavior serves as the average of the behaviors for the German and Arab cultures. Neutral was used during a non-evaluated user familiarization phase (Tables 1 and 2). Since our work focuses on analyzing the impact of differences between the cultural groups and not measuring the exact effect of the conditions, therefore, the exact definition of the values for the behavior aspects was not a priority and are defined relative to each other (e.g. high gaze is higher than low gaze).

Scenario: Within each condition, the user enters the AR environment with two agents engaged in a conversation. Starting from a position of 2.8 meters away from the agents, the user is instructed to approach the agents to a comfortable distance to join their conversation as a listener (Fig. 1). As the user approaches, the agents adjust their conversational formation with an interpersonal distance based on the condition. In addition, the agents' amount of eye contact with the

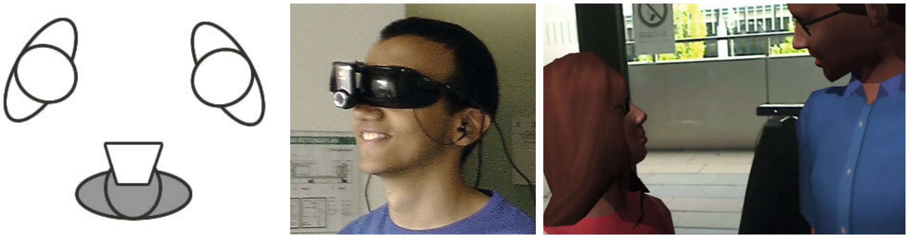


Fig. 1. Schematic setup of the experiment showing the position of the user and virtual agents (left), participant during interaction (middle), and an example of the user's view showing the virtual agents in the AR environment (right)

interlocutors is adapted to the condition. Other nonverbal behaviors, such as gestures and body postures, are left constant during all conditions.

Experimental Setup: The experiment is arranged in a room with five-meter width and six meter depth. The setup contains two main pieces of hardware equipment for tracking and AR display: (1) A Microsoft Kinect[1] is put on a pedestal of one meter height and placed at a centered position against the wall. We use the OpenNI framework and NITE middleware[2] to track the users in front of the Kinect and to get the user's position in relation to the sensor. (2) A Vuzix[3] Head Mounted Display (HMD) that includes earphones and a head tracker (measuring the head orientation via an accelerometer and compass). Using the tracking technologies, we incorporated the virtual agents into an AR environment. We apply the user's position (tracked by the Kinect) and head orientation (from the HMD's tracker) to the virtual camera used to compute the 3D transformations for rendering the virtual agents. To simulate the user's real world view, a webcam is attached to the HMD. The video image is captured, augmented with the virtual agents, and then presented back to the user's eyes. In addition, for accuracy, the head tracker is calibrated for each user, before the study, so that head orientation and virtual camera orientation are in line.

Throughout the study, we use the *NeXuS*[4] and the SSI framework [17] to measure and record four biosignals (electrocardiography, respiration, skin conductivity and blood-volume-pulse). Using these signals, we were then able to compute the user's level of physiological arousal during the study.

Procedure: The experiment begins by giving a description of the study and the equipment to the participant. After fitting the equipment to their body, they are asked to stand at the initial position in the room. Thereafter, the participant is acquainted with the environment and the gibberish dialog by allowing them to experience condition 0 of Table 2. Only then, the participant will randomly conduct each of the four different conditions, described in Table 2.

[1] http://www.xbox.com/kinect

[2] http://www.openni.org

[3] http://www.vuzix.com/consumer/products_wrap_1200vr.html

[4] http://www.humankarigar.com/wireless_nexus10.htm

5 Results and Discussion

For our evaluation study, we recruited 20 participants: 10 from Egypt (5 females, 5 males) and 10 from Germany (3 females, 7 males), while all of them had little or no experience with augmented reality environments.

The analysis of the physiological recordings yielded some interesting tendencies. From the ECG signals, we extracted the heart rate from the interval between successive R waves. Since the physical load was steady across the conditions an increase of the heart rate can be taken as a sign for an increased stress level of the participants. To observe differences between the four conditions (Table 2), we calculated for each participant, the mean heart rate over the period of each condition. In order to reduce the effect of individual differences we also subtracted the mean heart rate extracted from the whole session. Then we compared the average of the normalized values between Egyptian and German participants. For the conditions, where gaze and distance are adjusted to the same country (3+4), the values are within a range from $-1.1hz$ to $-0.6hz$, which is below the average heart rate. This indicates that participants generally were in a relaxed mood even when they were facing the conditions of the other culture, as long as eye gaze and distance are of the same culture. In case of the mixed conditions (1+2), for both groups, the normalized heart rate increased above zero when the eye gaze was adjusted to the other culture ($0.2hz$ for German and $0.4hz$ for Egyptians), but not in the case where the distance was adjusted to the other culture ($-1.1hz$ for German and $-1.4hz$ for Egyptians). This suggests that the eye gaze aspect has a higher impact on the level of physiological arousal than distance. However, we predict that a clearer picture on the statistical analysis of our results will reform when analyzing the results of additional participants. Therefore, we neither deny nor approve our hypothesis at this stage.

6 Conclusion and Future Work

In this paper we presented a pilot evaluation study to investigate the physiological arousal of users towards augmented reality virtual agents that resembles cultural behaviors. We demonstrate this by integrating two social signals (interpersonal distance and eye-gaze) for two different cultures (Arab and German) into virtual characters. The study is still in running process, however, we reported the analysis of the initial results from 20 participants (10 Arab and 10 German). The initial results, from the physiological responses of users, revealed interesting tendencies (such as a higher heart-rate average when interacting with agents of culturally inconsistent non-verbal behavior). The participants also perceived the conditions with a culturally appropriate eye gaze as more relaxing even if they are experiencing non-appropriate distance activity. Future work are directed towards exploring social attraction towards agents simulating cultural behaviors.

Acknowledgments. This work was partially funded by the European Commission within the 7th Framework Program under grant agreement eCute.

References

1. 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)
2. Endrass, B., André, E., Rehm, M., Lipi, A.A., Nakano, Y.: Culture-related differences in aspects of behavior for virtual characters across germany and japan. In: Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems, AAMAS, Richland, SC, vol. 2, pp. 441–448 (2011)
3. Obaid, M., Niewiadomski, R., Pelachaud, C.: Perception of Spatial Relations and of Coexistence with Virtual Agents. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 363–369. Springer, Heidelberg (2011)
4. Magnenat-Thalmann, N., Papagiannakis, G., Chaudhuri, P.: Applications of Interactive Virtual Humans in Mobile Augmented Reality. In: Furht, B. (ed.) Encyclopaedia of Multimedia, 2nd edn. Springer (2008)
5. Kastanis, I., Slater, M.: Reinforcement learning utilizes proxemics: An avatar learns to manipulate the position of people in immersive virtual reality. ACM Trans. Appl. Percept. 9(1), 3:1–3:15 (2012)
6. Jan, D., Traum, D.: Dynamic movement and positioning of embodied agents in multiparty conversations. In: Proc. of the Workshop on Embodied Language Processing, EmbodiedNLP 2007, Stroudsburg, PA, USA, pp. 59–66. Association for Computational Linguistics (2007)
7. Pedica, C., Högni Vilhjálmsson, H., Lárusdóttir, M.: Avatars in conversation: The importance of simulating territorial behavior. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 336–342. Springer, Heidelberg (2010)
8. Llobera, J., Spanlang, B., Ruffini, G., Slater, M.: Proxemics with multiple dynamic characters in an immersive virtual environment. ACM Trans. Appl. Percept. 8(1), 3:1–3:12 (2010)
9. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Interpersonal distance in immersive virtual environments. Personality and Social Psychology Bulletin 29(7), 819–833 (2003)
10. Mumm, J., Mutlu, B.: Human-robot proxemics: physical and psychological distancing in human-robot interaction. In: Proc. of the 6th Int. Conf. on Human-robot Interaction, HRI 2011, pp. 331–338. ACM, New York (2011)
11. Hofstede, G., Hofstede, G.J., Minkov, M.: Cultures and Organisations. Software of the Mind. Intercultural Cooperation and its Importance for Survival. McGraw Hill (2010)
12. Hall, E.T.: The Hidden Dimension. Doubleday (1966)
13. Ting-Toomey, S.: Communicating across cultures. The Guilford Press, New York (1999)
14. Hofstede, G.J., Pedersen, P.B., Hofstede, G.: Exploring Culture - Exercises, Stories and Synthetic Cultures. Intercultural Press, Yarmouth (2002)
15. Byrne, D.: The attraction paradigm. Academic Press, New York (1971)
16. Damian, I., Endrass, B., Huber, P., Bee, N., André, E.: Individualized Agent Interactions. In: Allbeck, J.M., Faloutsos, P. (eds.) MIG 2011. LNCS, vol. 7060, pp. 15–26. Springer, Heidelberg (2011)
17. Wagner, J., Lingenfelser, F., André, E.: The social signal interpretation framework (SSI) for real time signal processing and recognition. In: Proc. of Interspeech 2011, Florence, Italy, pp. 3245–3248 (2011)

Frown More, Talk More: Effects of Facial Expressions in Establishing Conversational Rapport with Virtual Agents

Joshua Wong Wei-Ern and Kevin McGee

National University of Singapore
{cnmjwwe,cnmmk}@nus.edu.sg

Abstract. How can conversational agents be better designed to build rapport with human beings? Related work on creating rapport through conversational agents has largely focused on nonverbal contingent envelope feedback. There is relatively little known about how forms of emotional feedback play a role in building rapport between agents and humans. This paper describes a study in which people told stories to an agent that provided emotional feedback in the form of facial expressions. Rapport was measured through the length of the stories, the fluency of their speech, and the user's own subjective experience. Surprisingly, results indicated that inappropriate emotional feedback *increased* story length, which was the opposite of previous studies on envelope feedback that had shorter stories in unresponsive conditions. This paper explains the factors particular to emotional feedback that could cause this difference.

Keywords: rapport, emotional feedback, virtual agents.

1 Introduction

One of the characteristics of enjoyable or memorable conversations is the feeling of rapport we have with the other person. Tickle-Degnen and Rosenthal [1] describe rapport as a quality of the interaction between individuals, that can be thought of as having three essential components: *mutual attentiveness*, *positivity*, and *coordination* between participants. Many people can be drawn to and benefit from intelligent agents who are able to establish a relationship of trust and rapport while in conversation with them. How then can conversational agents be better designed to support or build rapport with human beings?

2 Related Work

So far, the majority of work done on to evaluate rapport-building conversational agents has focused on *envelope feedback*. There are two possible forms of feedback that a listener can provide to the speaker that will enhance the quality of the interaction. These are known as *envelope* versus *emotional feedback* for the case of nonverbal behaviours [2]. In *envelope feedback*, the response of the listener is unrelated to the actual content of the conversation, but rather serves as a comment or aid to the

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 419–425, 2012.

process of communication. Examples include head nods to indicate that the listener has heard and understood the message and eye gaze behavior that tracks the speaker's movements. For *emotional feedback*, the listener responds directly to the content of the conversation. This usually takes the form of an emotional display of some sort, such as facial expressions, verbal exclamations, narrative interjections and others.

Work on rapport-building conversational agents that focus on envelope feedback include the work of Bailenson and Yee which used an agent that mimicked user's head movements while reading a persuasive message to them [3], and Gratch and colleagues' RapportAgent which used postural mimicking, gaze behaviour and head movements while listening to the users describe a story [4, 5], and investigated various aspects of rapport including contingency [6] and comparisons with rapport among humans showing facial expressions [7].

For studies that have tried to build agents that can display emotional feedback or specific feedback that responds to the content of the conversation, Krumhuber et al. [8] studied the effects of genuine versus faked smiles of synthetic characters in a simulated job interview setting, and Bickmore and colleagues developed relational agents for helping users through a personal fitness training program [9], and for engaging in small talk as a real estate agent trying to determine user preference for a house [10]. Other studies have overlapped with similar fields such as human-agent empathy, in the contexts of sales persuasion [11] and student motivation in teaching scenarios[12].

3 Research Problem

Most of the studies that specifically look at rapport and building relationships with agents have focused on envelope feedback. There is as yet little information about how emotional feedback plays a role in establishing and maintaining rapport in a conversation with an agent. This study therefore seeks to answer the question: How is rapport affected when an agent displays emotional feedback? Modelled on the initial studies done by Gratch et. al in the creation of the RapportAgent [4], it focuses on dyadic interactions in a storytelling situation, where one participant retells a story to a listener. The key findings in the Rapport Agent studies were that people talked significantly longer and more fluently to an agent that responded with appropriate contingent envelope feedback than to an agent which displayed non-contingent envelope feedback. This study attempted to reproduce similar conditions to the Gratch study, but used emotional feedback (specifically in the form of facial expressions) instead of the envelope feedback that the Rapport Agent used. (For a more detailed description of the study, see [13].)

Similar to the Gratch study, for comparison purposes, this study sought to answer three research questions:

RQ1. How would a listening agent that displays appropriate emotional reactions affect the length of the conversation compared to an inappropriate agent?

RQ2. How would a listening agent that displays appropriate emotional reactions affect the fluency of the speaker compared to an inappropriate agent?

RQ3. How would a listening agent that displays appropriate emotional reactions affect the self-reported rapport felt by the speaker compared to an inappropriate agent?

4 Method

4.1 Participants

There were 36 people recruited in this study, 27 female and 9 male university undergraduates. To eliminate the impact of any pre-existing friendships on the rapport built during conversation, the participants were screened beforehand and paired to strangers or casual acquaintances at most.

4.2 Session Protocol

Participants were split into pairs and took turns to be a Speaker and a Listener, with the order chosen randomly. Participants were told a cover story that they were helping to evaluate a computer program that accurately captures all the movements and facial expressions of one person and displays them on screen (using an avatar) to another person. While the Listeners waited at the workstation showing the Speaker's face (Fig. 1), the Speakers were taken to another room and asked to watch a 3½-min Tom & Jerry cartoon. They were told that they would later be describing the story to the Listener, and that the system would be evaluated by the Listener's story comprehension. After the video, the Speakers were led to a console in an empty room and encouraged to describe the video they just saw to the avatar they believed represented the Listener's facial expressions, while the Listener observed them through videoconference software. After finishing the story, the Speaker filled out a post-experiment questionnaire, the participants changed roles, and the procedure repeated with different Tom and Jerry video clip of the same length.

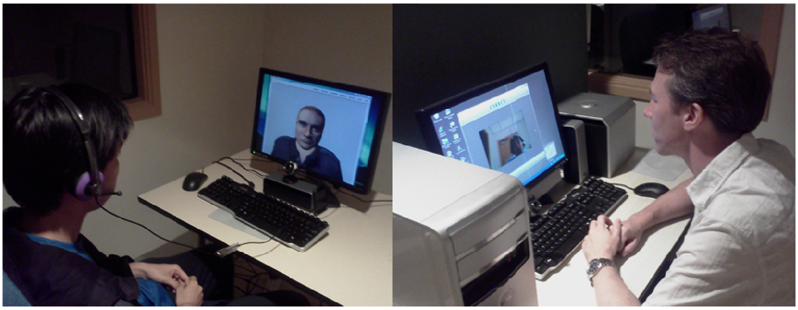


Fig. 1. Experimental Setup

4.3 Experimental Setup

In this study, an avatar was set up to display facial expressions as a person was talking to it. This avatar was controlled by a human Wizard-of-Oz, constrained to follow certain rules that mimicked the behavior of an intelligent agent. The avatar used in this study was originally developed by the Smartbody project [14]. It was modified for this study to be able to display two long-lasting moods as 'idle' animations, and several instantaneous emotional expressions, following the FACS system developed by Ekman et. al. [15, 16] and exaggerated slightly for effect. There were two experimental conditions for the avatar which was controlled by the agent: *appropriate* and *inappropriate* emotional feedback.

In the *appropriate* condition, the avatar would be displaying a small smile as an idle expression. The agent would also mirror the positive emotional expressions of the Speaker as they were speaking. It would also monitor backchannel opportunities and show either a broad smile or surprised expression, as relevant to the story. In the *inappropriate* condition, the avatar would have a slight frown as an idle expression. Furthermore, when the Speaker displayed a facial expression, or when the agent detected a backchannel opportunity, the agent had a 50% chance of not displaying any reaction at all, and another 50% chance of displaying an incongruous emotional reaction. The expressions used for incongruous reactions were sadness and puzzlement.

4.4 Data Gathering

At the beginning of the experimental session, basic demographic data was collected from the participants – age, gender and ethnicity. For behavioral measures, and to answer research questions 1 & 2, the Speaker was recorded via the webcam and microphone. Data collected can be grouped into three categories:

1. *Duration of interaction*: This includes total time to tell the story, the number of words in the story, and the number of meaningful words in the story.
2. *Speech fluency*: Two groups of measures are used – speech rate and amount of disfluencies. For speech rate, both overall speech and fluent speech rates (meaningful words per second) were measured. Likewise, for disfluencies, both disfluency rate (number of disfluencies per second) and disfluency ratio (number of disfluencies against overall word count) were measured.
3. *Self-report measures:* A post-experiment questionnaire was filled out by the Speaker, measuring their self-reported feelings for rapport.

5 Results

The primary interesting result was that participants exposed to an inappropriate condition spoke much longer than those who were faced with appropriate facial expressions, which was surprisingly opposite of the results from studies on envelope feedback. Other results showed that the differences in speech fluency were nearly negligible, and for rapport questionnaire, more participants reported positive indications of rapport for the appropriate condition than for the inappropriate condition.

Similar to the Gratch et. al study, because of the small sample size, non-parametric statistics were used to analyse the results: Mann-Whitney U for scale variables (length of interaction, speech fluency), and Chi-square for nominal variables (forced-choice questionnaire items). While none of the behavioral measures achieved statistical significance, there was a sharp difference with marginal statistical significance between the two conditions with regards to the Time Taken to tell the story (U = 103.0, p =

0.099). On average, those experiencing the Appropriate condition took shorter times to tell the story (67 secs) versus those in facing the Inappropriate agent (108 secs).

For the self-reported measures of rapport in the questionnaire, only two results were statistically-significant. Firstly, more people felt they were understood when faced with appropriate facial expressions than inappropriate ones (60% in the appropriate condition versus 10% of participants in the inappropriate condition, p = 0.024). Secondly, no one felt like they had a connection with the other person in the inappropriate condition, compared to about one-third in the appropriate condition who did feel some connection (p = 0.019).

6 Discussion

The results above raise several points for discussion. Although none of the behavioral results had full statistical significance, the overall trends for story length both in absolute time and word count indicate that the people who are faced with inappropriate facial expressions are taking *more* time and words to tell the story than people who are faced with appropriate facial expressions, which is different from the studies on envelope feedback. Reasons for this could include conversational grounding problems, social anxiety, or language barriers.

In the conversational grounding process [17], participants try to make sure that what has been said is what has been understood. Speakers can look for negative evidence – evidence that they have been misunderstood or misheard - or positive evidence that they have been understood. In the experiment, many of the Speakers who took part seemed hesitant in their speech, speaking in rapid bursts and then pausing for confirmation. This can be explained by the grounding process. In the appropriate condition, the agent would smile, laugh or look (pleasantly) surprised when backchannel opportunities came up. In all cases, this was a form of positive feedback that the Listener has understood. Furthermore, in the appropriate condition, there was no negative evidence to show that the agent did not understand the Speaker's utterances. However, in the inappropriate condition, the agent would only respond half the time when presented with backchannel opportunities. This could be taken by the Speaker as a sign of inattention, and thus a lack of positive evidence that the agent was sharing the same common ground. Furthermore, during the times that the agent did respond, it was with either a sad expression or a puzzled one. Both were negative evidences of misunderstanding. Thus, there was greater hesitancy and many more efforts at conversational repair with agent showing inappropriate emotional feedback than the agent with appropriate emotional feedback.

Another possible explanation for this result is the social anxiety. Social anxiety has been defined as a condition in which "some people, especially those who are shy or easily embarrassed, feel anxious in almost any situation in which they might be evaluated." [18] As the Speakers were clearly told that they would be recorded while they re-told the story of a video they only saw once to a stranger, it is quite possible that the experimental setup may have been ripe for social phobias to be triggered. Previous studies have shown that virtual agents could also cause anxiety in human users [19], and that an audience with hostile/bored facial expressions caused greater levels of anxiety than one showing appreciative expressions [20]. While not explicitly recorded, several participants did indicate that they were feeling anxious about the storytelling task – one or two even going so far as to request repeated viewing of the video before they were willing to go ahead and describe it to another person. Observations of the participants as they told stories indicated that participants in the inappropriate condition tended to speak jerkily – with rushed sentences followed by longer pauses – compared to those in the appropriate condition. This could have been an indication of the level of nervousness and embarrassment they felt. However, as this experiment did not control for social anxiety, further research would need to be done to disentangle this as a possible cause for the story length result.

A third explanation for the difference in story length is a possible language barrier in operation, which may have had opposite effects in the two conditions. The study was done in a university in Singapore, a multiracial, multilingual society which also serves as a transit hub for many international students within the region. As such, participants had a varying level of familiarity with English. A few were clearly translating from their native tongue to English, indicated by occasional usage of foreign words, followed by English equivalents. The lowered familiarity with English would tend to cut short unnecessary words in the story (thus explaining the relative brevity of the appropriate condition), but if they detect that the Listener is not understanding them due to the facial expressions, they would instead spend more effort and words to try and make themselves understood, which could account for much longer times in the inappropriate condition.

7 Conclusion

These results represent early findings on how rapport with a virtual character can be affected by emotional feedback in the form of facial expressions. The main results show that people faced with inappropriate emotional feedback spend more time and words to tell stories than people faced with appropriate feedback, which runs counter to findings in envelope feedback studies. Possible explanations include difficulties in conversational grounding, social anxiety, and language barriers. Further research would be necessary to refine the current study on facial expressions, and to look at the effects of the different factors uncovered in this study.

References

1. Tickle-Degnen, L., Rosenthal, R.: The Nature of Rapport and Its Nonverbal Correlates. Psychological Inquiry 1(4), 285–293 (1990)
2. Cassell, J., Thorisson, K.R.: The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. Applied Artificial Intelligence 13, 519–538 (1999)
3. Bailenson, J.N., Yee, N.: Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. Psychological Science 16, 814–819 (2005)

4. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S.C., 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)
5. 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)
6. Kang, S.-H., et al.: Does the contingency of agents' nonverbal feedback affect users' social anxiety? In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, Portugal, vol. 1. International Foundation for Autonomous Agents and Multiagent Systems, Estoril (2008)
7. Wang, N., Gratch, J.: Don't just stare at me! In: Proceedings of the 28th International Conference on Human Factors in Computing Systems. ACM, Atlanta (2010)
8. Krumhuber, E., et al.: Effects of Dynamic Attributes of Smiles in Human and Synthetic Faces: A Simulated Job Interview Setting. Journal of Nonverbal Communication 33, 1–15 (2009)
9. Bickmore, T., Picard, R.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interactions, 12(2), 293–327 (2005)
10. Bickmore, T., Cassell, J.: Relational Agents: A Model and Implementation of Building User Trust. In: CHI (2001)
11. Koster, T.: The Persuasive Qualities of an Empathic Agent. In: 5th Twente Student Conference on IT. University of Twente, Twente (2006)
12. Johnson, W.L., Rizzo, P., Bosma, W., Kole, S., Ghijsen, M., van Welbergen, H.: Generating Socially Appropriate Tutorial Dialog. In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 254–264. Springer, Heidelberg (2004)
13. Wong, J.W.-E.: Establishing Rapport with Conversational Agents: Comparing the Effect of Envelope versus Emotional Feedback. In: Department of Communications and New Media. National University of Singapore, Singapore (2012)
14. Thiebaux, M., et al.: SmartBody: Behavior Realization for Embodied Conversational Agents. In: Autonomous Agents and Multi-Agent Systems (AAMAS). International Foundation for Autonomous Agents and Multiagent Systems, Estoril (2008)
15. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press, Inc., Palo Alto (1978)
16. Ekman, P.: The directed facial action task. In: Coan, J.A., Allen, J.J.B. (eds.) Handbook of Emotion Elicitation and Assessment. Oxford University Press, Oxford (2007)
17. Clark, H.H., Brennan, S.E.: Grounding in Communication. In: Levine, L.B.R.J.M., Teasley, S.D. (eds.) Perspectives on Socially Shared Cognition, pp. 127–149. American Psychological Association, Washington D.C (1991)
18. Myers, D.: Social Psychology. McGraw-Hill College (1999)
19. Rickenberg, R., Reeves, B.: The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In: SIGCHI Conference on Human Factors in Computing Systems, The Hague, The Netherlands (2000)
20. Pertaub, D.-P., Slater, M.: An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. Presence: Teleoperators and Virtual Environments 11(1), 68–78 (2001)

Characters with Personality!

Karel Van den Bosch[1], Arjen Brandenburgh[2], Tijmen Joppe Muller[1],
and Annerieke Heuvelink[1]

[1] TNO
{karel.vandenbosch,tijmen.muller,annerieke.heuvelink}@tno.nl
[2] VU University Amsterdam
arjen.brandenburgh@gmail.com

Abstract. Serious games offer an opportunity for learning communication skills by practicing conversations with one or more virtual characters, provided that the character(s) behave in accordance with their assigned properties and strategies. This paper presents an approach for developing virtual characters by using the Belief-Desire-Intentions (BDI) concept. The BDI-framework was used to equip virtual characters with personality traits, and make them act accordingly. A sales game was developed as context: the player-trainee is a real-estate salesman; the virtual character is a potential buyer. The character could be modeled to behave either extravert or introvert; agreeable or non-agreeable; and combinations thereof. A human subjects study was conducted to examine whether naïve players experience the personality of the virtual characters in accordance with their assigned profile. The results unequivocally show that they do. The proposed approach is shown to be effective in creating individualized characters, it is flexible, and it is relatively easy to scale, adapt, and re-use developed models.

Keywords: intelligent agents, behavior modeling, training, personality, virtual characters, serious gaming, BDI.

1 Introduction

This paper is about a method for developing intelligent virtual characters with a flexible and easily expandable domain of discourse to be used for conducting dialogues with a human player in serious games. In particular it is about modeling the personality characteristics of a virtual character in such a way that it behaves in accordance with its assigned personality.

Serious games offer an opportunity to develop a contextually rich and flexible environment for training skills (e.g. Korteling et al., in press; Michael, 2006). They are considered to be effective for learning tasks that rely upon communication skills, like legal profession, project management, human resources (e.g. negotiating employment terms: Broekens et al., 2012), intelligence, sales, and many other professions (e.g. Core et al., 2006). Communication skills can be practiced in serious games by conducting conversations (e.g. interviews, sales conversations, negotiation talks) with one or more virtual characters (Cassell et al, 2003).

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 426–439, 2012.

Such practice conversations usually have one or more objectives, like (a) familiarizing the player-trainee with various types of conversational partners (e.g. cooperative versus intractable; empathic versus reserved; extravert versus introvert), (b) enabling practice in various conversational strategies (e.g. fact-driven versus emotion-driven); and (c) make the player-trainee experience how different communication styles (e.g. asking rather than telling; providing facts instead of opinions) affect the partner and the course of the conversation.

Although agents evidently do not possess human properties, users perceive agents to have these qualities when interacting with them (Reeves and Nass, 1996). Moreover, people do so automatically and mindlessly. The convincingness of a virtual character to a human interlocutor is strongly influenced by its functionalities, and very little by its features (Catrambone et al., 2004).

The quality of the virtual characters in the game determines the quality of the learning situation. Proper training requires characters that behave in accordance with the assigned properties and strategies, and use these to respond to player-trainee actions in a consistent manner. Virtual characters that fail to act in a consistent and believable fashion, and that do not respond in a logical way to their conversational partner may bring about a training with no or limited learning opportunities. It is even possible that it induces the player-trainee to adopt inappropriate conversation strategies, resulting in negative transfer of training.

The classic approach towards developing dialogue in serious games is to make use of scripts. In scripted dialogue, communications of a virtual character (from now on called "Non-Playing Character", or NPC) and the player-trainee are predefined as option sets. An NPC selects its response based on the option chosen by the player. This approach usually results in (large) decision trees that define both flow and content. The advantage of scripting is that it is a robust technique, and it allows complete control over the dialogue. However, if learning requires more than simple conversations, the dialogue tree grows exponentially, which makes adding new content both complex and laborious. Furthermore, the deterministic nature of scripted conversation often results in stereotyped and rigid non-playing characters, leading to a predictable game experience.

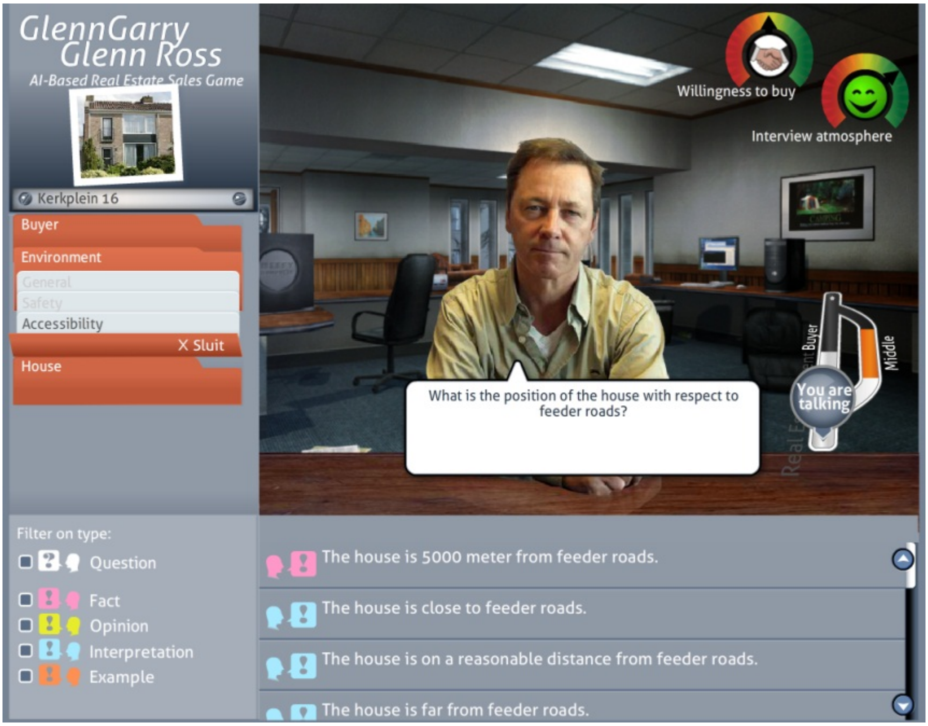
An alternative to scripting dialogues is the Belief-Desire-Intention (BDI) model of human behavior, proposed by Bratman (1987). In a BDI model, a character is not instructed to act upon a statement in the conversation, but rather upon the *belief* that the statement brings about. The belief then triggers a goal in the Non-Playing Character. What goal is triggered depends upon the context, and upon the properties of the NPC. For example, an agreeable NPC may decide to comply with a partner's suggestion to change the topic, whereas a non-agreeable person may decline the suggestion and hold on to the current topic. The advantage of BDI over scripts is that behavior is more flexible and reusable. It has been shown that with BDI it is possible to develop intelligent Non-Playing Characters that behave autonomously and realistically in simulations and games (Shenandarkar et al., 2006; Van den Bosch et al., 2009).

The research presented here makes use of the BDI approach to model Non-Playing Characters equipped with (combinations of) personality traits that make them behave in a fashion that is typical for individuals with such a personality profile.

The next section describes the game that is used as research context, followed by an account of the personality traits selected for the study and how these traits have been accommodated in the model. The model is tested experimentally in a human subjects study. The implications of the results are discussed in the final section.

2 The Game

A 'sales' game was developed, inspired by the Glengarry Glen Ross movie (1992). In the movie, four salesmen working at a real-estate agency become desperate when the corporate announces that all except the top two salesmen will be fired. Superior sales skills (e.g. listening, persuading, negotiating) are of the essence. The player in the present GGR-game is a real-estate salesman; a lead (a potential buyer) is a BDI-based NPC. See Figure 1 for an impression.



What is the position of the house with respect to feeder roads?

The house is 5000 meter from feeder roads.

The house is close to feeder roads.

The house is on a reasonable distance from feeder roads.

The house is far from feeder roads.

Fig. 1. Impression of the game's interface, showing the dossier with discussion topics (left), the player's communication options for the selected topic (below), and the NPC (potential buyer) that just asked a question and awaits a reply (middle right).

The principles of the game will be explained below. An elaborated account of the game can be found in Muller et al., [submitted]. Each NPC has its own belief base (e.g. knowledge of the house in question, wishes/demands), goal base (e.g. requesting information, deciding whether to buy the house, choosing topics to discuss), and plan base (strategies to achieve its goals). In a sales conversation, the player can ask questions to discover the NPC's wishes and can communicate information, interpretations and opinions in order to convince the NPC that the house suits its needs. The player can do so by emphasizing qualities of the house desired by this buyer, or by providing appropriate anecdotic material. The NPC uses its sets of beliefs, goals, and strategies to respond. But it is not only the player who can take the initiative in the dialogue. The NPC can also initiate communication. For example, the NPC can ask and answer questions; give opinions, take a decision to visit the house; terminate the conversation; etc. As both the player as the NPC can take the initiative in the dialogue, a 'turn-taking' mechanism was developed (visible on the right in Figure 1).

2.1 Domain Ontology

A large number of concepts related to 'house-buying' are represented in a predicate ontology, providing the structure of topics that the player and NPC can talk about. These, of course, include properties of the house (e.g. number of rooms; surface area; maintenance state), but also other topics that are typically addressed in house-buying negotiations (e.g. safety of the neighborhood; access to public transport/ motorways, etc.). Both the player and the NPC can refer in their communications to any of the concepts defined in the ontology.

2.2 Building Blocks

In order to assign meaning to the lemmas in the predicate ontology, a series of 'building blocks' were developed: *Fact*, *Interpretation*, *Opinion*, *Wish*, *Importance*, *Illustration*, and *Argumentation*. These building blocks define the dialogue.

- The *Fact* building block defines information about a given predicate (e.g. kitchen has a surface area of 12 m^2). Facts can be used by both player and NPC.
- *Interpretation* defines a subjective value about a predicate (e.g. kitchen has a large surface area). Interpretations can be used by both player and NPC.
- *Opinion* represents an opinion of either player or NPC (e.g. kitchen is fine), expressed by a number between 0 and 1.
- A *Wish* defines the desired value or range of values of a predicate and is used by the NPC only. A wish can be expressed as fact values (e.g. KitchenSurface [11, 20]) or interpretation values (KitchenSurface, [large, very large]).
- *Importance* defines the importance of the predicate to the NPC. It expresses how much importance the NPC attaches to the predicate (e.g. KitchenSurface) when judging the house.
- *Argumentation* is a building block to be used by the player only. It can be used as instrument to influence the NPC's opinion on a specific predicate. For example, the player argues "the nosy neighbor increases neighborhood safety". This may indeed affect the NPC's opinion on safety. However, it may also affect its opinion on tranquility.

- *Illustration* is a remark to be used by an NPC to provide hints to the player about its underlying motives. For example, an NPC may respond to the message that the house has an open kitchen with "Great, so I can keep chatting with my friends in the living room while cooking."

The following communication types were distinguished: *Tell*, *Ask*, and *Acknowledge*. This means that a player or NPC could tell a fact, wish, interpretation, opinion, etc., they could ask for them, and they could acknowledge a given fact, wish, etc. in a number of ways.

Sentence templates are used to interface message types, building blocks, and predicates in complete sentences to the player. For example:

```
Ask(Fact(FeederRoads))
```

uses the following sentence template to bring across the NPC's question to the player:

"What is the position of the house with respect to feeder roads?"

Likewise, the sentence template

"The house is close to feeder roads."

is used as player's dialogue option to convey the following information to the NPC:

```
Tell(Interpretation(FeederRoads, nearby))
```

The NPC continuously updates three parameters indicating the result of the game play. These involve: (1) its opinion on the suitability of the house in question, (2) its opinion about the conversation, and (3) the status of its information need. It is the task of the player to enquire about the NPC's wishes, to maintain a good atmosphere (to make the NPC find the conversation pleasant), to deliver relevant and positive information (to provide the NPC's with the needed information, and to establish a positive attitude in the NPC towards the house). The interface displays the status of the first two parameters to the player (see upper right corner of Figure 1); the status of the NPC's information need indicates whether the NPC is ready to take a decision (to visit the house or to place a bid) or whether it desires more information. The status of information need is not shown to the player.

3 Modeling Personality

Personality can be defined as a dynamic and organized set of characteristics that uniquely influences a person's behavior in various situations. In sales training, learning to recognize and utilize the relationships between personality and behavior is usually an important part of the program (e.g. McFarland et al., 2006; Sujan et al., 1988). Serious gaming offers the opportunity to familiarize trainees in a realistic setting with different simulated personalities. Furthermore, trainees may use such games to try different approaches in communicating with the various simulated personalities, experience the outcomes, and adjust their strategies accordingly. However, a precondition for meeting the potential of serious games is that the behavior of the modeled personalities accurately reflects the behavior of real humans with such a personality profile.

The question is then: is it possible to model NPCs in such a way that they behave in accordance with their assigned personality?

The literature shows that the study of personality has a broad and varied history in psychology, with an abundance of theoretical traditions. It is obviously beyond the scope to review all proposed dimensions and categorizations of personality here. Instead, the purpose of the present study is to investigate whether it is possible to capture one, or a few, well known personality traits in our behavior model of the NPC, and to examine whether the model generates behavior for the NPC that people experience as characteristic for that personality trait.

The literature shows interesting previous work. Bevacqua et al. (2010) used Eysenck's three-factor model to develop Sensitive Artificial Listeners that in dialogues respond in accordance with their assigned personality by means of non-verbal behavior (e.g. facial expressions, back channeling). In contrast, our study attempts to express personality through verbal behavior in human-agent dialogue.

Bostan (2010) uses a needs framework to define a series of 27 'psychological needs' as the driving force of behavior. Needs are affected by many factors, like feelings, desires, emotions, and also personality traits. Personality Inventories are used to build character profiles, like 'male dominant leader', or 'character of trustfulness'. In our approach, traits are used directly to generate behavior rather than indirectly through needs.

3.1 Selection of Personality Traits

Perhaps the best known framework of personality is 'the Big Five' (Costa & McCrae, 1992). The Big Five factors are openness, conscientiousness, extraversion, agreeableness, and neuroticism. Conscientiousness is exemplified by being disciplined, organized, and achievement-oriented. Neuroticism refers to degree of emotional stability, impulse control, and anxiety. Extraversion represents a high degree of sociability, assertiveness, and talkativeness. Openness is characterized by a strong intellectual curiosity and a preference for novelty and variety. Finally, agreeableness refers to being helpful, cooperative, and sympathetic towards others. Of course, the five traits are a gross categorization; they represent human personality at a broad level of abstraction.

From the Big Five, we selected *extraversion* and *agreeableness* to be modeled. The two traits were selected because they distinguishably affect behavior in a conversation: extravert personalities (compared to introvert people) use more social and positive language. They give more compliments and are more focused on reaching agreement. Furthermore, they tend to produce less formal but more complex sentences (Dewaele et al., 1999; Furnham, 1990). With respect to agreeableness, Mehl (2006) showed that agreeable persons explicitly communicate their understanding, are more open to suggestions of others, and are more likely to express appreciation (Graziano & Eisenberg, 1997). Mairesse et al. (2007) found evidence that agreeableness is a trait that can be observed by others.

3.2 Using Personality Traits to Shape Conversation

The next step is defining how the two selected traits affect the nature of conversation. Three main aspects of a conversation can be distinguished: *Form*, *Content* and *Strategy*. The form of a conversation refers to how sentences are being derived and put together, and how sentences are being pronounced. An extravert personality is, for example, likely to have a poorer lexical composition than an introvert personality. The content of communication refers to the information being exchanged. For example, a conversation may cover many topics or only a few; exchange of information may pertain to social information or domain-specific, and so on. Strategy, finally, defines how persons act and respond to the actions of their conversation partner. It determines, for instance, the topic(s) being addressed, initiative to shift to topics, to disclose or conceal certain information.

In a spoken dialogue, the form affects the nature of conversation significantly. For example, a friendly message spoken in a sarcastic tone can bring across a totally different meaning than what is actually being said. Form can also affect textual communication. In general, form can have a large impact on the nature of conversation, while the instruments to shape messages can be very subtle (e.g. tone, prosody, sarcasm, double-denial, etcetera). It was considered that it is still too difficult to model the use of these instruments to shape the nature of conversation as a function of personality trait. It was therefore decided not to include the aspect of form in the model, but instead focus on modeling how the selected personality traits affect the NPC's choice of content and strategy of communication.

3.3 Implementation of Personality Traits

This section addresses the way the selected personality traits extraversion and agreeableness have been modeled to affect the communication behavior of the NPC in the sales conversation (see Brandenburgh, 2012). In order to prevent stereotyped and rigid characters, the influence of a personality trait on behavior has been modeled in terms of probability. Thus, an extravert NPC is just more likely to spontaneously tell details than an introvert NPC, it is not a certainty.

Extraversion

Selection of communication type: Extravert NPCs are more likely to tell information; introvert NPCs are more likely to ask for information.

Selection of information type: Extravert NPCs are more likely to express wishes and opinions (e.g. "I'd like the house to be easy to reach by public transportation."); introvert NPCs are more likely to tell, or ask for, facts (e.g. "How many minutes walking distance is the house from public transportation?").

Selection of type of acknowledgment: Extravert NPCs are more likely to disclose their opinion implicitly when giving an acknowledgement for received information ("Great!"), whereas introvert NPCs are more likely to just acknowledge that the received information is understood ("Okay.").

Selection of wish expression: Extravert NPCs are more likely to generate a wish in subjective, interpretative terms (e.g. “I’d like a large bathroom.”), whereas introvert NPCs are more likely to express wishes in factual terms (e.g. “I’d like the bathroom to be $12m^2$.”).

Selection of detail of information: Extravert NPCs are more likely to provide details (e.g. opinions, illustrations) when telling information or answering questions (e.g. “That's rather large for two persons, a bit of a waste of the space.”, after acknowledging the information that there is a large bathroom); introvert NPCs are more likely to stick to pure facts.

Selection of type of opinion: Extravert NPCs are more likely to select a positive opinion (spontaneously or when asked for, e.g. “I think the condition of the bathroom is great.”); introvert NPCs are more likely to select a negative opinion (e.g. “I think the condition of the kitchen is not all that good.”). Clearly, the NPC will always share their opinion truthfully, so an extravert NPC will simply not select the opinion on a negatively valued fact; vice versa for introvert NPCs.

Agreeableness

Selection of information content: Agreeable NPCs are more likely to discuss the house or the environment; non-agreeable NPCs are more likely to tell about themselves.

Commitment to current topic: Agreeable NPCs are more likely to continue the conversation on the current conversation topic even if they find another topic more important (e.g. asking questions about the bathroom facilities, surface, condition, et cetera); non-agreeable NPCs are more likely to propose a new topic for discussion (e.g. moving from bathroom surface to kitchen condition).

Tractability: Agreeable NPCs are more likely to accept when the player proposes a topic change (e.g. switching from kitchen to bathroom); non-agreeable NPCs are more likely to refuse such requests.

Compliance: Agreeable NPCs are more likely to adopt a player’s opinion as their own; non-agreeable NPCs are more likely to stick to their own opinion.

Selection of type of acknowledgment: Agreeable NPCs are more likely to acknowledge received information (e.g. by giving a neutral acknowledgement like “Okay.” or, more probable, a qualitative acknowledgment like “Great!” or “That’s too bad.”); non-agreeable NPCs are more likely to refrain from acknowledgments.

Table 1 gives an impression in pseudo code of how this is formalized.

4 Evaluation

The model outlined above is intended to generate behavior for NPCs that is in accordance with the assigned personality profile in terms of the traits extraversion, agreeableness, and combinations thereof. The question is, of course, do people recognize the assigned personality when interacting with such characters?

Table 1. Impression of formalization of **Selection of information type** in pseudo code

```
if (agreeable) then             /* Boolean based on probability
  if (extravert) then             /* Boolean based on probability
    if (probability .5 ) then     /* With equal chance:
      return: tell(wish)          /* Tell a wish
    else
      return: tell(opinion)       /* Tell an opinion
    end if
  else                            /* Act introvert:
    return: tell(fact)            /* Tell fact about the house
  end if
else                              /* Act non-agreeable:
  return: tell(fact)              /* Tell fact about itself
end if
```

When complex characteristics are incorporated into an agent framework (such as emotions or personality), an appropriate approach has to be selected in order to evaluate the effects of these modifications. Merely running scenarios is not sufficient to establish the effects. Norling et al. (2002) argue that data from subjective judges must be collected and analyzed statistically. This approach was used in this study.

The question whether people recognize the assigned personality in virtual characters was investigated in a human subjects experiment. Subjects played the game several times. In each session, the buyer had a different personality profile. After each game session, participants were asked about their impression of the buyer. The hypothesis is that that player's impression matches the implemented personality traits of the buyer agent.

4.1 Methods

Subjects. 30 (under)graduate students (18 male, 12 female) were recruited from the VU University Amsterdam. Their mean age was 21.7 (SD = 2.45). Subjects were paid 30 Euro for participation.

Materials. For this study, it is important that the subject forms his or her impression of the NPC on its behavior, not on its appearance. For example, a friendly face may make the player feel that the buyer is agreeable. In order to eliminate appearance as contaminating factor, the game's interface just showed the NPC's silhouette.

Design. A 4x2 factorial within-subjects design was used. Orthogonal combinations of agreeableness and extraversion produced four buyer NPCs, each with a different personality profile, see Table 2. Subjects played the game four times: each time the buyer NPC had a different personality profile. The order was randomized for each subject.

Table 2. Values indicate the probability that the NPC shows extravert or agreeable behavior

Condition	Extraversion	Agreeableness
Low_E - Low_A	.1	.1
High_E - High_A	.9	.9
Low_E - High_A	.1	.9
High_E - Low_A	.9	.1

Procedure. Subjects were told that were going to play a series of game sessions and that after each session they would be asked about their impressions of the character in the game. A short familiarization session, with the NPC modeled as neutral in terms of agreeableness and extraversion, was used to introduce the game to the subjects.

Then the experiment proper started. Each game session took about 10-15 minutes to complete. After completing a session, subjects were asked to fill out a personality questionnaire. In addition, a series of statements regarding the (kind of) NPC behaviors were presented to the subject. The subject was asked to rate the applicability of the statements for the agent in that particular session. Then the next session of the game was started. This continued until all four games session were completed. At the end of the experiment, subjects were debriefed.

Measurements. To assess how the player evaluated the NPC's personality, the Dutch translation of the HEXACO questionnaire was used (Lee & Ashton, 2009; De Vries et al., 2009). Only the 32 items measuring extraversion and agreeableness were used. Subjects rated items on a 5-point Likert scale. They were specifically instructed to use the extremes of the scale when they considered this appropriate. If subjects argued that the experience during the gameplay was insufficient to form an adequate impression, they were told to score to the best of their ability.

For exploration purposes, subjects were presented a series of statements regarding the NPC's behavior and asked to rate the applicability on a 5-point Likert scale. Although the statements were unfortunately not designed to systematically capture the manipulated behaviors (see §3.3), it was afterwards possible to connect statements and manipulations. For example, the statement "the buyer took initiative" was considered to reflect extraversion. The statement "the buyer asked a lot of questions" was consistent with introvert behavior (extravert buyer agents, by design, spontaneously reported information; introvert buyer agents were more inclined to ask questions). "The buyer behaved friendly" was considered to reflect agreeableness, and "the buyer made clear what he wanted" was considered to indicate non-agreeableness (see 'tractability' in §3.3).

4.2 Results

For all 30 participants, the mean score on extraversion and agreeableness was calculated for each experimental condition. There were no missing data. Figure 2 shows the results. An analysis of variance was executed with the mean scores on extraversion and agreeableness as dependent variables, and with condition (4) as within-subjects variable. Results show that scores differed significantly over conditions (F=26.43; *df* =3; p<.01).

The mean scores on extraversion were calculated for conditions 2 and 4 (high extraversion, HE), and for conditions 1 and 3 (low extraversion, LE), and entered in paired t-tests. Likewise, the mean scores on agreeableness were calculated for conditions 2 and 3 (high agreeableness, HA), and for conditions 1 and 4 (low agreeableness, LA), and also entered in paired t-tests.

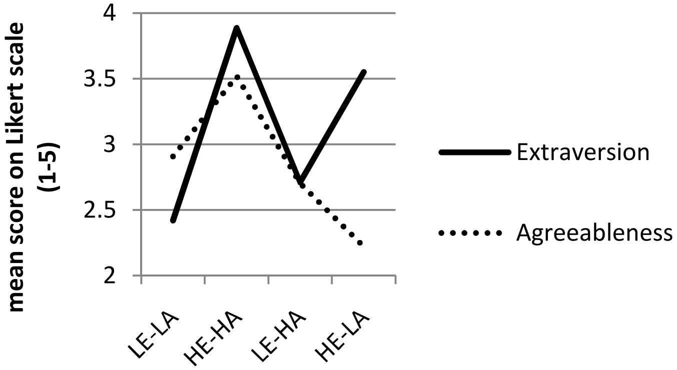


Fig. 2. Mean ratings on Extraversion and Agreeableness, split by condition

Results show that players rated high-extraversion characters as considerably more extravert than low-extraversion characters (3.72 versus 2.56), a significant difference ($t(29)=8.22$; $p<.01$). Furthermore, players rated high-agreeable characters as considerably more agreeable than low-agreeable characters (3.11 versus 2.56). This too was a significant difference ($t(29)=2.63$; $p<.05$).

The results above confirm that subjects perceive NPCs in accordance with their assigned personality traits. Do subjects also recognize the constituent behaviors, indicative for these personality traits, in their interactions with the NPCs? This question was addressed by examining the data on the rated statements. Table 3 shows the results.

Subjects indicated to recognize the behaviors typical for extraverts more in extravert buyer agents ($t(29)=16.2$, $p<.01$) and found behaviors that are typical for introverts to be more applicable to introvert agents than to extravert buyer agents ($t(29)=11.2$, $p<.01$). Furthermore, subjects indicated to recognize the behaviors typical for agreeable persons more in agreeable buyer agents ($t(29)=3.3$, $p<.01$). Behaviors that are typical for non-agreeable persons were not considered to be more applicable to non-agreeable buyer agents.

Table 3. Mean Likert ratings for statements concerning extravert, introvert, agreeable, and non-agreeable behavior

	ratings for statements concerning ... behavior			ratings for statements concerning ... behavior	
Condition ↓	**extravert**	**introvert**	Condition ↓	**agreeable**	**non-agreeable**
HE	4.65	4.45	HA	3.55	2.84
LE	1.88	2.08	LA	2.77	2.98

4.3 Discussion

This study was to test the question whether the model was able to correctly express personality traits into behavior. The results unequivocally show that players recognize the assigned personality of the character when interacting with them. Only one measurement was at variance. In the LE-HA condition, the players rated the NPC at a fairly low level of agreeableness, where a high score on this trait was expected (see Figure 2). Inspection of individual data, nor observations of the experiment leader, revealed an explanation for this unexpected result.

The HEXACO Observer Report proved to be an appropriate instrument for measuring how participants experience the behavior of their conversational partner in terms of personality traits. Although a number of subjects said in the debriefing that they found it difficult to rate the NPC given the limited exposure, the results show that the instrument is sensitive enough to detect and categorize even partial impressions.

5 Conclusion

Serious games are an immersive and attractive vehicle for learning and practicing communication skills required for many professions. Games may, for example, be used to train people in conducting sales conversation skills through dialogues with virtual characters. In such applications, the quality of the Non-Playing Character is decisive for the quality of the game as a learning tool. If characters act according to their modeled properties and strategies, this gives the game the potential for being a good learning environment.

In this paper we proposed the BDI-approach to create individualized NPCs, each with their own personality profile. By defining relationships between the building blocks that underlie the NPC's actions (e.g. Facts, Wishes, Opinion, et cetera), its modes of communication (Tell, Ask, and Acknowledge), and its personality traits (defined as parameters on the dimensions 'extraversion' and 'agreeableness'), we were able to construct NPCs that act in accordance with their personality profile. The results of the experimental study support this claim.

The strength of using BDI over the alternative method of scripting behavior is its flexibility. If a trait like extraversion is assigned to a character through scripting, then separate decision trees have to be worked out in depth for extravert and introvert NPCs. Furthermore, scaling the level of a trait like extraversion would be a difficult, if not unfeasible, job. BDI instead allows the developer to use the aforementioned relationships between traits and the building blocks of behavior in a more general fashion, making it easier to scale, adapt, and re-use the developed model.

The proposed approach opens up various paths for future research. One possibility is to expand the number of personality traits and to validate the models with respect to those constructs (does NPC behavior indeed represent the modeled (mix of) traits?) and to purpose of application (do the developed NPCs adequately prepare trainees to deal with various personalities in real-life communication?). Another possibility is to investigate whether the architecture proposed here can also be used for modeling other properties affecting behavior, like emotion. Acknowledging the need for future work, the present research may prove an important step towards developing individualized NPCs, each equipped with their own traits, properties, and strategies.

Acknowledgement. This research was partly funded by the GATE research program, http://gate.gameresearch.nl/.

References

Bevacqua, E., de Sevin, E., Pelachaud, C., McRorie, M., Sneddon, I.: Building credible agents: Behaviour influenced by personality and emotional traits. In: Proc. of the Int. Conf. on Kansei Engineering and Emotion Research, KEER 2010 (2010)

van den Bosch, K., Harbers, M., Heuvelink, A., van Doesburg, W.: Intelligent Agents for Training On-Board Fire Fighting. In: Duffy, V.G. (ed.) ICDHM 2009. LNCS, vol. 5620, pp. 463–472. Springer, Heidelberg (2009)

Bostan, B.: Explorations in Player Motivations: Virtual Agents. In: Proceedings of the ICEC Conference, Seoul, Korea (2010)

Brandenburgh, A.: Influence of Personality on the Behavior of Conversational Agents. Master's Thesis, VrijeUniversiteit, Amsterdam (2012)

Bratman: Intentions, Plans and Practical Reason. Harvard University Press, Cambridge (1987)

Broekens, J., Harbers, M., Brinkman, W., Jonker, C., van den Bosch, K., Meyer, J.J.C.: Virtual Reality Negotiation Training Increases Negotiation Knowledge and Skill. In: Proceedings of the Conference on Intelligent Virtual Agents (IVA), Santa Cruz, CA (accepted for publication)

Cassell, J., Bickmore, T.: Negotiated collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. User Modeling and User-Adapted Interaction 13(1), 89–132 (2003)

Catrambone, R., Stasko, J., Xiao, J.: ECA as User Interface Paradigm:Experimental Findings within a Framework for Research. In: Pelachaud, C., Ruttkay, Z. (eds.) From Brows to Trust:Evaluating Embodied Conversational Agents, pp. 239–267. Kluwer Academic Publishers, Dordrecht (2004)

Core, M., Traum, D., Lane, H.C., Swartout, W., Gratch, J., van Lent, M., Marsella, S.: Teaching Negotiation Skills through Practice and Reflection with Virtual Humans. Simulation 82(11), 685–701 (2006)

Costa, P.T., McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Psychological Assessment Resources, Odessa (1992)

Dewaele, J.M., Furnham, A.: Extraversion: the Unloved Variable in Applied Linguistic Research. Language Learning 49(3), 509–544 (1999)

Furnham, A.: Language and personality. Handbook of Language and Social Psychology (1990)

Graziano, W.G., Eisenberg, N.: Agreeableness: aDimension of Personality. In: Hogan, R., Johnson, J., Briggs, S. (eds.) Handbook of Personality Psychology, pp. 795–824. Academic Press, San Diego (1997)

Korteling, J.E., Helsdingen, A.S., Theunissen, N.C.M.: Serious gaming @ work: Learning Job-Related Competencies using Serious Gaming. In: Derks, D., Bakker, A.B. (eds.) The Psychology of Digital Media @ work. Psychology Press, London (in press, 2012)

Lee, K., Ashton, M.C.: The HexacoPersonality Inventory-Revised (2009), `http://www.hexaco.org`

Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research 30(1), 457–500 (2007)

McFarland, R.G., Challagalla, G.N., Shervani, T.A.: Influence Tactics for Effective Adaptive Selling. Journal of Marketing 70(4), 103–117 (2006)

Michael, D.: Serious games: Games that educate, train, and inform. Thomson Course Technology, Boston (2006)

Muller, Heuvelink, Swartjes, Van den Bosch: A BDI model for Open Dialogue, Glengarry Glen Ross (submitted)

Norling, E., Soneberg, L.: An Approach to Evaluating Human Characteristics in Agents. In: Proceedings of the RASTA 2002 Workshop, AAMAS 2002 (July 2002)

Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press (1996)

Shendarkar, A., Vasudevan, K., Lee, S., Son, Y.: Crowd Simulation for Emergency Response using BDI Agent Based on Virtual Reality. In: WSC 2006, Proceedings of the Winter Volume, pp. 545–553 (2006)

Sujan, H., Weitz, B.A., Sujan, M.: Increasing Sales Productivity by Getting Salespeople to Work Smarter. Journal of Personal Selling & Sales Management, 9–19 (1988)

de Vries, R.E.: Ashton en Kibeom Lee, M.C.: De Zes Belangrijkste Persoonlijkheidsdimensies en de HexacoPersoonlijkheidsvragenlijst. Gedrag&Organisatie 3 (2009)

A Formal Architecture of Shared Mental Models for Computational Improvisational Agents

Rania Hodhod[1,2], Andreya Piplica[1], and Brian Magerko[1]

[1] School of Literature, Communicationa and Culture, Georgia Institute of Technology
[2] Faculty of Computer and Information Sciences, Ain Shams University
{rhodhod,piplica,magerko}@gatech.edu

Abstract. This paper proposes a formal approach of constructing shared mental models between computational improvisational agents (*improv agents*) and human interactors based on our socio-cognitive studies of human improvisers. Creating shared mental models helps improv agents co-create stories with each other and interactors in real-time interactive narrative experiences. The approach described here allows flexible modeling of non-Boolean (i.e. fuzzy) knowledge about scene and background concepts through the use of fuzzy rules and confidence factors in order to allow reasoning under uncertainty. It also allows improv agents to infer new knowledge about a scene from existing knowledge, recognize when new knowledge may be divergent from the other actor's mental model, and attempt to resolve this divergence to reach cognitive consensus despite the absence of explicit goals in the story environment.

Keywords: Improv agents, shared mental models, computational creativity.

1 Introduction

While there have been systems that have explored improv theatre as a model for creating interactive narratives [1, 2, 3], to the extent of our knowledge none have focused on the co-construction of story between an AI and a human interactor (i.e. where neither agent has privileged or pre-authored story knowledge). Achieving this goal requires the agents to be able to construct *shared mental models* (i.e. shared understandings about scene content) while collaboratively performing an improvised story. Shared mental models (SMMs) are a cognitive construct that incorporates the development of mutual beliefs from individuals' mental models until a common mental model is reached by the group, either explicitly or implicitly [5, 6, 7]. Agents in a co-constructive interactive narrative also must be able to reason about ambiguous knowledge in an uncertain environment and to reach a shared understanding about scene elements with the other actor *without any collaboration outside of the performance*. We have developed an interactive narrative within the domain of the Old West based on the improv game *Three Line Scene*, which focuses on establishing the *platform* (i.e. the characters, setting, and joint activity of a scene) in only three lines of dialogue. *Three Line Scene* allows users to provide gestural input through Kinect to an AI-controlled

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 440–446, 2012.

avatar that is in a scene with another AI-controlled character. This paper describes a formal approach to shared mental models for interactive narrative agents that is flexible enough to allow human interactors to act as an equal co-creator in an improvised scene.

2 Shared Mental Models in Improvisation

The ambiguous actions in a scene (e.g. if one actor holds their fists one on top of the other and moves them from side to side, another actor could interpret this as either raking or sweeping among other possibilities) and the ease with which they can be misinterpreted can cause an improviser to develop a mental model that differs from the other improvisers' models. The state where improvisers' mental models differ is called *cognitive divergence* [4]. Improvisers repair their divergent mental models to reach a state of *cognitive consensus*, where everyone either implicitly or explicitly agrees on a shared mental model, through the process of *cognitive convergence*[1] [6]. Cognitive consensus can be thought of as the process of "getting on the same page."

Improvisers employ *repair strategies* to deal with divergences as they occur [4], which are either *other-oriented* or *self-oriented***.** Other-oriented repair strategies aim to affect another's mental model through presenting new concepts (*presentation*) or correcting divergences (*clarification*). Self-oriented repair strategies try to align one's own mental model with those of others. For example, an actor may state an unsure idea about what is going on in the scene so that others may confirm it (*verification*). Alternatively, the actor may introduce new, vague information to the scene to observe how the others react, hoping that this will reveal some enlightening information (*blind offer*). Repair strategies help improvisers update their mental models and approach a cognitive consensus that reflects their common understanding regarding how key issues are defined and conceptualized, which is essential in story co-creation.

3 Computational Shared Mental Models

Improvisers interact through the process of proposing and responding to *offers* (i.e. proposals made by improvisers in a performance to add something to a scene) [8]. While making or responding to offers, an improviser is able to model other actors' mental models in the scene, evaluate the outcome of actions, and update goals, which can be referred to as theory of mind [6].

Based on our understanding of how human improvisers construct shared mental models in an improvised scene, we developed a shared mental model for interactive narrative agents that preserves the above-mentioned cognitive behavior related to theory of mind. The model consists of three components, as described below: beliefs, commitments, and reasoning and decision making modules.

The *beliefs component* models the agent's beliefs about itself, about others, and about others' beliefs (see Section 4). Those beliefs are associated with confidence factors that show how much the agent thinks its belief is correct (see Section 4). Confidence factors (CFs) are fuzzy values (degrees of membership values on a scale from 0 to 1) that allow the agent to compute the strength of its beliefs in a specific world frame, such as the emerging platform (i.e. initial scene details). CFs guide the actions that the agent takes either towards advancing the scene (by adding new information to the scene) or correcting cognitive divergences (by taking steps to repair its mental model). The *commitments component* encapsulates any obligations (commitments) the agent might have towards others. Finally, the *reasoning and decision making component* provides the process for reasoning about fuzzy knowledge and dynamically updates beliefs, checks for any inconsistencies that exist (i.e. a divergence), and allows the agent to decide on its next action(s). Consequently, the mental model may turn to one of the previously described repair strategies to resolve an observed divergence in order to "get on the same page" (i.e. reach cognitive convergence).

We utilize a hybrid model to describe the components of a computational shared mental model which incorporates fuzzy logic that allows reasoning about degrees of truth rather than exact knowledge. This logical representation has mapped well to Lakoff's prototype theory [10], which describes a view on how humans represent concepts in the world, and the way we have seen improvisers describe their own characters in a scene [12]. We represent knowledge about the story domain (e.g. the association between characters and joint activities) as degrees of association (DoA), which are essentially bi-directional fuzzy memberships (to know more about how DoA is utilized in this work, please see [13]). By representing domain knowledge this way, agents can compute the *iconicity* of memberships. Agents do not always want to present an element (e.g. a character) that has the highest DoA to the other elements in its mental model as there might also be iconic elements that hold low DoA values (e.g. a *monk* has a higly iconic low DoA with *talking*). Rather, the agent should present something that is iconic that conveys a lot of information about the agent's mental model. When an agent makes an iconic presentation, another improviser is less likely to misinterpret the presentation, which reduces cognitive divergences and aids in repairing the shared mental model. For example, only a few characters in the Old West, like an outlaw, are highly associated with robbing a bank. Therefore, robbing a bank is an iconic activity for an outlaw.

Iconicity calculations are similar to those used in [9] for *Party Quirks*, but these are updated to account for the different element categories of the knowledge structure: actions, characters, and joint activities. An agent can use the iconicities of the elements in its mental model to determine the probability that a certain element will be selected from its knowledge structure category given its current mental model. The necessity of probabilistic selection is raised to model the unpredictable nature of human behavior in interpreting context in a scene. Moreover, probability helps the agent to estimate the level of confidence it might have in its mental model.

4 Application of Fuzzy Rules

Improv actors execute motions on stage that are ambiguous. Even a simple motion like waving a hand can be interpreted multiple ways (e.g. waving to someone, erasing a board, or cleaning a window). Improv agents need to entail knowledge and future actions after observing these ambiguous presentations from a human (via Kinect) or other AI agent. To handle this ambiguity, agents need to measure their confidence in their entailed knowledge because it may be based on assumptions that diverge from others' mental models.

Fuzzy rules (a procedural representation in the form of "*if... then...*" rules for handling the ambiguous kinds of knowledge we have seen in our observations of improvisers) are designed based on the way humans reason about the likeliness of a (random) event to occur that is: the higher the probability of an event, the more certain we are that the event will occur. However, even if this is the case other elements should be considered in order to come up with a confidence value that really reflects the whole situation, such as iconicity in this work. For example, the agent would use the following fuzzy rules to determine its confidence in the character it thinks another agent is portraying:

Fuzzy Rule One: *If* agent_B thinks that action_1 done by agent_A has medium probability to occur and has low iconicity to character_X portrayed by agent_B, *then* agent_B's confidence in agent_A portraying character_X is low.
Fuzzy Rule Two: *If* agent_B thinks that action_2 done by agent_A has high probability to occur and medium iconicity to Character_Z portrayed by agent_B, *then* agent_B's confidence in agent_A portraying character_Z is medium.

For this purpose, we use Trapezoidal and Triangular Membership Functions that use three fuzzy values (low, medium, high), which provides high quality results when compared to other membership functions, see Fig. 1. The x axis represents the inputs of the probabilistic or iconicity values φ. The y axis represents the degree of membership μ of element φ in the fuzzy sets *low*, *medium*, and *high*, where each term in μ(probability) is characterized by a fuzzy set in a universe of discourse U=[0, 1].

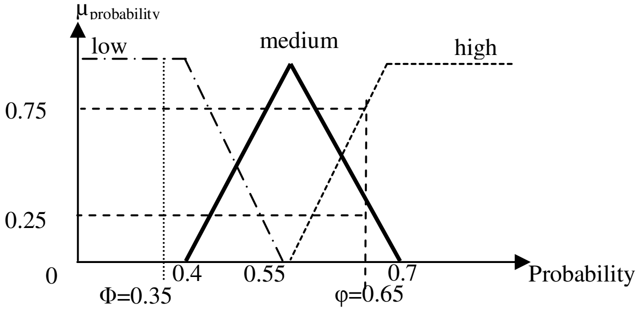


Fig. 1. Diagrammatic representation of fuzzy probabilities using Trapezoidal and Triangular Membership Functions. The low, medium, and high fuzzy values are used to determine the agent's confidence based on the probability and iconicity inputs.

In order to illustrate the computation of the confidence factor using fuzzy rules 1 and 2, consider a scene set in the Old West where agent_B believes that agent_A is presenting the joint activity *apprehending a criminal*. Agent_B wants to extrapolate this knowledge to learn what character agent_A might be portraying. Agent_B considers that agent_A might be portraying the character *sheriff* or the character *outlaw*. Assume *<apprehending a criminal, sheriff>* has *medium iconicity* = 0.4 *and <apprehending a*

criminal, outlaw> has *low iconicity* = 0.22. Based on the iconicities of those characters with the joint activity *apprehending a criminal*, the probabilistic values for agent$_A$ portraying a *sheriff and an outlaw are* φ=0.65 and φ=0.35 consecutively. These values will act as the inputs to the Trapezoidal and Triangular Membership Functions to compute their membership to the fuzzy sets: low, medium, and high. It is worth noting that we are using the same membership function shown in Fig. 1 for both iconcity and probability fuzzy variables. In order to compute the certainty factors for the portrayed characters, the agent need to apply the following three steps for Fuzzy Rule One and Fuzzy Rule Two:

Step1: Fuzzify inputs: Resolve all fuzzy statements in the antecedent to a degree of membership between 0 and 1.

Fig. 1 shows that the probabilistic value 0.65 cuts the medium and high fuzzy sets (y axis) in 0.25 and 0.75 respectively. This means that the probability of being a *sheriff* character has the following degrees of membership: $\mu_{\text{low}}(0.65)=0$, $\mu_{\text{medium}}(0.65)=0.25$ and $\mu_{\text{high}}(0.65)=0.75$. Similarly, the *outlaw* character has the following degrees of membership: $\mu_{\text{low}}(0.35)=1$ and $\mu_{\text{medium}}(0.35)=0$, $\mu_{\text{high}}(0.35)=0$.

Repeat Step 1 for the iconicity values 0.4 and 0.22, which provides the following results: $\mu_{\text{low}}(0.4)=1$, $\mu_{\text{medium}}(0.4)=0$, $\mu_{\text{high}}(0.4)=0$ for the *sheriff* character, and $\mu_{\text{low}}(0.22)=1$, $\mu_{\text{medium}}(0.22)=0$ and $\mu_{\text{high}}(0.22)=0$ for the *outlaw* character.

Step 2: Apply fuzzy operators to multiple part antecedents: The "fuzzy and" operator is the minimum of the degree of memberships in the antecedents, while the "fuzzy or" operator is the maximum of the degree of memberships in the antecedents.

Applying this step on the antecedent part of Fuzzy Rule One, we will find that the *sheriff* character has $\mu_{\text{low}}(0.4)=1$ for iconicity and $\mu_{\text{medium}}(0.65)=0.25$ for probability. Next, apply the "fuzzy and" operator to these results; $CF_{\text{rule1}}= \min\{1, 0.25\} = 0.25$.

Repeat Step 2 for Fuzzy Rule Two. We will find that the *sheriff* character has $\mu_{\text{medium}}(0.4)=0$ for iconicity and $\mu_{\text{high}}(0.65)=0.75$ for probability. Again, apply the "fuzzy and" operator to these results; $CF_{\text{rule2}}= \min\{0, 0.75\} = 0$.

Step 3: Defuzzify outputs: For a group of rules, defuzzify the outputs by aggregating all the rules' outputs to produce one 'crisp' value using the Centroid Defuzzification Method. Final crisp value for a group of rules = $\sum_{i=1}^{n} m_i w_i \ / \ \sum_{i=1}^{n} m_i$, where m_i is the membership of the output of each rule, and w_i is the centre of gravity of each fuzzy value area. Applying the Centroid Defuzzification Method to the values obtained from steps 1 and 2, we obtain: $CF_{(\text{rule1and rule2})}= (0.25*0.4+0*0.55) / (0.25+0) = 0.4$

Now the agent can use this confidence factor in the representation of knowledge in his mental model as shown in the following form of logical predicate:

bel(agent$_B$, is_a(agent$_A$, sheriff), medium, 0.4)

In plain English, this predicate can be read as: "agent$_B$ believes that agent$_A$ is a *sheriff* with medium confidence 0.4" It is worth of noting that in the transformation process of probability to confidence, iconicity acts as an adjustment factor. The same procedure would be followed to represent a shared belief about the outlaw character as shown below:

mutual_belief (agent$_B$, bel(agent$_A$, is_a(agent$_B$, cowboy), high, 0.7))

In plain English, this predicate can be read as: "Agent$_B$ has a shared belief that agent$_A$ believes that agent$_B$ is a *cowboy* with high confidence 0.7" Fuzzy rules are also needed to update the agent's confidence about his beliefs after each interaction with the other agents. An example of these fuzzy rules is shown below:

Fuzzy Rule Three: *If* agent$_B$ has low confidence about agent$_A$'s motion *and* has medium confidence about agent$_A$ as character$_Y$, *Then* decrease agent$_B$'s confidence in character$_Y$.
Fuzzy Rule Four: *If* agent$_B$ has high confidence about agent$_A$'s motion *and* has medium confidence about agent$_A$ as character$_X$, *Then* increase agent B confidence in character$_X$.

The computational implantation of this rule is achieved via using the *serial combination function (*SCF= CF_1*CF_2*) and* the *parallel combination function (*PCL= $CF_1+CF_2-(CF_1*CF_2)$*)* for rule 3 and rule 4 respectively, where CF_1 and CF_2 are the certainty factors for the first and second statement in the antecedents part of the rule.

The agents' confidence about the knowledge in their mental models keeps changing based on the actions they take. Reaching cognitive consensus requires the agents to understand each other. In fact, shared mental models can be measured in terms of the degree of overlap or consistency among team members' knowledge and beliefs [11].

5 Discussion

This paper presents a computational shared mental model for improv agents based on preliminary modeling efforts and studies of human improvisers. The goal of our model is to a) formally represent of our findings of how human improvisers negotiate shared mental models and b) support intelligent agents's ability to improvise scenes with each other or with a human interactor. This model provides the flexibility for improv agents to infer and extrapolate to new knowledge from their interaction based on their current shared mental models. Improv agents can be employed in games environments where they can reason about fuzzy uncertain knowledge and interact with human without relying privileged knowledge or communication. The fuzzy rules can be applied to other domains and can also be edited to include any number of factors (evidences) that might affect the generated confidence. In its current state, this approach does not capture all of the complexities of a full theory of mind because assessments are only based on degrees of association. For example, agents with the kind of shared mental models described here cannot reason about privileged knowledge (e.g. knowledge that only one agent knows because the other agent was not present to hear it), which a full theory of mind can account for.

References

1. Hayes-Roth, B., Van Gent, R.: Story-Making with Improvisational Puppets and Actors. Technical Report KSL-96-09. Stanford University, Palo Alto, CA (1996)
2. Swartjes, I., Vromen, J.: Emergent Story Generation: Lessons from Improvisational Theater. In: Proc. of AAAI Fall Symposium on Intelligent Narrative Technologies, Arlington, VA (2007)

3. Zhu, J., Ingraham, K., Moshell, J.M.: Back-Leading through Character Status in Interactive Storytelling. In: André, E. (ed.) ICIDS 2011. LNCS, vol. 7069, pp. 31–36. Springer, Heidelberg (2011)
4. Fuller, D., Magerko, B.: Shared Mental Models in Improvisational Theatre. In: Proc. of the 8th ACM Conference on Creativity and Cognition, pp. 269–278. ACM Press (2011)
5. Magerko, B., Dohogne, P., Fuller, D.: Shared Mental Models in Improvisational Digital Characters. In: Proc. of the 2nd International Conference on Computational Creativity, pp. 33–35 (2011)
6. Si, M., Marsella, S.C., Pynadath, D.V.: THESPIAN: An architecture for interactive pedagogical drama. In: Proc. of the Twelfth International Conference on Artificial Intelligence in Education, pp. 595–602. IOS Press (2005)
7. Van De Kieft, I., Jonker, C., Van Riemsdijk, M.B.: Improving User and Decision Support System Teamwork: An Approach Based on Shared Mental Models. In: Proc. of the 22nd IJCAI Workshop on Explanation-Aware Computing, pp. 61–70 (2011)
8. Brisson, A., Magerko, B., Paiva, A.: A Computational Model for Finding the Tilt in an Improvised Scene. In: Proc. of the 4th International Conference on Interactive Digital Storytelling, Vancouver, Canada (2011)
9. Magerko, B., DeLeon, C., Dohogne, P.: Digital Improvisational Theatre: *Party Quirks*. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 42–47. Springer, Heidelberg (2011)
10. Lakoff, G.: Women, Fire, and Dangerous Things. Univ. of Chicago Press, Chicago (1987)
11. Cannon-Bowers, J.A., Salas, E., Converse, S.A.: Shared mental models in expert team decision making. In: Castellan Jr., N.J. (ed.) Current Issues in Individual and Group Decision Making, pp. 221–246. Lawrence, Hillsdale (1993)
12. Magerko, B., Manzoul, W., Riedl, M., Baumer, A., Fuller, D., Luther, K., Pearce, C.: An Empirical Study of Cognition and Theatrical Improvisation. In: Proc. of ACM Conference on Creativity and Cognition, Berekely, CA (2009)
13. O'Neill, B., Piplica, A., Fuller, D., Magerko, B.: A Knowledge-Based Framework for the Collaborative Improvisation of Scene Introductions. In: Proc. of the 4th International Conference on Interactive Digital Storytelling, Vancouver, Canada (2011)

A Reasoning Module to Select ECA's Communicative Intention

Jeremy Riviere, Carole Adam, and Sylvie Pesty

Grenoble University - LIG, Grenoble, France
{Jeremy.Riviere,Carole.Adam,Sylvie.Pesty}@imag.fr

Abstract. In the context of a ECA-human interaction, we have created a BDI-like reasoning engine based on the agent's mental states. This reasoning engine first aims to trigger the agent's emotions from its goals, beliefs, ideals and the notion of responsibility. Then this engine selects the agent's communicative intention from its mental states and a set of dialogue rules. The integration of "Stimulus Evaluation Checks" from Scherer's appraisal theory allows us to associate the selected communicative intention with a multimodal expression. We present a test-scenario involving an argument between the user and the ECA MARC, currently used to evaluate the perceived sincerity and believability of the ECA behaviour.

Keywords: ECA, dialogue, BDI, emotion.

1 Introduction

In the field of Embodied Conversational Agents (ECA), one of the main bottlenecks can be translated into a very simple question: *how to decide what to express and how to express it?* A lot of researchers have proposed ECA's models and implementations to trigger emotions *via* appraisal and express them through multimodalities, regardless of language [1,2]. Other researchers have chosen to achieve the ECA's communicative intention through cognitive and reactive backchannels [3], while a few researchers combine emotion and language to achieve the ECA's intention [4] but not in the context of a ECA-human dialogue. In this context, what is needed is a module that lets the agent reason on its mental states, select its communicative intention while taking its emotions into account and compute a way to achieve it (both **verbally** - using a generic language - **and non-verbally**).

In this paper, we present a BDI-like (Belief, Desire, Intention - [5]) reasoning module that selects the communicative intention of an ECA from its mental states. This reasoning module corresponds to the first module of the standard SAIBA architecture [6], the communicative intention planner. This paper is structured as follows. In a first part, we describe the underlying mechanisms of the reasoning module, its architecture and the three different ways to select the agent's communicative intention. In a second part, we present our implementation of the reasoning module in Prolog and its deployment in the MARC ECA

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 447–454, 2012.

[2]. We also present a test-scenario involving an argument between the user and the ECA, and we conclude and outline the perspectives of our work.

2 Proposition: The Reasoning Module

Our reasoning module (see Figure 1) is composed of three sub-modules following the classical perception - decision - action loop of BDI architectures, plus an additional appraisal sub-module to trigger *complex* emotions and compute a facial expression. The four sub-modules are:

1. the *perception module*: updating the agent's mental states;
2. the *appraisal module*: from agent's mental states, triggering agent's emotions and evaluating Scherer's SEC to compute facial expression;
3. the *deliberation module*: selecting the agent's communicative intention;
4. the *intention planning module*: computing and executing a plan to reach this intention.

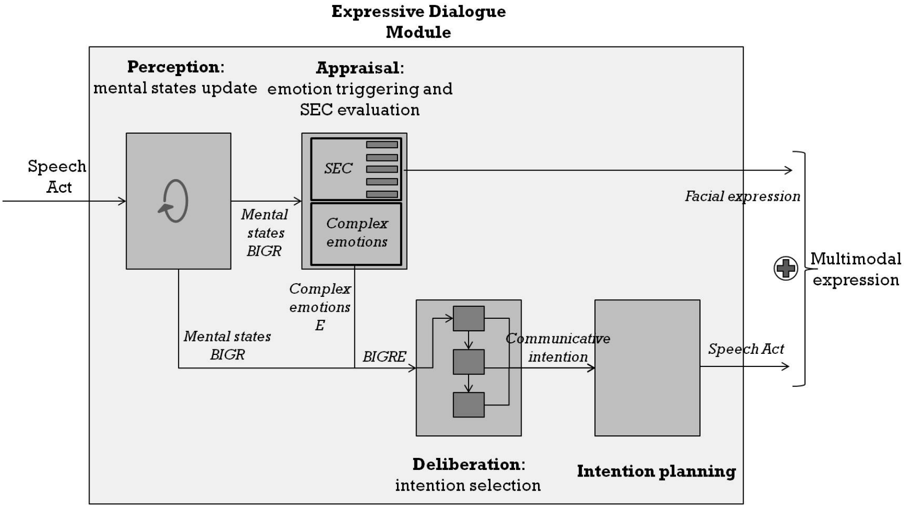


Fig. 1. General architecture of the reasoning module

We use here a specific BDI logical framework presented in [7]: *beliefs* (B), *ideals* (I), *goals* (G), *responsibility* (R) and *complex emotions* (E), forming what we call the BIGRE model. These modal operators allow us to represent utterances in terms of the mental states that they express, particularly *complex* emotions [7,8]. *Complex* emotions are based upon the agent counterfactual reasoning and upon reasoning about responsibility, skills and social norms. The notion of responsibility is central here in order to provide an analysis of counterfactual emotions such as regret and disappointment.

2.1 Perception Sub-module

During an agent-human dialogue, human utterances are translated into speech acts[1] whose perception triggers an update of the agent's mental states in the knowledge base (KB) *via* inference rules. The agent's KB is divided into static domain knowledge (library of domain plans, ontology and specific knowledge, *e.g.* about politics) and dynamic knowledge (the agent's mental states and its beliefs on the user's ones, deduced from interactions with them). The agent also updates its KB after itself performing a speech act.

2.2 Appraisal Sub-module

The appraisal sub-module aims at, first, triggering the *complex* emotions from their logical definition in terms of the agent's BIGR mental states, and second, computing a dynamic facial expression to accompagny the speech act achieving the communicative intention selected by the reasoning module. Its inputs are the agent's BIGR mental states including the belief about the user's speech act.

So on one hand, the agent's *complex* emotions E are triggered from the BIGR. For example, the emotion of gratitude is triggered when the agent i's BIGR contain $Goal_i\varphi \wedge Bel_i Resp_j\varphi$ [7]; its intensity is derived from the priority of the goal. This triggering depends on the degree of one aspect of the agent's personality, its emotionalism[2]. On the other hand, this sub-module appraises each speech act *w.r.t.* 5 "Stimulus Evaluation Checks" (SEC) introduced by Scherer's appraisal theory [9,10] and that we adapted to the speech act theory:

1. novelty of the speech act (was it expected in the dialogue scheme?);
2. intrinsic pleasantness (depending on the type of act, *e.g.* `Refuse` *vs.* `Accept`, and the propositional content);
3. congruence with the agent's goals and attribution of responsibility;
4. coping potential (can the agent influence the speech act's consequences?);
5. compatibility with the agent's ideals.

This appraisal process consists of this sequence of SEC, evaluated in turn. A dynamic facial expression can be computed from this appraisal, since Scherer showed that each SEC can be linked to a temporal sequence of Action Units (AU) [11]. So for each assessed event, the facial expression of each evaluated SEC is computed, combined with the previous one and expressed by the ECA; the global expression lasts while the ECA answers to the user (for a concrete example, see Figure 2).

Within Scherer's theory, we can identify some SEC that are part of our *complex* emotions definition: the goal congruence is represented by the *goal* operator, the attribution of responsibility by the *responsibility* operator, and norms can be assimilated with the *ideals*. Thus we can say that *complex* emotions form a subset of the emotions triggered by the SEC evaluation process.

[1] For now we only have a very limited grammar of human inputs that was designed *ad hoc* for the demo scenario.

[2] How much it is easy (or difficult) for the agent to feel an emotion,

2.3 Deliberation Sub-module

The communicative intention of the ECA is selected from its mental states (BIGR+E) *via* practical reasoning. We define three kinds of communicative intentions: the "emotional" and "obligation-based" intentions useful to local dialogue regulation [12], and the "global" intention.

The "emotional" intention is the intention to express an emotion. During each dialogue turn, the agent's emotions are triggered from its B, I, G, and R mental states as updated by the perception sub-module. Then the "emotional" intentions are selected from the triggered emotions depending on the agent's expressiveness (a very expressive agent will intend to express all its emotions while a less expressive agent will only intend to express the most intense ones). The achievement of an "emotional" intention is possible *via* the appropriate expressive speech act (the speech act library is described in Section 2.4). These intentions have *the highest priority*: an agent will first try to achieve its stronger "emotional" intention (*i.e.* express its most intense emotion) before considering other intentions. They participate in the local regulation of dialogue by enabling a more natural interaction [13] between the ECA and the human user.

The "Obligation-Based" Intention. The major drawback of most dialogue engines is that they are unable to regulate dialogue at a "local" level because they favour the agent's global intention. For example, an agent which has the intention to know something will keep asking the same question while ignoring user's input that does not answer it, including possible user's questions. Traum and Allen [16] have introduced the concept of discourse obligation to compensate this drawback: each act sent or received by the agent corresponds with a number of discourse obligations the agent has to follow. The obligations thus represent social norms guiding the agent's behaviour and making it reactive at the discourse level. We have integrated in our module a number of obligation rules that allow the regulation of the dialogue at the "local" level. For example, when the user `Asks` for something (or `Offers` something), the agent has to either `Accept` or `Refuse`, depending on its goals and plans. The second type of communicative intentions are these "obligation-based" intentions that are selected from the obligation rules. They have *priority over* the "global" intentions, but are only selected if the agent has no "emotional" intention.

The "global" intention corresponds to the global level of dialogue: it is the intention that gives the direction of dialogue and defines its type ([14], *e.g.* deliberation, persuasion). The agent adopts the global intention to pursue a certain goal when it is committed to achieve this goal. Such a commitment can be *public*, *via* commissive acts such as `Promise`, `Accept` etc., or *private*, as the agent can commit on one of its goals after practical reasoning on its knowledge *w.r.t.* the domain plans it knows. These types of commitment are consistent with Cohen and Levesque's definition [15]: our public commitment matches their social commitment and our private commitment matches their internal commitment. An exemple of *private* commitment can be that the agent has the goal to make up with the user and knows the plan to do it: in the appropriate context (*e.g.* if the agent does not have other incompatible "global" intentions and commitment), it adopts this intention.

2.4 Intention Planning Sub-module

The way to achieve the selected communicative intention is planned by the intention planning sub-module according to a plan-based approach of dialogue [17,18]. We provide in our reasoning module a library of Multimodal Conversation Acts (MCA) based on the Speech Acts theory [19,20]. These MCA are the plans' operators (actions): they are described in terms of preconditions and effects to enable backward-chaining planning. The agent looks for all the sub-actions which satisfy a given intention; then looks for all the false preconditions of these sub-actions; actions establishing these false preconditions are added to the plan; and so on.

Our library of MCA constitutes an extension of our previous work on Expressive MCA [8], since we provide here 38 MCA from four categories, namely assertives (`Affirm`, `Deny`...), directives (`Ask`, `Suggest`...), commissives (`Promise`, `Accept`...) and expressives (`Apologize`, `Satisfy`...). We identify different effects upon receiving or upon sending each MCA. For example, the definition of the `Rejoice` MCA from the point of view of the agent a in a dialogue with the human h is:

To `Rejoice`:

Preconditions: $Goal_a\varphi \wedge Bel_a Resp_a\varphi \overset{déf}{=} Rejoicing_a\varphi$
The agent a "feels" the emotion of rejoicing; it is responsible for having achieved its goal φ.

Effects upon sending: $Bel_a Bel_h Rejoicing_a\varphi$
The agent a believes that human h believes that a feels rejoicing about φ.

Effects upon receiving: $Bel_a Goal_h\varphi \wedge Bel_a Bel_h Resp_h\varphi$
The human h has just expressed rejoicing to agent a about φ, so a believes that h has achieved its goal φ and that h feels responsible for this.

When the agent receives a MCA performed by the human, it deduces from this MCA the human's mental states (including emotions in the case of Expressive MCA). This kind of "reverse appraisal" has been discussed in psychology [21] and tackled in the virtual agent field [22,23].

In the case of "emotional" and "obligation-based" intentions, the built plan usually includes only one MCA. For example, if the agent's "emotional" intention is to express gratitude, the plan will include the MCA `Thank` or `Congratulate` depending on the emotion's intensity. In the case of "global" intentions, the domain plans in the agent's KB may be necessary. These domain-dependent actions are also described in terms of preconditions and effects. For instance, if the agent intends to book a train for the user, it has to know that to book a train it needs information about the time and date of departure and destination. Then it can use the same planning mechanisms to decide what are the appropriates MCA (e.g. `Ask`) to get this information.

If a plan is already known (because it was computed earlier), this plan is updated from the last performed action and the actions which cannot be done (for example, if the user refused to give an answer, the agent has to find another plan to get the desired information).

3 Implementation in MARC

MARC [2] is an ECA following the SAIBA framework and able to communicate *via* several modalities (including facial expressions incoded by Action Units). We have implemented the reasoning module in Prolog and deployed it in MARC in order to evaluate it within a test-scenario. This scenario is within the context of a dialogue between a companion ECA (acted out by MARC) and the user. The scene takes place after an argument between MARC and the user. MARC first `Regrets` the fight (the plan to achieve its "emotional" intention) and `Asks` for forgiveness (start of the plan to achieve its "global" intention). The user then has three choices: he can `Accept` to forgive MARC ("forget about it"), `Refuse` ("you cannot be my personal agent anymore!", see Figure 2), or `Ask` for some time to think about it.

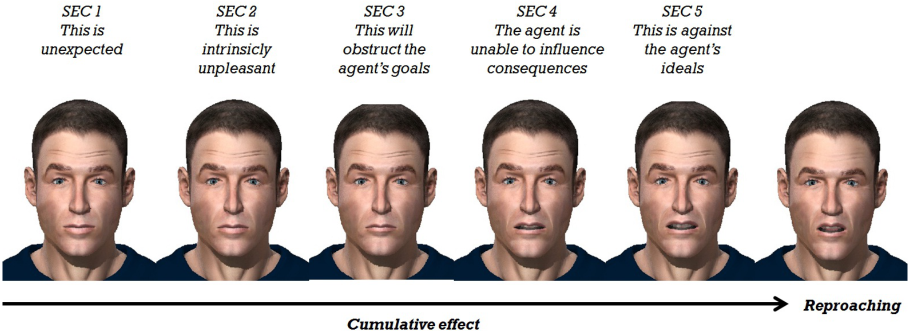

SEC 1
This is unexpected
SEC 2
This is intrinsicly unpleasant
SEC 3
This will obstruct the agent's goals
SEC 4
The agent is unable to influence consequences
SEC 5
This is against the agent's ideals
Reproaching
Cumulative effect

Fig. 2. The user just `Refused` to forgive MARC. From left to right, the expression of the evaluation of each SEC and its cumulative effect. The total expression lasts while MARC `Reproaches` the user for his refusal.

The point here is to observe the agent's reaction following the user's answer. In the case where the user `Asks` for some time to think about it, four communicative intentions are selected in sequence:

1. "emotional" intention: MARC expresses its disappointment (`Complain`) because the user does not want to grant its request;

2. "obligation" intention: MARC has to answer the user's request; since its goal is to have an answer to its own request, MARC `Refuses` to wait.
3. "emotional" intention: MARC `Apologises` because it believes that the user expected an acceptance and it is responsible for not having accepted his request.
4. "global" intention: MARC keeps following its current "global" intention to make up with the user and `Asks` again for forgiveness.

This scenario is currently used to evaluate the sincerity and believability of the agent behaviour perceived by users. The first evaluation's results are promising.

4 Conclusions and Perspectives

In this paper we have introduced a BDI-like reasoning module that aims to select and plan the ECA's communicative intention. The intention is selected from the agent's emotions (triggered from its beliefs, goals, ideals and the notion of responsibility - the BIGRE model), the user's speech acts and our set of dialogue rules. A plan consisting of a sequence of MCA is then computed to achieve this intention. To compute a multimodal expression to accompagny the MCA, we have implemented 5 of Scherer's SEC linked to facial expressions. We have implementated this reasoning module in Prolog and integrated in the MARC ECA. A user evaluation using MARC is currently in progress to prove that both the communicative intention selected by our reasoning module and the multimodal expression computed from events appraisal have a positive influence on the perceived believability and sincerity of ECA.

Acknowledgments. This work is supported by the French National Research Agency (ANR), CECIL project (*www.irit.fr/CECIL*), grant number ANR-08-CORD-005.

References

1. Marsella, S.C., Gratch, J.: EMA: A process model of appraisal dynamics. Cognitive Systems Research 10, 70–90 (2009)
2. Courgeon, M., Martin, J.-C., Jacquemin, C.: MARC: a Multimodal Affective and Reactive Character. In: 1st Workshop on Affective Interaction in Natural Environments (AFFINE 2008), Chania, Crete (2008)
3. 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)
4. Aylett, R., Dias, J., Paiva, A.: An affectively driven planner for synthetic characters. In: International Conference on Automated Planning and Scheduling (ICAPS), pp. 2–10. AAAI Press (2006)
5. Rao, A.S., Georgeff, M.P.: Modeling Rational Agents within a BDI-Architecture. In: 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR 1991), pp. 473–484. Morgan Kaufmann publishers Inc, San Mateo (1991)

6. Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S.C., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: 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)
7. Guiraud, N., Longin, D., Lorini, E., Pesty, S., Riviere, J.: The face of emotions: a logical formalization of expressive speech acts. In: 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011), pp. 1031–1038. International Foundation for AAMAS (2011)
8. Riviere, J., Adam, C., Pesty, S., Pelachaud, C., Guiraud, N., Longin, D., Lorini, E.: Expressive Multimodal Conversational Acts for SAIBA Agents. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 316–323. Springer, Heidelberg (2011)
9. Scherer, K.R.: Emotion. In: Introduction to Social Psychology, pp. 151–195. Wiley-Blackwell (2001)
10. Scherer, K.R.: Appraisal considered as a process of multi-level sequential checking. In: Appraisal processes in emotion: Theory, Methods, Research, pp. 92–120. Oxford University Press, New York and Oxford (2001)
11. Scherer, K.R., Ellring, H.: Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? Emotion 7(1), 113–130 (2007)
12. Baker, M.J.: A Model for Negotiation in Teaching-Learning Dialogues. Journal of Artificial Intelligence in Education 5(2), 199–254 (1994)
13. Bates, J.: The role of emotion in believable agents. Communication of ACM 37(7), 122–125 (1994)
14. Walton, D., Krabbe, E.: Commitment in Dialogue: Basic concept of interpersonal reasoning. State University of New York Press, Albany (1995)
15. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence 42(2-3), 213–261 (1990)
16. Traum, D.R., Allen, J.F.: Discourse Obligations in Dialogue Processing. In: 32nd Annual Meeting on Association for Computational Linguistics (ACL 1994), pp. 1–8. Association for Computational Linguistics, Stroudsburg (1994)
17. Perrault, C.R., Allen, J.F.: A Plan-based Analysis of Indirect Speech Acts. Computational Linguistics 6, 167–182 (1980)
18. Allen, J.F., Perrault, C.R.: Analyzing Intention in Utterances. Artificial Intelligence 15(3), 143–178 (1980)
19. Searle, J.R.: Speech acts: an essay in the philosophy of language. Cambridge University Press, New York (1969)
20. Searle, J.R., Vanderveken, D.: Foundations of illocutionary logic. Cambridge University Press, New York (1985)
21. Hareli, S., Hess, U.: What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. Cognition and Emotion 24(1), 128–140 (2009)
22. de Melo, C.M., Zheng, L., Gratch, J.: Expression of Moral Emotions in Cooperating Agents. In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) IVA 2009. LNCS, vol. 5773, pp. 301–307. Springer, Heidelberg (2009)
23. de Melo, C., Gratch, J., Carnevale, P.: Reverse Appraisal: Inferring from Emotion Displays who is the Cooperator and the Competitor in a Social Dilemma. In: The 33rd Annual Meeting of the Cognitive Science Society (CogSci 2011), Boston (2011)

Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors

Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini,
Alesia Egan, Albert (Skip) Rizzo, and Louis-Philippe Morency

University of Southern California
Institute for Creative Technologies, Los Angeles, California
scherer@ict.usc.edu

Abstract. Modern virtual agents require knowledge about their environment, the interaction itself, and their interlocutors' behavior in order to be able to show appropriate nonverbal behavior as well as to adapt dialog policies accordingly. Recent achievements in the area of automatic behavior recognition and understanding can provide information about the interactants' multimodal nonverbal behavior and subsequently their affective states. In this paper, we introduce a perception markup language (PML) which is a first step towards a standardized representation of perceived nonverbal behaviors. PML follows several design concepts, namely *compatibility and synergy*, *modeling uncertainty*, *multiple interpretative layers*, and *extensibility*, in order to maximize its usefulness for the research community. We show how we can successfully integrate PML in a fully automated virtual agent system for healthcare applications.

Keywords: perception, standardization, multimodal behavior analysis, virtual human system.

1 Introduction

Human face-to-face communication is a complex bi-directional multimodal phenomenon in which interlocutors continuously emit, perceive and interpret the other person's verbal and nonverbal displays and signals [9,5]. Interpreting a person's behavior to understand his or her intent requires the perception and integration of a multitude of behavioral cues, comprising spoken words, subtle prosodic changes and simultaneous gestures [13].

Many recent achievements in automatic behavior analysis enable automatic detection, recogniton or prediction of nonverbal behavioral cues, such as laughter [15], voice quality [14], backchannels [8], or gestures [17]. For the rapid advancement of virtual agent systems it is crucial to establish an infrastructure that allows researchers to efficiently integrate these sensing technologies and share their new developments with other researchers. In this paper we introduce perception markup language (PML) as a first step towards a standard representation of perceived nonverbal behavior. The standardization of PML was inspired by efforts

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 455–463, 2012.

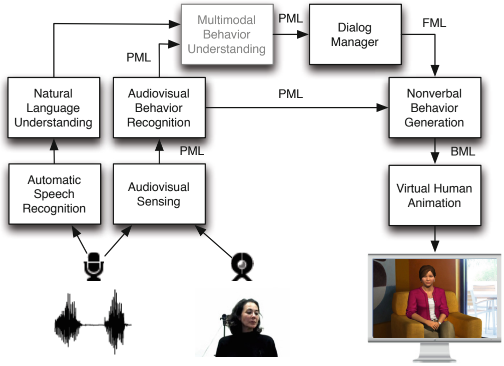

Multimodal Behavior Understanding
PML
Dialog Manager
FML
PML
Natural Language Understanding
Audiovisual Behavior Recognition
PML
Nonverbal Behavior Generation
PML
BML
Automatic Speech Recognition
Audiovisual Sensing
Virtual Human Animation

Fig. 1. Schematic overview showing how perception markup language (PML) can be used in a virtual human architecture

in the field of nonverbal behavior generation where behavior markup language (BML) and functional markup language (FML) have been introduced in order to enable standardized interfaces for virtual human behavior animation [6,4].

We show, in Section 3, how PML can interface between sensing and other modules (see Figure 1). PML enables interactive virtual humans to react to the user's nonverbal behaviors. With PML a virtual human system can, for example, provide a wide range of verbal and nonverbal backchannel feedback such as a head nod or para-verbals (e.g., uh-oh) that signal attention, comprehension, (dis-)agreement or emotional reaction to the perceived utterance. This promotes enhanced bi-directional conversations that improve the fidelity of the interaction.

We implemented PML in a real-world healthcare application called "Ellie". Ellie is designed to help detect behaviors related to depression and post-traumatic stress disorder (PTSD) and offers related information if needed. We show in this paper how PML can successfully interface between sensing and higher-level modules such as the dialog manager (DM) and nonverbal behavior generation (NVBG).

2 Perception Markup Language

In this section we describe our perception markup language (PML), which takes full advantage of the well-established XML standard. We first express the four design concepts behind PML, then give a formal definition through two examples and finally describe PML interaction with DM and NVBG.

2.1 Design Concepts

In the following we discuss the main design concepts behind PML.

Compatibility and Synergy. Significant effort has been dedicated to building standards for virtual human animation (e.g., FML, BML) [6,4,16], speech and language representation (e.g., VoiceXML[1]) and user emotional state (e.g., EmotionML[2]). When designing PML, we carefully analyzed previous standards to learn from their experience and keep PML as compatible as possible. For example, we followed naming conventions of the BML standard whenever possible, as we envision a close interaction between the two standards. Also, instead of reimplementing a textual analysis layer to PML, we plan to work closely with existing speech standards such as VoiceXML. By following these guidelines we not only accelerate the development of a standard, but we also make PML more accessible to the community.

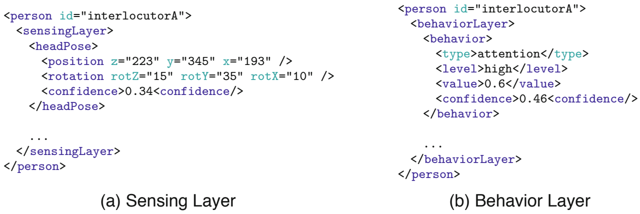
Modeling Uncertainty. One of the biggest differentiators between PML and previous standards for virtual human animation (e.g., BML, FML) is the requirement of modeling the inherent uncertainty in sensing and interpreting human nonverbal behaviors. The same visual gesture such as a gaze away can be interpreted as a thinking moment or a disengagement behavior. The language needs to be able to handle multiple hypothesis with their own uncertainty measure. Also, the audio and visual sensing is prone to noise and errors (e.g., due to occlusion or quick movement) which may result in observations with low confidence (i.e., high uncertainty). The correct handling of uncertainty within such modules not only leads to more robust predictions, but might even improve generalization capabilities.

Multiple Interpretative Layers. When building computational representations of nonverbal communication, people naturally identify multiple layers of interpretation. A concrete example of this is seen in the SAIBA framework [6] which defines separate layers for the behaviors (BML) and their dialog functions (FML). Since the BML is processed by an animation module (e.g., SmartBody) to create the final animation parameters, we can even identify a third layer which includes these animation parameters sent to the realization engine. PML follows the same logic by predefining three layers: sensing, behaviors and functions. Further, these layers allow for the versatile use of PML messages. Some components might solely be interested in higher level observations, while others might analyze rawer data.

Extensibility. Since the field of human behavior recognition and understanding is a constantly growing and developing one, we expect that the XML schema of PML will require multiple updates and revisions even after a deployable version is attained. As technologies develop, the language should develop and adapt to changing requirements. Through collaboration with researchers developing new technologies and researchers using the language, the standard elements of PML will be expanded. The schema can also be automatically converted into code usable for various programming languages following a few processing steps, rendering PML an easily maintainable and extensible markup language.

[1] http://www.w3.org/TR/voicexml30/

[2] http://www.w3.org/TR/emotionml/



```
<person id="interlocutorA">
  <sensingLayer>
    <headPose>
      <position z="223" y="345" x="193" />
      <rotation rotZ="15" rotY="35" rotX="10" />
      <confidence>0.34<confidence/>
    </headPose>

    ...
  </sensingLayer>
</person>
```

(a) Sensing Layer

```
<person id="interlocutorA">
  <behaviorLayer>
    <behavior>
      <type>attention</type>
      <level>high</level>
      <value>0.6</value>
      <confidence>0.46<confidence/>
    </behavior>

    ...
  </behaviorLayer>
</person>
```

(b) Behavior Layer

Fig. 2. PML sample sensing layer (left) and behavior layer (right)

2.2 Perception Markup Language Specification

Based on these design concepts, we developed the perception markup language (PML) which is a multi-layer representation of perceived nonverbal behaviors and their uncertainty. PML contains two main sections: `<header>` refers to the meta-data section (e.g. time stamps and information source) and `<person>` encloses the perceived nonverbal behaviors of a specific person. PML predefines three different layers: `<sensingLayer>`, `<behaviorLayer>` and `<functionLayer>`.

The `sensingLayer` layer provides information from the multiple sensing technologies about the audiovisual state of the users such as their gaze direction, their intonation or their body posture. The `behaviorLayer` layer represents the nonverbal behaviors recognized by integrating temporal information from one or more sensing cues. For example, this layer integrates head and eye gaze to estimate attention behavior or, head and arm motion to estimate fidgeting and rocking behaviors. The `functionLayer` provides information about the user's intent, functional role or affective state of the associated behavior. These higher level concepts are usually estimated by integrating verbal and nonverbal behaviors with contextual information. This paper focuses on the first two layers, keeping the `functionLayer` as future work where we plan to interface this layer with other high-level markup languages such EmotionML and FML. The remainder of this section explains the different parts of a typical PML message.

Message Header `<header>`. The header of a PML message includes meta-data such as the time stamp that is used for synchronization and a list of datasources `source` identified by a `name` and `id`. The datasource `id` is reflected in the observations within the message `<person>`.

Message Body `<person>`. The message body refers to a single person specified with a unique identifier `id`. As discussed, the information associated with each person is separated into multiple layers. The `<sensingLayer>` layer provides information about the current instantaneous audiovisual states. It includes, but is not limited to, fields such as: `gaze, headPose` and `posture`. Each item of the `<sensingLayer>` provides varying information relevant to the field, for example `gaze` provides information on the vertical and horizontal rotation of the eyes and `headPose` provides the coordinates and rotation degrees of the head in the current moment. The `confidence` represents one offered approach to model the uncertainty in sensing technologies. Other fields such as `covariance` can be used to specify the full covariance matrix. An example of the `<sensingLayer>` is seen in Figure 2 (a). The `<behaviorLayer>` layer includes information gathered over longer time periods or inferred complex behaviors. Information such as the attention behavior of the perceived person is transmitted within a `behavior` item. Again the modularity of the approach is seen in the example below, where `behavior` items are structured similarly in order to have an easily extensible approach that can grow with the development of the technology and the demands of connected components. Each `behavior` is identified with a `type`, and the values `level` (categorical; low, mid, high) and `value` (continuous; $\in [0, 1]$) indicate the behavior strength. Again, `confidence` indicates the certainty associated with the items as seen in Figure 2 (b).

2.3 PML Interaction with Other Modules

This section describes an example of how PML can interact with dialog management and the nonverbal behavior generation.

Processing PML in DM. The dialogue manager (DM) is tasked to keep track of the state of the conversation, and then decides when and what action to execute next. The DM can select actions that drive the virtual character to say a particular line or execute some specified nonverbal behavior. In this example we use an information state based DM (see [19]) designed to support flexible mixed initiative dialogues and to simplify the authoring of virtual characters. Every time an event is received, to find which action to execute in response, the DM simulates possible future conversations and then selects the action that achieves the highest expected reward. To support relatively high frequency PML events in our virtual agent system, a forward search is not initiated for each PML event that is received, but instead, the message *silently* updates the information state. This in turn affects the action selection procedures of the DM. So, if the audiovisual sensing module identifies a lack of user attention a dialog policy will be triggered to inquire about this possible lack of engagement. Two similar examples are shown in the supplemental video.

Processing PML in NVBG. Whereas, the nonverbal behavior generator (NVBG) [7] automates the generation of physical behaviors for our virtual humans, including nonverbal behaviors accompanying the virtual humans dialog, responses to perceptual events as well as listening behaviors. The handling of PML messages represents a different use case than generating nonverbal behavior for the virtual human's own utterances. In the case of the PML messages, NVBG is deciding on how to respond to perceptual signals about the human's behavior. A human's responses to others' nonverbal behavior, such as mirroring behavior and generic feedback, can in large be measured automatically as opposed to having an explicit communicative intention like an utterance.

Specifically, NVBG's response is determined by a perceptual analysis stage that leads into the behavior analysis and BML generation stages discussed previously. How the virtual human reacts towards actual PML messages in an example interaction is shown in Section 3.2.

3 Use Case: Virtual Human for Healthcare Application

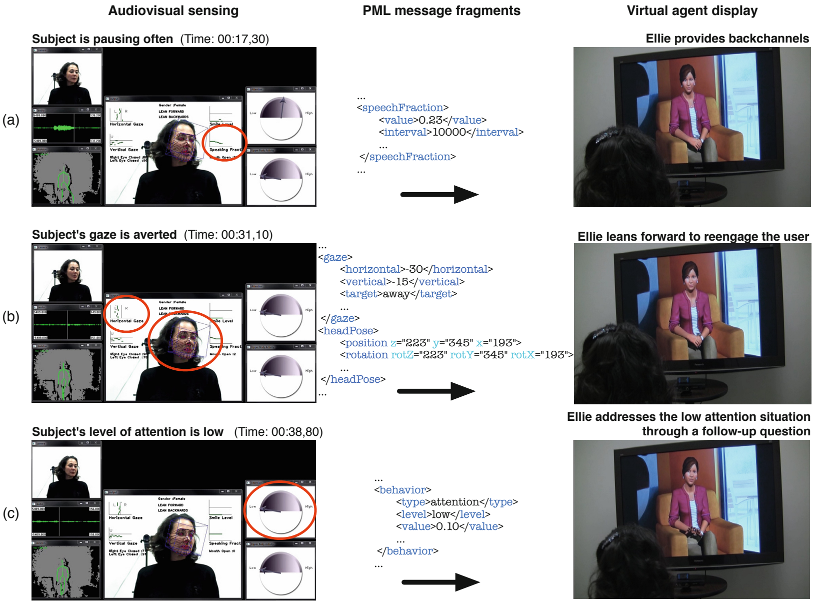
The use case scenario in this paper is aimed at subjects and patients interacting with *Ellie*, a virtual human healthcare provider. The interactive sensing system detects depression and PTSD relevant indicators and can offer a faster screening process to a population that often experiences significant wait-times before seeing a real clinician or therapist. The behavioral cues, or indicators include, but are not limited to, behaviors such as lack of expressivity in speech [10,3], constant and ongoing gaze aversion and lack of mutual gaze [21,10], as well as increased amounts of anxiety expressed by a rocking motion or fidgeting [2,11].

3.1 Implementation Details

Figure 1 shows the interactional loop of our virtual agent. The *automatic speech recognition* is performed CMU's pocket Sphinx [1] with a push-to-talk approach. The acoustic and language models were trained using transcribed interactions from our pre-study. The recognized utterance is sent to the *natural language understanding* module which recognizes speech acts such as question, statement and backchannel. For the *audiovisual sensing* component we have developed a flexible framework for sensing, based on the social signal interpretation framework (SSI) by [20]. We integrated the following sensing technologies: Cogito Health's *Cogito Social Signal Platform* (CSSP) to extract the speaking fraction as well as other audio features for the speaker, OMRON's *OKAO Vision* for the eye gaze signal, and *CLM FaceTracker* by [12] for facial tracking and head position and orientation, and Microsoft Kinect skeleton tracker.

For the *audiovisual behavior recognition* module we implemented memory-driven rule-based system which integrates audio-visual information over time to get higher-level signals such as attention (measured by the gaze signal and face orientation) and activity (measured by body pose information).

The *dialogue manager* has the task to keep track of the state of the conversation and decides when and what action to execute. We use an information state based dialogue manager (see [19]) designed to support flexible mixed initiative dialogues and simplify the authoring of virtual characters. In our scenario, the verbal and nonverbal messages are integrated directly by the dialogue manager instead of having a separate *multimodal behavior understanding* module (as originally shown in the Figure 1). We used the same technique described in Section 2.3 to automatically generate virtual human nonverbal behavior based on the generated utterances (sent by the dialogue manager through the FML messages) and the perception messages (PML). We then use the SmartBody animation module [18] to analyze the BML messages and produce the animation parameters. The final virtual human animations are created using the Unity game engine.


Audiovisual sensing
PML message fragments
Virtual agent display
Subject is pausing often (Time: 00:17,30)
(a)
Ellie provides backchannels
Subject's gaze is averted (Time: 00:31,10)
(b)
Ellie leans forward to reengage the user
Subject's level of attention is low (Time: 00:38,80)
(c)
Ellie addresses the low attention situation through a follow-up question

```
...
<speechFraction>
      <value>0.23</value>
      <interval>10000</interval>
      ...
 </speechFraction>
...
```

```
...
<gaze>
      <horizontal>-30</horizontal>
      <vertical>-15</vertical>
      <target>away</target>
      ...
 </gaze>
<headPose>
      <position z="223" y="345" x="193">
      <rotation rotZ="223" rotY="345" rotX="193">
      ...
 </headPose>
...
```

```
...
<behavior>
      <type>attention</type>
      <level>low</level>
      <value>0.10</value>
      ...
 </behavior>
...
```

Fig. 3. Display of multimodal nonverbal behavior analysis (i.e. Multisense; left column) shown with corresponding PML message fragments (middle column) and virtual agent reactions (right column). Times indicate position in supplementary video.

3.2 Example and PML Analysis

Figure 3 exemplifies a detailed analysis of a typical interaction with our virtual agent and highlights several key moments when PML messages are used. In these key moments, *Ellie* reacts towards the subject's nonverbal behavior in a way that would not have been possible without the information provided by PML. She for example, exhibits a *head nod* when the subject is pausing a lot in the conversation to encourage the subject to continue speaking (see Figure 3 (a)). In Figure 3 (b), the subject exhibits low attention by looking away. A PML message with this information is sent to NVBG and the virtual agent is signaled to *lean forward*, as an effort to engage the subject. Figure 3 (c) shows an instance where PML signals the DM that the subject's *attention level* is low. This message triggers a branching in the dialog policy[3].

4 Conclusions

We introduced the perception markup language (PML), a first step towards standardizing perceived nonverbal behaviors. Further, we discussed how the PML messages are used to either change dialog policies or the virtual agent's nonverbal behavior. We provided a detailed walkthrough of a current version of our

[3] For further information and examples refer to: http://projects.ict.usc.edu/pml

system with the help of an example interaction, which is provided in full as a supplementary video to the submission of this paper.

In the long run, PML will enable collaborations between currently often isolated workgroups as well as increase the reusability of previous findings and implementations.

Acknowledgements. This work was funded in part by the DARPA Detection and Computational Analysis of Psychological Signals (DCAPS) program and in part by the National Science Foundation grant #IIS-1018055. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices, vol. 1 (2006)
2. Fairbanks, L.A., et al.: Nonverbal interaction of patients and therapists during psychiatric interviews. J. Abnorm. Psychol. 91(2), 109–119 (1982)
3. Hall, J.A.: et al. Nonverbal behavior in clinician-patient interaction. Appl. Prev. Psychol. 4(1), 21–37 (1995)
4. Heylen, D., Kopp, S., Marsella, S.C., Pelachaud, C., Vilhjálmsson, H.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)
5. Kendon, A. (ed.): Nonverbal Communication, Interaction, and Gesture. Selections from Semiotica Series. Walter de Gruyter (1981)
6. 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)
7. 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)
8. Ozkan, D., Morency, L.-P.: Modeling wisdom of crowds using latent mixture of discriminative experts. In: Proc. of ACL HLT 2011, pp. 335–340. Association for Computational Linguistics (2011)
9. Pentland, A.: Honest Signals - How they shape our world. MIT Press (2008)
10. Perez, J.E., Riggio, R.E.: Nonverbal social skills and psychopathology. Nonverbal Behavior in Clinical Settings, pp. 17–44. Oxford University Press (2003)
11. Pestonjee, D.M., Pandey, S.C.: A preliminary study of psychological aftereffects of post-traumatic stress disorder (ptsd) caused by earthquake: the ahmedabad experience. Technical Report WP2001-04-01, Indian Institute of Management (2001)
12. Saragih, J.M., et al.: Face alignment through subspace constrained mean-shifts. In: Proc. of ICCV 2009, pp. 1034–1041. IEEE (2009)
13. Scherer, S., et al.: A generic framework for the inference of user states in human computer interaction: How patterns of low level communicational cues support complex affective states. JMUI, Special Issue on: Conceptual frameworks for Multimodal Social Signal Processing, 1–25 (2012)

14. Scherer, S., et al.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. In: CSL (2012)
15. Scherer, S., et al.: Spotting laughter in naturalistic multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. ACM TiiS: Special Issue on Affective Interaction in Natural Environments 2(1), 4:1–4:31 (2012)
16. Schröder, M.: The SEMAINE API: A component integration framework for a naturally interacting and emotionally competent Embodied Conversational Agent. PhD thesis, Saarland University (2011)
17. Suma, E.A., et al.: Faast: The flexible action and articulated skeleton toolkit. In: Virtual Reality, pp. 247–248 (2011)
18. Thiebaux, M., et al.: Smartbody: behavior realization for embodied conversational agents. In: Proc. of AAMAS 2008, pp. 151–158 (2008)
19. Traum, D.R., Larsson, S.: The information state approach to dialogue management. In: Kuppevelt, J., Smith, R.W., Ide, N. (eds.) Current and New Directions in Discourse and Dialogue, vol. 22, pp. 325–353. Springer (2003)
20. Wagner, J., Lingenfelser, F., Bee, N., André, E.: Social signal interpretation (ssi). KI - Kuenstliche Intelligenz 25, 251–256 (2011), doi:10.1007/s13218-011-0115-x
21. Waxer, P.: Nonverbal cues for depression. Journal of Abnormal Psychology 83(3), 319–322 (1974)

Flexible Conversation Management Using a BDI Agent Approach

Wilson Wong, Lawrence Cavedon, John Thangarajah, and Lin Padgham

School of Computer Science and IT, RMIT University, Melbourne, Australia
{wilson.wong,lawrence.cavedon,john.thangarajah,lin.padgham}@rmit.edu.au

Abstract. We describe a BDI (Belief, Desire, Intention) goal-oriented architecture for a conversational virtual companion embodied as a child's Toy, designed to be both entertaining and capable of carrying out collaborative tasks. We argue that the goal-oriented approach supports both structured conversational activities (e.g., story-telling, collaborative games) as well as more "free-flowing" engaging dialogue with variation and some unpredictability. BDI plans encode the knowledge required for the structured engagements, with the use of multiple plans for conversational goals providing variation in the interactions.

1 Introduction

A conversational virtual companion [1] must be able to engage the user, which in turn requires the ability to support both structured conversation-based activities (e.g., story-telling, collaborative games) as well as more "free-flowing" chatty dialogue. Unlike task-based dialogue, the purpose of what we refer to as *Conversational Activity* is not simply to successfully perform a task (e.g., book a flight), but to actually engage and entertain the user over an extended period.

We describe a BDI architecture for a conversational agent that supports both task-oriented dialogue as well as more "chatty" conversations. The BDI agent model has been used successfully in a range of applications (e.g., see [2, Ch. 10]) that require a mix of reactive behaviour and goal-directed reasoning. This design model supports different means for achieving a goal depending on context and other factors [3]. The mixed reactive/proactive model enables the management of coherent conversational activity while still being responsive to unexpected user input. Scripted BDI plans provide knowledge of how to perform different types of *Conversational Activities*. BDI agent-based approaches to dialogue management have been previously proposed (e.g., [4,5]); however, these have typically been for task-oriented conversations (e.g., accessing email or managing an appointment). A novelty of our approach is the use of the BDI framework to provide variability in the way a goal is progressed, as well as in the conversational content.

The content for our agent appears in the form of *Conversational Fragments*, which are effectively templates of adjacent pairs of utterances. These are dynamically assembled into sequences to construct conversations. However, unlike chatbots, the conversation is strongly guided by the plans of the Conversational

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 464–470, 2012.

Activity which provides the narrative framework and coherence typically missing from chatbot-generated dialogue; our approach also supports *mixed-initiative*.

2 Architectural Overview

Our conversational infrastructure is implemented in the context of an interactive Toy, designed as a virtual companion for children. The Toy contains a *Dialogue Manager* (DM) which is composed of a *Conversation Management* (CM) component that interacts with an *Activity Management* (AM) component. These are both implemented using a BDI agent-oriented methodology.

The AM selects and instantiates specific *Conversational Activities* which direct the structure and the kind of content for the Toy utterances, while the CM manages the specific details of choosing utterances and interpreting input. The CM has dialogue processing strategies built as plans. For example, there are plans designed to handle errors or low-confidence results from speech recognition; plans to handle utterance content and update the information state; and plans to manage concurrent conversational threads and select which of a number of candidate responses to output.

The CM is designed to be multi-domain and extensible via Conversational Activity modules. These modules are designed to guide conversation around particular activities within a content domain and encapsulate the plans and data required for this. A conversational activity module contains: a knowledge-base segment; a set of conversational fragments; a collection of plans to handle the particular conversational activities of the module; and, for each top level conversational activity, an input grammar which specifies the form of the input to be interpreted as a trigger for starting the activity. The input grammar is specified using regular expressions that can be matched against user input. For example, the input grammar for the story-telling activity is: "`* tell * story *`" which results in instantiation of a goal to initiate a story-telling activity.

The input handling component of the DM analyses and extracts weighted keyphrases, topics, sentiments and requests from the user inputs. We use the Stanford Parser [6] for part-of-speech tagging, Morphadorner[1] for lemmatisation, and the dictionary-based approach for detecting sentiment [7] and request. It analyses whether the input is a response that (1) matches one of a set of templates for continuation of the current conversational activity (OK), (2) is a specific request, i.e., matches an input grammar to some conversational activity (Specific-Request), (3) is an expression of negative sentiment but without any specific request (Negative), or (4) is not able to be understood (Not-Understood). The analysed input is then provided to the AM for decision as to what to do next, which can be to continue with the current activity, or to suspend/abort it, in which case either a new Conversational Activity will be instantiated, or an existing suspended one will be resumed.

[1] http://morphadorner.northwestern.edu

For generating output, the `CM` receives weighted contextual information that has been built up from both inputs and utterances during the interaction, as well as information provided by the plans within the current Conversational Activity. It uses the *Fragment Library* to find appropriate utterance templates, instantiates any variables using both contextual information and the *Knowledge Base*, and then strategically determines the response to be uttered. The *Fragment Library* contains Conversational Fragments, which are pre-scripted[2] pieces of dialogue, which may be tagged as relevant to particular goals, and may contain both input and response variables (e.g. `$FOOD`, `$ANIMAL` etc), to allow scripting of more general purpose fragments. A generic response variable `$ETC` is used when any response is considered acceptable. The use of Conversational Fragments avoids the need for full natural language generation and allows the Toy to generate quite flexible interactions by choosing amongst relevant fragments in a non-deterministic, but nevertheless guided manner.

When interacting with the child, the Toy suggests possible Conversational Activities such as a cooking game/role play, a story, a quiz, etc. These activities are represented as BDI goal-plan structures (i.e., a set of plan templates in the agent's plan library) which guide the different aspects of the activity and the selection of fragments for the Toy to utter in pursuit of that activity. Importantly, the specific utterances are **not** part of the activity structure. Rather, it includes goals and plans to move the conversation in particular directions. The plans can provide contextual information which is used by the `CM` to select appropriate outputs. Analysis of the child's input also provides data that is used to determine how to progress the activity. For example, keyphrases from the input may help to guide plan selection within a particular activity. This notion of Conversational Activity helps to keep the dialogue cohesive, while allowing flexibility. It also meets the requirement that an engaging interaction should provide interesting tasks for the child while staying controlled by them [8]. We note that activities can be resumed or paused to allow switching between them, either to follow the child's topic requests or to insert personalised contributions, for example.

3 Activity Management

Central to our architecture is the library of goal-plan structures for directing coherent interaction with the user. Our architecture assumes a BDI plan library, with plans that have a *trigger* (the goal[3] they will achieve), a *context condition*, which determines the situation under which this plan is to be used, and a *body* which contains the plan code, which we can think of as *plan-steps* [3]. Some of these steps will be subgoals, which trigger the selection of plans to achieve them. This gives rise to a goal-plan tree, where a goal can have many possible plan options to achieve it, and a plan may contain many (sub)goals.

[2] We have recently developed techniques for automatically mining content from web question-answer forums.

[3] Goals are often called and implemented as *Events* in BDI agent platforms.

As analysed in [3], this provides a very large number of possible executions within a relatively compact structure. According to the example in [3], a goal-plan tree with depth 3, 2 plans per goal, and 4 subgoals per plan can result in over 2 million execution paths. In our case, this equates to 2 million different potential conversations resulting from a single activity tree of this size.

It is this diversity of execution paths which we exploit to achieve the desired variability, while retaining coherent, goal-oriented dialogue. As variability itself is a key aim in our design, we require multiple plans that can be applicable to each sub-goal. These are then chosen somewhat randomly to avoid the child obtaining the same response to similar input utterances.

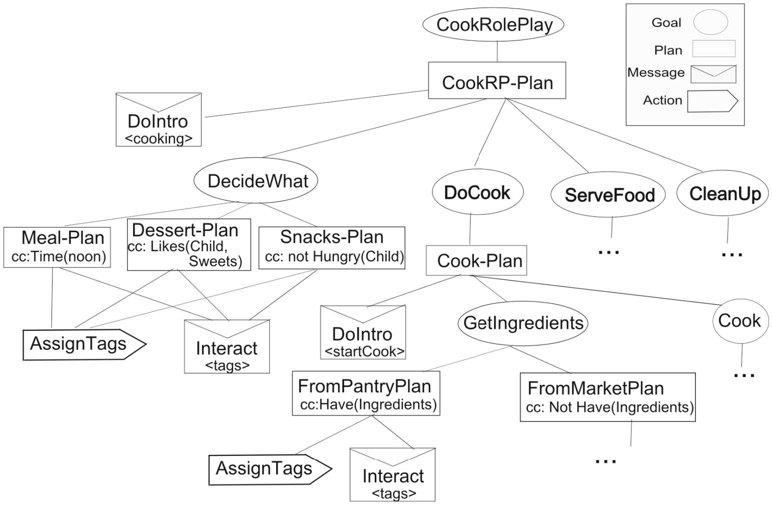


Fig. 1. Example activity: Cooking role play

Processing Conversational Goals: Figure 1 shows a (partial) example goal-plan tree for the cooking role-play activity in the Toy. The top-level goal has a single plan which guides the structure of the activity. It is possible to have different plans to choose from at the top level, providing even more variety. This plan has a sequential set of subgoals, each with a set of plans to choose from, and so on. Prior to executing the subgoals the plan first sends a DoIntro message[4] to the CM, carrying information about the current activity (Cooking), and triggers a plan in the CM to select a suitable introductory fragment for this activity. This introductory fragment (without any expected response) will be prefixed to the next system output. It will then decide what to cook using one of the plans that achieve the DecideWhat subgoal. This involves performing some interactions with the user, by assigning suitable tags (AssignTags action) that will be used

[4] The DoIntro message is essentially a subgoal that triggers another plan.

when selecting output fragments and posting an `Interact` message to the `CM` to perform the interaction. This results in the `CM` determining an output fragment and analysing the user response, which is then provided back to the plan in the form of phrases and a response category. Assuming the response is `OK`, when the plan has completed its interactions, it decides (based on the terms collected) what food it believes is going to be prepared, and the activity progresses onto the next subgoal `DoCook`, which is managed in a similar way. The `DoIntro` goal assists in smoothly moving between sub-activities.

Managing Activities: An important capability of the Toy is to be responsive to the user. The agent has to be able to drop the current Conversational Activity if this is inferred from user input. This may occur via an explicit demand (e.g., *"I want to do something else"*), detecting that the child has drifted away from responsiveness, or opportunistically (e.g., to insert a joke pertinent to the previous utterance). Our intention-based BDI approach incorporates a framework for suspending plans (and thereby activities), including infrastructure for reasoning about whether to resume or terminate a suspended activity later on. This provides a seamless approach to inserting "digressions", which are common in natural conversation flow. The BDI architecture thus provides the framework for transitioning between Conversational Activities. However, we require such transitions to be conversationally smooth and coherent. To facilitate this we use a *semantic relatedness measure* [9], and *transition fragments*. The latter are specially designed fragments to provide the conversational glue between activities (e.g. *"OK, lets finish the* `$FROMACTIVITY` *now and do something else."*).

4 Example Interaction

Figure 2 shows part of an example interaction with the Toy. The agent started with the chat activity at Line 4 and the keywords from the subsequent user inputs and system outputs were collected to form the context. We will highlight several features discussed above using the line numbers:

`Line 13`: Select the cooking roleplay based on context;
`Line 15`: `CookRP-Plan` guides conversation; `GetIngredients` subgoal invoked;
`Line 16`: Explicit request from user matches trigger for story telling activity;
`Line 17`: Cooking activity suspended; Agent chooses story using relatedness;
`Line 24`: Story ends; Cooking activity is resumed using a transition fragment;
`Line 27`: Question detected; the question-answering activity invoked;
`Line 36`: Cooking activity ends; Agent selects new Conversational Activity.

Figure 3 shows two conversations that evolve from the same starting point, and use the same Conversational Activity structure. However, the choice of different plans to achieve the subgoals, as well as different fragments to output for the `Interact` goal, results in different interactions (i.e., variability).

```
04: SYSM: do you like chocolate?
05: USER: i love chocolate.
06: SYSM: i love it too. i wish i could eat it for breakfast...
07: USER: chocolate for breakfast? yuck!
08: SYSM: what do you eat for breakfast instead of chocolate?
09: USER: toast.
10: SYSM: what do you like to eat on your toast.
11: USER: ice cream.
12: SYSM: yum yum! delicious.
13: SYSM: do you want to make some ice cream?
14: USER: yeah.
15: SYSM: do you have the recipe with you?
16: USER: can you please tell me a story?
17: SYSM: sure. susie the ice cream moose was...
...
23: SYSM: that's the end of the story.
24: SYSM: let's continue with our cooking. so, do you have the recipe?
25: USER: no.
26: SYSM: you'll need milk and chocolate.
27: USER: where does chocolate come from?
28: SYSM: chocolate comes from cacao trees.
29: SYSM: let's continue with our cooking, do you need to get out a pan?
...
34: USER: it's done!
35: SYSM: yum yum! your ice cream is ready.
36: SYSM: now, let's talk about the zoo, ok?
```

Fig. 2. An example conversation

```
           USER: i really like cooking.
(a)        SYSM: cooking is great.
           USER: yeah. i agree.
           SYSM: what do you want to cook for your friends.

           USER: i really like cooking.
(b)        SYSM: if we do some cooking, we can share it with  your friends.
           USER: i love cooking and sharing.
           SYSM: let's decide what to cook.
```

Fig. 3. Examples showing variability in system responses

5 Conclusion

We have described a goal-oriented approach to managing Conversational Activities between a user and a conversational virtual companion. We have performed an initial evaluation, with human subjects rating the system for coherence, interesting variability in outputs, and overall appropriately sustained focus over a conversation. Overall rating by the subjects was positive, notwithstanding relatively limited coverage in terms of content and knowledge base. We have recently developed techniques to automatically mine content from web forums [10], which should significantly alleviate this problem. We are also developing a model of *proactive engagement*, whereby the conversational agent actively monitors user engagement [11] and applies conversational strategies when required. Another important avenue of research is how to incorporate explicit strategies for long-term relationship-building [12], including gathering and making effective use of personalised information.

Acknowledgement. This work is partially supported by the Australian Research Council and Real Thing Entertainment Pty. Ltd. under grant number LP110100050.

References

1. Cavazza, M., Camara, R., Turunen, M.: How was your day? a companion ECA. In: Proc. AAMAS, Toronto (2010)
2. Wooldridge, M.: Introduction to MultiAgent Systems, 2nd edn. Wiley & Sons (2009)
3. Winikoff, M., Padgham, L.: Developing Intelligent Agent Systems: A Practical Guide. Wiley Series in Agent Technology. Wiley and Sons (2004)
4. Nguyen, A., Wobcke, W.: An agent-based approach to dialogue management in personal assistants. In: Proc. IUI, San Diego, California (2005)
5. van Oijen, J., van Doesburg, W., Dignum, F.: Goal-based communication using bdi agents as virtual humans in training: An ontology driven dialogue system. In: Proc. AAMAS Workshop on Agents for Games and Simulation, Toronto (2010)
6. Klein, D., Manning, C.: Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the ACL (2003)
7. Tang, H., Tan, S., Cheng, X.: A survey on sentiment detection of reviews. Expert Systems with Applications 36(7), 10760–10773 (2009)
8. Brandtzaeg, P., Folstad, A., Heim, J.: Enjoyment: Lessons from Karasek, pp. 55 – 65. Springer (2006)
9. Macias-Galindo, D., Wong, W., Thangarajah, J., Cavedon, L.: Coherent topic transition in a conversational agent. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (InterSpeech), Oregon, USA (2012)
10. Wong, W., Thangarajah, J., Padgham, L.: Health conversational system based on contextual matching of community-driven question-answer pairs. In: Proc. CIKM Demo Track, Glasgow, pp. 2577–2580 (2011)
11. Castellano, G., Pereira, A., Leit, I., Paiva, A., McOwan, P.: Detecting user engagement with a robot companion using task and social interaction-based features. In: Proc. ICMI, Cambridge, USA (2009)
12. Bickmore, T., Picard, R.: Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer Human Interaction (ToCHI) 12(2), 293–327 (2005)

Synthesising and Evaluating Cross-Modal Emotional Ambiguity in Virtual Agents

Matthew P. Aylett[1,2] and Blaise Potard[2]

[1] University of Edinburgh,
matthewa@inf.ed.ac.uk
http://homepages.inf.ed.ac.uk/matthewa/
[2] CereProc Ltd. Edinburgh

Abstract. Emotional ambiguity, when more than one emotion appears present at a given time, or several emotions are superimposed, is common in human interaction and effects such as irony can be intentionally created through a mismatch of such emotional signals. High quality emotional speech synthesis offers a means for testing the effect of combining differences in vocal emotion, facial expression and text content in a virtual agent. In this paper we combine high quality emotional speech synthesis with a video rendered non-naturalistic virtual agent. Vocal emotion and text content combined to increase or decrease the emotional valence (positivity) of an utterance, while emotional facial expressions did not affect valence, but interacted with vocal emotion altering emotional activation in the lax and stressed vocal condition.

Keywords: speech synthesis, unit selection, expressive speech synthesis, emotion, prosody, facial animation.

1 Introduction

In this work we address the challenges of synthesising and evaluating cross-modal emotional ambiguity in virtual agents by:

1. Evaluating utterances using a parametric *activation/evaluation* space[1–3] (Figure 1a). This allows the evaluation of magnitude across two dimensions, activation - how active or passive a subject rates an utterance, evaluation - how positive or negative a subject rates an utterance. The experiment was carried out online using 12 native English speakers.
2. Combining three modalities: Textual content, emotional speech synthesis based on stressed/lax voice quality changes[4–7] and synthesised angry/neutral/happy facial expressions using a non-naturalistic animated head[8] (Figure 2), and evaluating the interactions between them.

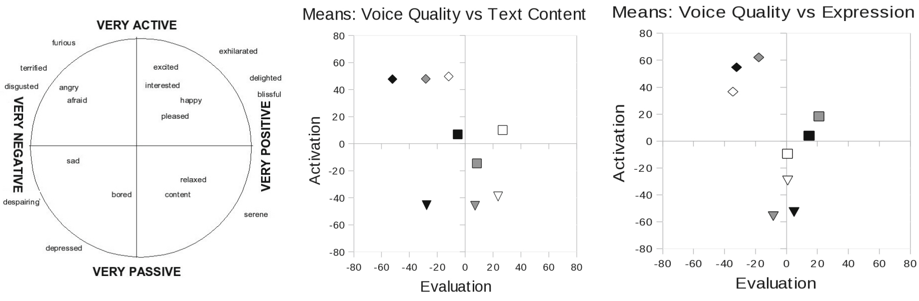
Our research questions are as follows:

RQ1: Are negative and positive features of the three modalities additive in the evaluation domain?

RQ2: Does a mismatch of features across modalities produce an ambiguous emotion? If so can it be distinguished from a neutral rendition?

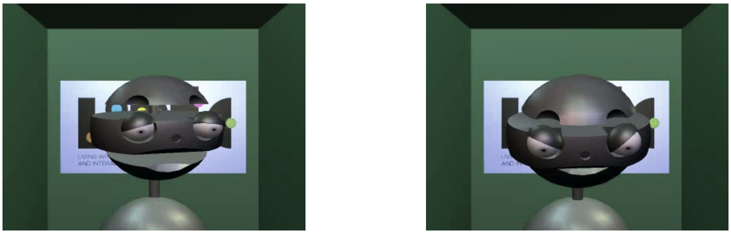
Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 471–473, 2012.

2 Results



a) b) c)

Fig. 1. a) Activation/Evaluation Space (radius 170). b) Mean activation/evaluation of materials by voice quality and text type. Voice quality significantly affected valence(F(2, 26)=17.93, p<0.001) and activation(F(2, 26)=98.53, p<0.001), Text type significantly affected valence only (F(2, 26)=25.47, p<0.001) c) Mean activation/evaluation of materials by voice quality and facial expression. Significant interaction between expression and voice quality (F(4, 26)=4.02, p<0.005). Diamond - Stressed VQ, Square - Neutral VQ, Triangle - Lax VQ. White - Positive Text/Happy Expression, Grey - Neutral Text/Neutral Expression, Black - Negative Text/Angry Expression.



Happy and Angry Facial Poses

Fig. 2. Virtual EMYS

3 Discussion

Results show that we can merge speech style and text content to create a different perception of the emotion in a message wich is additive in the evaluation dimension. In contrast, facial expression had no significant effect on valence. Instead it interacted in a complex way with voice quality, significantly affecting activation when it mismatched the underlying voice quality.

A mismatch between text content and voice quality does make the emotion more ambiguous (closer to the centre point of the activation/evaluation space). In addition, the happy facial expression causes an increase in activation for the lax condition but a decrease for the stressed condition[1,2].

[1] Four videos are available at:
http://homepages.inf.ed.ac.uk/matthewa/iva2012emys/

[2] This research was funded by the Royal Society through a Royal Society Industrial Fellowship.

References

1. Schlosberg, H.: A scale for judgement of facial expressions. Journal of Experimental Psychology 29, 497–510 (1954)
2. Plutchik, R.: The Psychology and Biology of Emotion. Harper Collinns, New York (1994)
3. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: FEELTRACE: An instrument for recording perceived emotion in real time. In: ISCA Workshop on Speech and Emotion, pp. 19–24 (2000)
4. Gobl, C., Chasaide, A.N.: The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189–212 (2003)
5. Schröder, M., Grice, M.: Expressing vocal effort in concatenative synthesis. In: ICPhS, pp. 2589–2592 (2003)
6. Aylett, M.P., Pidcock, C.J.: The cerevoice characterful speech synthesiser sdk. In: AISB, pp. 174–178 (2007)
7. Aylett, M., Pidcock, C.: UK patent GB2447263A: Adding and controlling emotion in synthesised speech (2012)
8. Ribeiro, T., Paiva, A.: The illusion of robotic life: principles and practices of animation for robots. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2012, pp. 383–390. ACM, New York (2012)

Immersive Interfaces for Building Parameterized Motion Databases

Yazhou Huang, Carlo Camporesi, and Marcelo Kallmann

University of California, Merced

Abstract. We present new interface tools for the interactive motion modeling and construction of parameterized motion databases for virtual agents. The tools provide different forms of visual exploration of database coverage, providing an intuitive way to model and refine motions by direct demonstration.

In interactive virtual training and assistance applications, virtual agents are driven by realistic motion synthesis techniques with parameterized variations in respect to a given scenario. Building such database can be complex, time-consuming and in many cases must be done by experts. Common solutions for the motion modeling process rely on hand-crafted motions [2, 12, 13], gestures synthesized with algorithmic procedures [5, 6], motion blending based on captured data [7, 8, 10, 11], etc. We improve the modeling phase of our existing interactive motion modeling framework by direct demonstration [1]. The frameworks is based on a full-scale 3D display facility, and has been extended to include intuitive interfaces to build, visualize, evaluate and refine a motion database in respect to the spatial coverage inside a simulated workspace, guiding the on-line programming of scenario-specific examples.

Interface Description. Our system targets situations where the user is able to model clusters of action or gesture motions by direct demonstration. This is done via either a wearable *gesture vest* [3] or Vicon tracking system for capturing upper-body motions together with data gloves for capturing hand motions. A WiiMote serves as an interface to create new cluster types, start/stop capture, playback, trim, annotate stroke points, save or delete motions. Clusters defined by examples is an important concept for specifying parameterized action or gesture types, and motions within a same cluster are blended to consistently represent variations of the same type. The user can also examine the database spatial coverage inside the virtual workspace with two visualization methods (described below) for guidance on improving the database coverage as needed. The coverage refers to how well each motion cluster is parametrized using *inverse blending* [4] from the discrete examples in order to satisfy specified spatial constraints. Fig 1 top left outlines framework pipeline, top right shows one motion being reviewed for editing.

Database Spatial Coverage Visualization. The ability to enforce constraints for new motions greatly depends on the existing variations among the example motions. In general, a small number of carefully selected example motions can provide good coverages for the regions of interest (ROIs) in the workspace. We propose two specific visualization methods rendering a palette of colors [9, 14, 15] inside the workspace to intuitively

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 474–476, 2012.

guide the user during the process of adding new motions to refine the database for improved coverage: Workspace Volume Visualization **(WV)** and Local Coverage Visualization **(LV)**. See Fig 1. WV conducts a coarse uniform sampling of the workspace and presents the overall spatial coverage with colored cubes for the entire workspace without the need to define an overly fine subdivision of the constraint space. Each cube represents a reaching target (spatial constraint), and a motion synthesized towards each cube is measured by reaching precision (error e^*) using a constraint evaluation function, and the value $e^*/e_{max}(\in [0,1])$ is mapped onto a hue color space then assigned to each cube. For a reasonably sized database WV takes a few seconds to generate, then the user can immediately spot areas with low coverage by the color of the cubes (red or orange), and add additional motion towards these areas. LV renders a transparent colored mesh geometry covering a small ROI, delimiting the coverage evaluation within its volume. It focuses on the local coverage visualization taking only milliseconds to be computed, and it is suitable for fine tuning coverage of smaller volumes when only small local regions are of interest. LV uses the same color mapping but applied to mesh vertices. LV follows the movement of the user's hand, its size and shape can be iteratively changed for either fast sweeping over large ROIs (a table surface) or for carefully checking small ROIs (buttons, etc). LV is also able to utilize motions dynamically added to the database without any pre-computation lag. Please refer [4] for details on motion synthesis and error evaluation with spatial constraints.

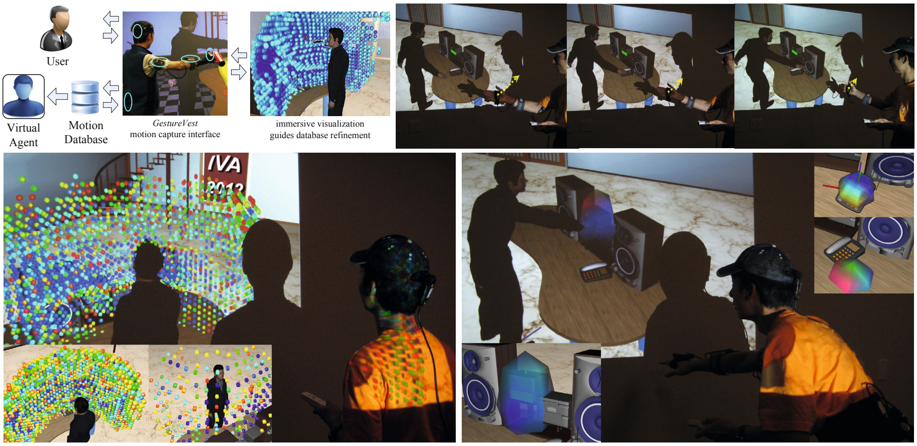

User
Virtual Agent
Motion Database
GestureVest motion capture interface
immersive visualization guides database refinement

Fig. 1. Top-left: framework pipeline. Top-right: user moves his hand to scroll through a motion being edited. Bottom-left: Workspace Volume Visualization mode gives an overview of database coverage, density and error threshold can be adjusted for clear viewing. Bottom-right: Local Coverage Visualization mode, ideal for checking small ROIs like dials and buttons.

Conclusions and Acknowledgments Our proposed tools greatly improve the process of interactive motion modeling and the overall approach constitutes a powerful approach for programming virtual agents. This work was partially funded by NSF award IIS-0915665.

References

1. Camporesi, C., Huang, Y., Kallmann, M.: Interactive Motion Modeling and Parameterization by Direct Demonstration. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 77–90. Springer, Heidelberg (2010)
2. Gebhard, P., Kipp, M., Klesen, M., Rist, T.: What are they going to talk about? towards life-like characters that reflect on interactions with users. In: Proc. of the 1st International Conference on Technologies for Interactive Digital Storytelling and Entertainment, TIDSE 2003 (2003)
3. Huang, Y., Kallmann, M.: Interactive Demonstration of Pointing Gestures for Virtual Trainers. In: Proceedings of 13th International Conference on Human-Computer Interaction, San Diego, CA (2009)
4. Huang, Y., Kallmann, M.: Interactive motion modeling and parameterization by direct demonstration. In: Proceedings of the 3rd International Conference on Motion in Games, MIG (2010)
5. Kallmann, M.: Analytical inverse kinematics with body posture control. Computer Animation and Virtual Worlds 19(2), 79–91 (2008)
6. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents: Research articles. Computer Animation and Virtual Worlds 15(1), 39–52 (2004)
7. Kovar, L., Gleicher, M.: Automated extraction and parameterization of motions in large data sets. ACM Transaction on Graphics (Proceedings of SIGGRAPH) 23(3), 559–568 (2004)
8. Mukai, T., Kuriyama, S.: Geostatistical motion interpolation. In: ACM SIGGRAPH, pp. 1062–1070. ACM, New York (2005)
9. Rodriguez, I., Peinado, M., Boulic, R., Meziat, D.: Bringing the human arm reachable space to a virtual environment for its analysis. In: IEEE International Conference on Multimedia and Expo. (2003)
10. Rose, C., Bodenheimer, B., Cohen, M.F.: Verbs and adverbs: Multidimensional motion interpolation. IEEE Computer Graphics and Applications 18, 32–40 (1998)
11. Rose III, C.F., Sloan, P.-P.J., Cohen, M.F.: Artist-directed inverse-kinematics using radial basis function interpolation. Computer Graphics Forum (Proceedings of Eurographics) 20(3), 239–250 (2001)
12. 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 23(3), 506–513 (2004)
13. Thiebaux, M., Marshall, A., Marsella, S., Kallmann, M.: Smartbody: Behavior realization for embodied conversational agents. In: Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS (2008)
14. Yang, J., Sinokrot, T., Abdel-Malek, K., Beck, S., Nebel, K.: Workspace zone differentiation and visualization for virtual humans. Ergonomics 51(3), 395–413 (2008)
15. Zacharias, F., Howard, I.S., Hulin, T., Hirzinger, G.: Workspace comparisons of setup configurations for human-robot interaction. In: IROS, pp. 3117–3122 (2010)

Creating Personalized and Distributed Virtual Learning Spaces through the Use of i-Collaboration 3.0

Eduardo A. Oliveira[1,2], Patrícia Tedesco[1], and Thun Pin T.F. Chiu[1]

[1] Informatic Centre, Federal University of Pernambuco (UFPE)
[2] Recife Center for Advanced Studies and Systems (CESAR),
Recife – PE – Brazil
{eao,pcart,tptfc}@cin.ufpe.br

Different from what happened a few years ago, distance learners are now much more familiar with Internet resources (in its various platforms and social networks – thus characterizing what is now known as digital nomadism) and are used to work in groups. With this paradigm change, it becomes even harder to do students of virtual courses get interested in a traditional VLE, where the interfaces are not appropriate to their particular needs and often collaborative tools do not allow these students to establish relations with other colleagues, unlike what happens naturally in social networks.

To contribute to the minimization of the challenges found in VLEs (communication difficulties, centralized access, interoperability and data integration, ...), that we believe also contribute to minimizing the various problems currently found in the distance education (motivation and isolation feeling), this work has as main objective to use an intelligent agent to contribute to the creation of distributed virtual learning spaces. The support will be provided through the proposed i-collaboration 3.0 model, an extension of i-collaboration model (1.0) [1,2,3,4].

I-Collaboration 3.0 tries to ensure decentralized (distributed) access to learning contents available in different Web 2.0 tools (Twitter, MSN, Blogs, ...) and social networks (Facebook, Orkut, ...) through an intelligent agent support. The intelligent agent in i-collaboration 3.0 integrates students' distributed data and personalize the learning contents (the students' are distributed in the Internet – the same student can learn using MSN, Facebook and Twitter, at the same time, for example), based on the particular tastes and needs of each student (identified through de student behavior in Twitter, MSN, ...). With the virtual learning spaces support, the students will be able to study through the Web, using platforms and environments that they already meet and frequently use.

To better understand i-collaboration 3.0 model, we can assume, as an example scenario, that Twitter, MSN, a blog (Blogger site), Facebook and Moodle (VLE) are integrated with the i-collaboration 3.0 model. To assume that the system is integrated with these environments is to assume that these environments are using i-collaboration 3.0. In the presented scenario, a single instance of an intelligent agent, which is provided by the i-collaboration 3.0 model, is available in each of these environments (such as a contact on MSN, as a user in Twitter, and as a chatterbot in Sites and VLEs, ...). Despite the fact that the intelligent agent appears in many

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 477–479, 2012.

different environments, the model provides a single agent to these environments, this ensures that the same intelligent agent will be used in all environments. The student talks across different environments, with the same bot. If a computer science student starts communicating with the intelligent agent in Gtalk, asking him about the main function of a program: "what is a main function?" he will get an answer about the main function, as requested. A few minutes later, the student goes to the MSN and asks the same thing to the intelligent agent: "main" (because he is still with doubts). At this time, the intelligent agent recognizes the student (that has communicated with him through Gtalk) asking him about the same thing (and in a few interval of time – context sensitive [5]). The intelligent agent infers about student question, student environment, studied contents, student profile [6] and answers him with new questions: "We do not talk about it?" "You need more help with this issue?". If student needs more help, the intelligent agent must suggest to this student related contents based on his doubts in programming introduction.

The advantage to provide a single intelligent agent in the system is in the fact that with only one agent, we can also have a single integrated database in the model (based on students' interaction with the agent in distributed environments). With a single database, the student, which communicates with the agent in Gtalk and in MSN, can now be identified on these and in any other environment that the intelligent agent is presented. If a student interacts with the intelligent agent through Facebook, the agent will know, referring to the historical database of the student that he has already communicated with him through Twitter and MSN, and that he demonstrated interest in studying programming concepts before.

As a way of enabling the various customers running separately and accessing the same database of the proposed model, all in a distributed way (even on different servers) and custom, we chose to work with RMI architecture. In this architecture, different customers of different environments (MSN, Twitter, Web, ...) can work in a distributed computing (multiple JVMs). The client (company or institution that wishes to have an instance of the i-collaboration 3.0 available) download the model API and implements a method for the environment that he wants and get an instance of the system. The main class has the following method signature: public String getResponse (String 'questionText, String userId, EnvironmentType environmentType). Any customer interested in using the i-collaboration 3.0 must use this method, stating the text of the student, the student ID and the environment that the student is communicating with the intelligent agent.

This work is actually implemented, and was tested during two months by two different developers (proof of concept) to check the integration of Twitter, MSN and Gtalk with i-collaboration 3.0, the calibration of the learning contents, the server stability and the content adaptation and personalization quality. The tests showed promising results (results obtained from logs analysis and lines of code written during integration processes – with and without i-collaboration support, comparative analysis). Since 03/26/12 i-collaboration 3.0 is being experimented by 65 undergrad students, working with a beginner's Programming course. The experiment ends in 05/26/2012. The system is running at Amazon Cloud Computer Servers.

Through the development of i-collaboration 3.0 concept and model, this paper presented a contribution to mitigate the problems that have made difficult the use of distributed personalized learning. The model seeks to deal with each student in a unique way, in the environment that the student feels better, thus motivating these students to learn. I-collaboration 3.0 supports the creation of virtual learning spaces.

References

[1] Oliveira, E.A., Tedesco, P.: Putting the Intelligent Collaboration Model in practice within the Cleverpal Environment. In: 2009 International Conference of Soft Computing and Pattern Recognition (IEEE), Malacca, Malaysia, pp. 687–690 (2009), doi:10.1109/SoCPaR.2009.13

[2] Oliveira, E.A., Tedesco, P.: i-Collaboration in Practice: Results from our investigation within the Cleverpal Environment. In: IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS 2009), Shanghai, China, November 20-22 (2009) 978-1-4244-4738-1/09

[3] Oliveira, E., Tedesco, P.:i-collaboration: Um modelo de colaboração inteligente personalizada para ambientes de EAD". Anais do XVIII Simpósio Brasileiro de Informática na Educação - SBIE, São Paulo - SP, pp. 412–421 (2007) ISBN 978-85-7669-157-0

[4] Oliveira, E.A., Tedesco, P.: i-collaboration: Um modelo de colaboração inteligente personalizada para ambientes de EAD. Revista Brasileira de Informática na Educação 18, 17–31 (2010)

[5] Vieira, V.: CEManTIKA: A Domain-Independent Sistema for Designing Context-Sensitive Systems. In: Tese de Doutorado, Centro de Informática – UFPE, Brasil (2008)

[6] MBTI – Myer Briggs Type Indicator, http://www.myersbriggs.org

Eliciting Gestural Feedback in Chinese and Swedish Informal Interactions

Jia Lu

Division of Cognition and Communication, Department of Applied IT,
Chalmers & Göteborgs universitet,
Göteborg, Sweden
jia.lu@gu.se

Abstract. In this paper, how people use gestural feedback to elicit further interaction is studied in Chinese and Swedish intercultural and mono-cultural interactions. The results can be used to make the intelligent virtual agents in Chinese and Swedish communication contexts behave more human-like.

Keywords: Feedback, eliciting, gestural, Chinese, Swedish, emotion, attitude.

1 Introduction

In human and virtual human interactions, both communicative gestures and eliciting feedback have been studied by many researchers [1] to [5], but there is not much research on eliciting gestural feedback. In this paper, gestural feedback that functions as an eliciting device is studied in the Chinese and Swedish mono-cultural and intercultural first encounters. The results can be used to make the intelligent virtual agents in Swedish and Chinese communication contexts behave more human-like.

2 Purpose

Three research questions are investigated. First, what is the typical eliciting gestural feedback used by Chinese and Swedish speakers in their mono-cultural interactions? Second, what is used when they speak in English in intercultural interactions? Third, what emotions and attitudes are expressed through the eliciting gestural feedback?

3 Data and Method

Because of the Swedish and Chinese cultural differences [6][7] and the cultural impact on interaction [8], the Swedish and Chinese participants were selected. Four Chinese, four Swedish, and eight Chinese-Swedish face-to-face dyadic dialogs were video-recorded. The communicative activity is first acquaintance meeting, because people employ many eliciting devices to get acquainted. Each recording lasts six to

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 480–482, 2012.

ten minutes. The data was annotated according to the GTS[1] (Göteborg Transcription Standard) [9] and the MUMIN multimodal coding scheme[2] [10]. Inter- and intra-coder reliability checking was done between six Chinese and Swedish annotators.

4 Result and Analysis

The data shows that most of the eliciting gestural feedback has vocal-verbal accompaniment. In the mono-cultural interactions, the Chinese speakers have seven times more eliciting gestural feedback than the Swedes. This might be because when the Chinese meet abroad, they feel stronger intimacy and they are more eager to find out how the other is doing. Both Chinese and Swedish speakers use *eyebrow rise* and *head forward*; whereas, g*aze movement* is typical Chinese and *posture forward* is typical Swedish. Many emotions are expressed, and one feedback can express more than one emotion. Uncertainty and surprise are the most frequent ones. This is probably due to the insecurity or uncertainty that they may feel in first acquaintance.

In the intercultural interactions, the Swedes use more eliciting gestural feedback than the Chinese, which is different from the mono-cultural results. This may be because the Swedes are local people and they feel more confident. Both Chinese and Swedish speakers use *eyebrow rise* and *head down-nod(s)* most frequently. The Swedes have more *eyebrow rise* than what they have in the mono-cultural interactions, and the Chinese start using *head forward* and *head tilt* [11]. This may be because of the participants' mutual influence [12] and co-activation [13]. Besides, the most frequently communicated emotions are uncertainty, surprise, and interest.

5 Conclusion

This paper primarily addresses three questions. First, what is the typical eliciting gestural feedback used by Chinese and Swedish speakers in their mono-cultural interactions? Second, what are used when they communicate in English in inter-cultural interactions? Third, what emotions and attitudes are expressed?

The Chinese interlocutors have more eliciting gestural feedback than the Swedes in the mono-cultural interactions; whereas, the Swedish speakers have more in the intercultural interactions. Most of the eliciting gestural feedback expressions are used

[1] The Gothenburg Transcription Standard was mainly created by Joakim Nivre, Jens Allwood, Leif Grönqvist, Magnus Gunnarsson, Elisabeth Ahlsén, Hans Vappula, Johan Hagman, Staffan Larsson, Sylvana Sofkova, and Cajsa Ottesjö in Department of Linguistics of Göteborg University. It is a standard for machine-readable transcriptions of spoken language first used within the research program Semantics and Spoken Language at Göteborg University. Recently it has been more popularly used as a transcription standard for the study of spoken language features and social activity patterns.

[2] The MUMIN multimodal coding scheme was mainly created by Jens Allwood, Loredana Cerrato, Laila Dybkær, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. It was originally created to annotate the multimodal communicative behaviors in the video clips. Recently, it has been used as a general instrument for the study of gestures and facial displays in interpersonal communication, in particular the role played by multimodal expressions for feedback, turn management and sequencing.

in combinations with vocal-verbal means to express more than one emotion and attitude. In the mono-cultural interactions, both Chinese and Swedish speakers use *eyebrow rise* and *head forward* as the most common eliciting gestural feedback. *Gaze movement* is typical Chinese and *posture forward* is typical Swedish. They are most frequently used to express uncertainty and surprise. Meanwhile, in the intercultural interactions, *eyebrow rise* and *head down-nod(s)* are most frequently used by both Chinese and Swedish interlocutors. The Swedes have more *eyebrow rise* than what they have in the mono-cultural interactions, and the Chinese start using *head forward* and *head tilt*. The most frequently communicated emotions are uncertainty, surprise, and interest. Finally, since the size of this study is relatively small and the activity is limited to informal first encounters, it still necessitates further investigation.

Acknowledgement. I would like to thank Jens Allwood, Elisabeth Ahlsén, Alexander Holender, Karl Johan Sandberg, Xi Chen, and Yansi Xu at University of Gothenburg.

References

1. Kita, S.: Cross-cultural variation of speech-accompanying gesture: A review. Language and Cognitive Processes - Lang Cognitive Process 24(2), 145–167 (2009)
2. Gullberg, M.: Methodological reflections on gesture analysis in SLA and bilingualism research. Second Language Research 26, 75–102 (2010)
3. Jokinen, K., Navarretta, C., Paggio, P.: Distinguishing the Communicative Functions of Gestures. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 38–49. Springer, Heidelberg (2008)
4. Miller, N., Resnick, P., Zeckhauser, R.: Eliciting Informative Feedback: The Peer-Prediction Method. Management Science 51(9), 1359–1373 (2005)
5. Smith, H., Fitzpatrick, G., Rogers, Y.: Eliciting reactive and reflective feedback for a social communication tool: a multi-session approach. In: Proceedings of the 5th Conference on Designing Interactive Systems (DIS 2004), pp. 39–48. ACM, New York (2004)
6. Hall, E.T.: Beyond Culture. Anchor, Garden City (1977)
7. Hofstede, G.: Cultures' Consequences: Comparing Values, Behaviors, Institutions, and Organizations across Nations, 2nd edn. Sage, Thousand Oaks (2001)
8. Endrass, B., Rehm, M., André, E.: Planning Smalltalk Behavior with Cultural Influences for Multiagent Systems. Computer, Speech, and Language 25(2), 158–174 (2011)
9. Nivre, J.: Göteborg Transcription Standard. Version 6.2, 38, Göteborg University, Department of Linguistics, Göteborg (1999)
10. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN Coding Scheme for the Annotation of Feedback. In: Martin, J.C., et al. (eds.) Multimodal Corpora for Modelling Human Multimodal Behaviour, vol. 41(3-4), pp. 273–287 (2007)
11. Lu, J., Allwood, J.: Unimodal and Multimodal Feedback in Chinese and Swedish Mono-cultural and Intercultural Interactions. In: Proceedings of the 3rd Nordic Symposium on Multimodal Communication (2011), `http://dspace.utlib.ee/dspace/handle/10062/22545`
12. Allwood, J., Lindström, N.B., Lu, J.: Intercultural Dynamics of Fist Acquaintance: Comparative Study of Swedish, Chinese and Swedish-Chinese First Time Encounters. In: Stephanidis, C. (ed.) HCII 2011 and UAHCI 2011, Part IV. LNCS, vol. 6768, pp. 12–21. Springer, Heidelberg (2011)
13. Allwood, J., Lu, J.: Unimodal and Multimodal Co-activation in First Encounters -— A Case Study. In: Proceedings of the 3rd Nordic Symposium on Multimodal Communication. Finland, Helsinki, May 27-28 (2011), `http://dspace.utlib.ee/dspace/handle/10062/22540`

Locus of Control in Conversational Agent Design: Effects on Older Users' Interactivity and Social Presence

Veena Chattaraman[1], Wi-Suk Kwon[1], Juan Gilbert[2], and Shelby Darnell[2]

[1] Department of Consumer Affairs, Auburn University, Auburn AL 36849, U.S.A
{vzc0001,kwonwis}@auburn.edu
[2] School of Computing, Clemson University, Clemson SC 29634, U.S.A
{juan,sdarnel}@clemson.edu

Abstract. This study examines the impact of locus of control in conversational agent design on the interaction experience of 61 older users (65+ years old) through a laboratory experiment. Results reveal that for older users, agent control facilitates greater interactivity and social presence than user control.

Keywords: Conversational agent, older users, locus of control.

1 Introduction

The U.S. senior population (aged 65+) will reach 71.5 million by 2030 [1]. Only 35% of U.S. seniors use the Internet [2]. Recent older user studies reveal that conversational agents (CAs) enhance their Internet adoption through enhanced perceptions of social support, trust, and hedonic and functional benefits [3-5]. However, no studies have examined the effect of locus of control on older users' interaction with CAs.

2 Conceptual Background

The concept of locus of control lies in matching the agent's abilities to the user's expectations [6]. The constructivist (high user control) to instructivist (high agent control) continuum [7, 8] provides a useful conceptualization in determining the locus of control. The former places more agency with the user, limiting the agent's role to supporting and guiding the user's cognitive abilities; whereas the latter places more agency with the agent which uses traditional modes of teaching/directing to deliver information/knowledge to the user [7]. No studies have examined the effect of the level of control exercised by an agent on older user experience. We propose and test whether the level of agent control will influence interactivity and social presence, reduce the cognitive and socio-psychological barriers faced by older users in web interaction, and contribute to increased website reuse intent among older users.

3 Method

The virtual agent was created using SitePal and employed in a mock e-commerce website. A laboratory experiment was conducted with a 3-condition (agent control

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 483–485, 2012.

[AC], user control [UC], no agent; see Table 1) between-subjects design. A total of -61subjects (65-83 years old, M = 70.8) participated in the study. Participants were randomly assigned to a condition and performed a task of purchasing a product.

Table 1. Locus of control manipulations

Locus of Control	Agent Control (AC)	User Control (UC)
Role of agent and user	• **Agent acts as doer/**performer. • Agent performs tasks • Agent proactively seeks information from user • Agent actively completes the task on the user's behalf	• **Agent acts as helper/**facilitator • User performs task • Agent passively reacts to user's expressed information needs • Agent provides information to help user perform the task

4 Evaluation

Following the task, participants completed a questionnaire with dependent measures: perceptions of interactivity [9], social presence [10], ease of use [11], ease of information search [12], usefulness [13], efficacy [14], trust [15], and social support [16], and intent to reuse the website, which were all rated on a 5-point Likert-type scale. Analysis of Variance results revealed that the use of an agent significantly increased users' perceptions of interactivity ($M_{\text{no agent}} = 3.23$, $M_{\text{UC}}= 3.58$, $M_{\text{AC}}= 3.95$; $F_{2,57}= 5.50$, $p < .01$) and social presence ($M_{\text{no agent}} = 2.83$, $M_{\text{UCagent}}= 3.22$, $M_{\text{AC}}= 3.45$; $F_{2,57} = 2.83$, $p < .10$), and the AC was more effective than the UC in enhancing interactivity and social presence. Regression analysis results also showed that users' cognitive barriers to using the website were significantly reduced through the increased social presence and interactivity via the agents. Specifically, perceived ease of use (β = .33, $p < .05$), ease of information search (β = .35, $p < .05$), and usefulness of the website (β = .30, $p < .05$) were all positively influenced by increased interactivity. Further, perceived usefulness was also positively influenced by perceived social presence (β = .46, $p < .001$). In addition, the increased interactivity and social presence via the agents also contributed to reducing users' socio-psychological barriers. Specifically, increased interactivity positively influenced users' perceived social support (β = .43, $p < .001$), trust (β = .27, $p < .05$), and efficacy (β = .40, $p < .001$); whereas increased social presence led to increased perceptions of social support (β = .42, $p < .001$) and trust (β = .48, $p < .001$). Finally, regression analysis also revealed that the cognitive and socio-psychological benefits of using the agents led to users' enhanced intention to reuse the website; perceived usefulness (β = .43, $p < .001$), and trust (β = .35, $p < .05$) were the most influential predictors of reuse intent. These results demonstrate that for older users, designing agents with greater control in the interaction is more effective in enhancing interactivity and social presence in the Internet interface.

References

1. Administration on Aging, Department of Health & Human Services, `http://www.aoa.gov/AoARoot/Aging_Statistics/index.aspx`
2. Pew Internet, `http://www.pewinternet.org/trends/User_Demo_7.22.08.htm`
3. Chattaraman, V.: Kwon, W.-S., Gilbert, J.: Social Presence in Online Stores: Building Social Support and Trust among Older Consumers. In: Evans, J. R. (ed.) Retailing Strategic Challenges and Opportunities in Uncertain Times: Special Conference Series 2009, vol. XII. Proceedings of the 12th Triennial National Retailing Conference of the Academy of Marketing Science and the American Collegiate Retailing Association, New Orleans (2009) (CD-ROM)
4. Kwon, W.-S., Chattaraman, V., Gilbert, J.: Effects of Conversational Agents in Retail Web sites on Aging Consumers' Interactivity and Perceived Benefits. In: Mynatt, E., Fitzpatrick, G., Hudson, S., Edwards, K., Rodden, T. (eds.) Proceedings of the 28th International Conference on Human Factors in Computing Systems, Atlanta (2010)
5. Kwon, W.-S., Chattaraman, V., Shim, S.I., Alnizami, H., Gilbert, J.: Older User-Computer Interaction on the Internet: How Conversational Agents Can Help. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part II. LNCS, vol. 6762, pp. 533–536. Springer, Heidelberg (2011)
6. Erickson, T.: Designing Agents as if People Mattered. In: Bradshaw, J.M. (ed.) Software Agents, pp. 79–96. MIT Press, Menlo Park (1997)
7. Baylor, A.L.: Permutations of Control: Cognitive Considerations for Agent-Based Learning Environments. Journal of Interactive Learning Research 12(4), 403–425 (2001)
8. Reiber, L.P.: Computers, Graphics, & Learning. Brown & Benchmark, WI (1994)
9. Liu, Y.: Developing a Scale to Measure the Interactivity of Websites. Journal of Advertising Research 43, 207–216 (2003)
10. Gefen, D.: Manage User Trust in B2C e-Services. E-Service Journal 2(2), 7–24 (2003)
11. Klein, R.: Internet-Based Patient-Physician Electronic Communication Applications: Patient Acceptance and Trust. E-Service Journal 5(2), 27–51 (2006)
12. Page-Thomas, K.: Measuring Task-Specific Perceptions of the World Wide Web. Behaviour & Information Technology 25(6), 469–477 (2006)
13. Ahn, T., Ryu, S., Han, I.: The Impact of the Online and Offline Features on the User Acceptance of Internet Shopping Malls. Electronic Commerce Research Applications 3, 405–420 (2004)
14. Tsai, M.-J., Tsai, C.-C.: Information Searching Strategies in Web-Based Science Learning: The Role of Internet Self-Efficacy. Innovations in Education and Teaching International 40(1), 43–50 (2003)
15. Klein, R.: Internet-Based Patient-Physician Electronic Communication Applications: Patient Acceptance and Trust. E-Service Journal 5(2), 27–51 (2007); adopted from Davis, F. D.: Perceived Usefulness, Perceived ease of Use, and User Acceptance of Information Technology. MIS Quarterly. 13(3), 319–340 (1989)
16. Zimet, G.D., Dahlem, N.W., Zimet, S.G., Farley, G.K.: The Multidimensional Scale of Perceived Social Support. Journal of Personality Assessment 52(1), 30–41 (1988)

Online Behavior Evaluation with the Switching Wizard of Oz

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,m.termaat,d.k.j.heylen}@utwente.nl

1 Introduction

Advances in animation and sensor technology allow us to engage in face-to-face conversations with virtual agents [1]. One major challenge is to generate the virtual agent's appropriate, human-like behavior contingent with that of the human conversational partner. Models of (nonverbal) behavior are pre-dominantly learned from corpora of dialogs between human subjects [2], or based on simple observations from literature (e.g. [3,4,5,6]).

Humans are particularly sensitive to flaws in the displayed behavior, both in form and timing [7,8]. This effect also occurs when certain behaviors are not animated, which is common in experimental settings where the behavior of the virtual agent is varied systematically only one or a few modalities [9,10]. This leads to biased perceptual ratings, which hampers progress in the design and implementation of behavior synthesis algorithms.

To this end, we propose a methodology and implementation that combines ideas behind the human Turing test with those of a Wizard of Oz setup. At the heart of our approach is a distributed (video-conferencing) setting with two human conversational partners. Each of the subjects is observed with camera and microphone and algorithms are employed to analyze the verbal and nonverbal behavior in real-time (similar to e.g., [11,12,13]). These observations are used as input to a behavior synthesis model. Both subjects are shown a virtual representation of the other (see Fig. 2), animated based on one of two sources: (1) directly on the observed behavior of the other, or (2) on the output of the synthesis model. Both sources share the same behavior animation capabilities and limitations. We can therefore analyze the effect of the quantity, type and timing of the nonverbal behaviors on the perception thereof. During a conversation, the source of animation of the representation of each subject switches occasionally.

The idea behind the framework is that, when the displayed behavior deviates from what is typically regarded as human-like, the observer should notice. In this case, he or she is instructed to press a button (the *yuck* button [10]). The ratings can be used to evaluate and improve the behavior synthesis models (e.g. [14]). As observations of the subjects are continuously recorded, the framework doubles as a tool for study into nonverbal behavior.

* This publication was supported by the Dutch national program COMMIT.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 486–488, 2012.

2 Switching Wizard of Oz

In the Switching Wizard of Oz (SWOZ) setting, two human subjects A and B, seated at distributed locations, are shown virtual representations of each other. The representation of B displays either the behavior performed by B, or behavior synthesized by an algorithm, based on audio or video observations of A. The behaviors displayed by the virtual representations can be discrete (e.g. nods) or continuous (e.g. head movement). During a conversation, the source of a virtual representation is switched at random time intervals. To evaluate the quality of behavior synthesis models, both subjects are presented with a yuck button which they press whenever they believe the displayed behavior does not originate from the other subject.

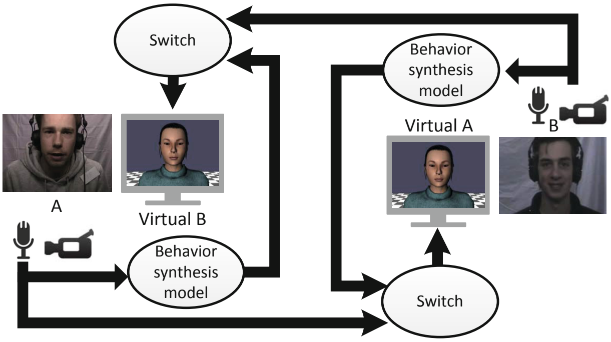


Fig. 1. Schematic representation of the Switching Wizard of Oz framework

Subject observation. The conversational partners are observed via sensors such as cameras, microphones, Kinects and gaze trackers. The observations are encoded into features in real-time. It should also be possible to regenerate the observed behavior on the virtual representation of the subject.

Behavior synthesis. These extracted features are subsequently used in a behavior synthesis algorithm, to determine whether or not certain behaviors should be animated. These algorithms can be manually engineered sets of classification rules (e.g. [3,5]) or machine learning classifiers trained on previously recorded corpus data (e.g. [6]). Based on the outcome of the algorithm or the observations of the actual conversational partner, the behavior is animated on a virtual agent. Behaviors can be verbal and nonverbal, discrete and continuous.

Behavior switching. The framework switches between the two sources at random time intervals. The displayed behavior should be continuous. For discrete events, this implies that the currently animated behavior should be finished and a new behavior should not be directly animated. For continuous behaviors, it should also be ensured that the displayed behavior is continuous so the switching moment will not be perceived as such to the observer. As the switching component of the framework is presented with the behavior of both the conversational partner and the algorithm, the switching time can be selected when the two sources are more or less similar, to allow for interpolation between the two.

References

1. Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., Schröder, M.: Generating Listening Behaviour. In: Emotion-Oriented Systems Cognitive Technologies - Part 4, pp. 321–347. Springer (2011)
2. Martin, J.C., Paggio, P., Kuehnlein, P., Stiefelhagen, R., Pianesi, F.: Introduction to the special issue on multimodal corpora for modeling human multimodal behavior. Language Resources and Evaluation 42(2), 253–264 (2008)
3. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 32(8), 1177–1207 (2000)
4. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: Proceedings of the Conference of the European chapter of the Association for Computational Linguistics, Budapest, Hungary, vol. 1, pp. 51–58 (April 2003)
5. Truong, K.P., Poppe, R., Heylen, D.: A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of Interspeech, Makuhari, Japan, pp. 490–493 (2010)
6. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems 20(1), 80–84 (2010)
7. McDonnell, R., Ennis, C., Dobbyn, S., O'Sullivan, C.: Talking bodies: Sensitivity to desynchronization of conversations. ACM Transactions on Applied Perception 6(4), A22 (2009)
8. Hodgins, J., Jörg, S., O'Sullivan, C., Park, S.I., Mahler, M.: The saliency of anomalies in animated human characters. ACM Transactions on Applied Perception 7(4), A22 (2010)
9. Poppe, R., Truong, K.P., Reidsma, D., Heylen, D.: Backchannel strategies for artificial listeners. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 146–158. Springer, Heidelberg (2010)
10. Poppe, R., Truong, K.P., Heylen, D.: Backchannels: Quantity, Type and Timing Matters. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 228–239. Springer, Heidelberg (2011)
11. Bailenson, J.N., Yee, N., Patel, K., Beall, A.C.: Detecting digital chameleons. Computers in Human Behavior 24(1), 66–87 (2008)
12. Edlund, J., Beskow, J.: Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. Language and Speech 52(2-3), 351–367 (2009)
13. Huang, L., Morency, L.-P., Gratch, J.: Virtual Rapport 2.0. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 68–79. Springer, Heidelberg (2011)
14. de Kok, I., Poppe, R., Heylen, D.: Iterative perceptual learning for social behavior synthesis. Technical Report TR-CTIT-12-01, University of Twente (2012)

Modeling the Multi-modal Behaviors of a Virtual Instructor in Tutoring Ballroom Dance

Hung-Hsuan Huang, Yuki Seki, Masaki Uejo, Joo-Ho Lee, and Kyoji Kawagoe

College of Information Science & Engineering, Ritsumeikan University, Japan
hhhuang@acm.org

1 Introduction

In learning sophisticated physical skills such as sport, gymnastics, or dance, it requires the learner to practice repeatedly for long period. Embodied conversational agents (ECAs) are thus suitable candidates for such tasks. The agent (virtual instructor) can be available at anytime and suffers much less constraints in the location where to set up. The learner can thus practice the tasks for unlimited times at their favorite place at their favorite time. If the knowledge and the interaction scenario of the agent are well designed, the lessons at acceptable level can be guaranteed constantly. Unlike human instructors, they never feel tired caused by long-time service and never loose patience on awkward learners. They also do not need to mind requesting the instructor to repeat the exemplary motion and never feel embarrassed to practice when they are still unskillful. On the other hand, ballroom dance is becoming a world-wide popular sport. The motion of ballroom dance involve the whole body of the dancers, head, torso, arm, and legs have to move simultaneously and synchronized with each other and music. Ballroom dance motion are so sophisticated and require man-to-man instruction and massive repeated practice for long time. This characteristics of ballroom dance make it an appropriate candidate for research on virtual instructors.

This paper presents an ongoing project aiming to develop such a virtual ballroom dance instructor. For building a believable virtual instructor, the first step is to know how a human instructor does. Therefore this project starts with a human-human tutoring experiment. The behaviors of the agent is then designed base on the analysis of the collected corpus.

2 Corpus Collecting Experiment

In order to build a believable virtual instructor, our first step is to investigate how a human professional instructor tutors the learner. Also, since the 2D virtual instructor is limited in the screen and can not touch the learner which is considered as quite normal behaviors in teaching physical skills. A human-human tutoring experiment is therefore conducted to collect the dance tutoring activities corpus in a simulated situation. Six male students of the dancesport club of our university are recruited as the learner subjects. Four of them have one-year experience in learning ballroom dance and two of them have two-year experience. The instructor and the learners are separated to two rooms and can not see each other directly. These two rooms are then connected by the

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 489–491, 2012.

telecommunication software, Skype so that the subjects can see and hear each other from remote. The learner's dance motion is recorded by NaturalPoint OptiTrack optical motion capture (captures at 100 fps) system for further analysis. They are instructed to dance a sequence of basic rumba steps which contains six counts. The interaction between the instructor and the learner is controlled to end around 30 minutes.

The corpus collected in the experiment is then labeled by two coders who are knowledgeable with ballroom dance but are not directly involved in this project. They are instructed to label the verbal and non-verbal behaviors both of the instructor and the learner subjects. 21 instructor and 17 learner predefined behaviors are annotation, and the annotation tool, Anvil [1]. The inter-coder agreement between two coders is generally above 90% and can be considered as reliable. The annotation done by the coder who has higher expertise in ballroom dance is adopted. The learners seldom spoke during the tutoring interaction. Most of them were the acknowledgments to the instructor's instruction, the requests for the repetition or the confirmation of what the instructor just said or demonstrated. From the observation during the experiment and the interview with the subjects, this may be caused by the relatively strict social relationship between the instructor and the learners of the ballroom dance community in Japan.

3 State Transition Model of the Virtual Instructor's Behaviors

The most frequently performed actions are Explain(21.2%), Demo(13.4%), Praise (10.6%), Progress(10.4%), Request-dance(9.4%), Beat(8.5%), and Spec-demo(6.5%). According to the analysis results, we designed a state transition model of the instructor's tutoring behaviors.The interaction starts with the greeting from the virtual instructor, it at first makes an introduction of what it is going to teach, demonstrate how to perform it, and then request the learner to mimic its motion. While the learner is dancing, the instructor evaluates the learner's performance and make real-time feedbacks. If the dancer is dancing well, the instructor praises the learner. If the learner's performance is too bad, the instructor may interrupt the learner's dance and repeat the explanation or demonstration. The process is repeated as a loop until the learner's performance is improved to some satisfying degree, then the virtual instructor progress to next steps. The interaction ends when all specified steps reach a satisfying level.

4 Conclusion and Future Works

This paper presents a methodology for modeling a virtual ballroom dance instructor. The behaviors of the instructor is modeled based on empiric results of a human-human tutoring experiment. The first future work is to build a prototype system using this model. Second, it is also essential to evaluate whether the instruction model is natural and whether it is effective in improving ball room dance learners' skill. The third step is to realize more sophisticated behaviors / interactions such as synchronizing verbal explanation and instructing dance movements in variable speed, for example, in the situation to dance with the learner together.

5 Related Works

Using virtual characters for instruction task is not a brand new idea. Earlier system like [2] proposed by Chua et al. is a virtual reality training system of Tai Chi, the learner wears a head mounted display (HMD) where one or more virtual instructors are projected to. The authors tried five layouts of the positions of instructor/learner, e.g. one instructor standing in front of the learner, four instructors surrounding the learner, etc. In this system, there was no interactive instruction and feedbacks, and the virtual instructor did not behave like an instructor but only worked like a video clip allowing the learner to mimic. The learner's motion are motion captured and compared to template motion based on Euclidean distance. Nakamura et al. proposed [3], a dance training system using demonstration video instead of an agent projected on a screen . The authors compared learners' performance between a fixed screen and a moving screen synchronized to the dance motion. Chan et. al. [4] evaluated the effects of three different ways of feedback to the learner's performance in a dance training system: an avatar of highlighted limbs where the learner did badly, a slow motion replay, and numeric scores. However, these present systems merely use CG characters for demonstrating exemplar motions but have no thought for actually utilizing the character as a "virtual instructor" teaching the learner like how a human instructor does. All of these systems do not have interactive instruction and the conversation between the virtual instructor and the learner. The virtual character, Steve [5] who teaches the procedure of operating a complex instrument is a pioneer work of virtual instructor for teaching a physical skill. However, the ballroom dance teaching task involves a much more complex and fast body motion than instrument operation.

References

1. Kipp, M.: Spatiotemporal coding in anvil. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008 (2008)
2. Chua, P.T., Crivella, R., Daly, B., Hu, N., Schaaf, R., Ventura, D., Camill, T., Hodgins, J., Pausch, R.: Trainning for physical tasks in virtual environments: Taichi. In: IEEE Virtual Reality, VR 2003 (2003)
3. Nakamura, A., Tabata, S., Ueda, T., Kiyofuji, S., Kuno, Y.: Dance training system with active vibro-devices and a mobile image display. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2005 (2005)
4. Chan, J.C., Leung, H., Tang, J.K., Komura, T.: A virtual reality dance training system using motion capture technology. IEEE Transactions on Learning Technologies PP(99) (2010)
5. Rickel, J., Johnson, W.L.: Steve: An animated pedagogical agent for procedural trainning in virtual environments. SIGART Bulletin 8, 16–21 (1998)

Hospital Buddy: A Persistent Emotional Support Companion Agent for Hospital Patients

Timothy Bickmore[1], Laila Bukhari[1], Laura Pfeifer Vardoulakis[1],
Michael Paasche-Orlow[2], and Christopher Shanahan[2]

[1] College of Computer and Information Science, Northeastern University, Boston, MA
[2] Boston Medical Center, Boston, MA
{bickmore,bukhari,laurap}@ccs.neu.edu,
{Michael.Paasche-Orlow,Christopher.Shanahan}@bmc.org

The hospital experience can be disempowering and disorientating. Patients are deprived of sleep deprivation and exposed to constant noise, frequent interruptions, an unfamiliar environment filled with changing health professionals and ancillary staff, as well as medications often fraught with physical or psychoactive side-effects. These conditions often lead to discomfort and anxiety, and commonly induce delirium (especially in older adults), a neuropsychiatric condition in patients that results in clinically significant cognitive and perceptional problems. Simultaneously, because patients are usually alone in their rooms until a medical intervention is required they often are bored and starved for personal attention.

To address these issues, we developed a computerized hospital companion agent designed to support a patient throughout their hospital stay. The Hospital Buddy talks using synthetic speech and animation to which the patient responds using a touch screen attached to an flexible articulated arm at the bedside. The agent chats with patients about their hospital experience - providing empathic feedback and emotional support - in addition to a range of topics that have medical and entertainment functions.

1 Design of the Hospital Buddy

The virtual agent interface used is described in [1]. The Hospital Buddy provides patients with a brief orientation dialogue followed by options for the top-level dialogue topics described below. Following the initial conversation, the agent walks off the screen until the patient beckons it again (*Can we talk again?*).

"Let me tell you what's been going on." This dialogue enables patients to discuss an event that just occurred to them in the hospital, such as: just waking up; just finishing a meal; just completing an interaction with a provider; just finished watching TV; family or friends just visited; or just had a procedure or test done. In each case, the agent would elicit how the patient felt about the event, and provided empathic feedback when warranted. In addition, following interactions with a provider, patients were prompted for a brief evaluation of the provider and the interaction.

"I want to tell you how I've been feeling." This dialogue enables patients to self-report different subjective health-related states—including pain and stress—and

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 492–495, 2012.

record them for later time-series display for their own use or to share with their providers. The agent also took these patient utterances as empathic opportunities to provide comfort when appropriate.

"Can we chat?" Finally, if the patient initiated this dialogue, the agent offered to tell them a story, selected from a list of 97 health-related stories, anecdotes, and jokes.

2 Pilot Acceptance Study

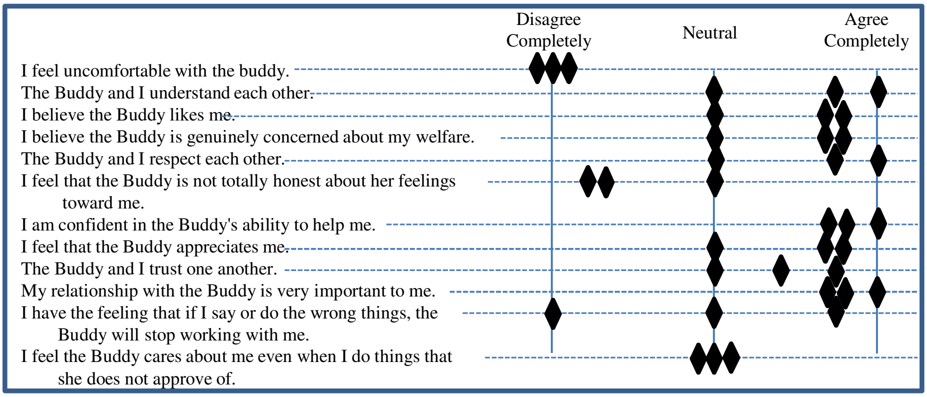
We conducted a preliminary pilot study to gage acceptance and use of the system by hospital patients when left in their room for 24 hours.

Participants. Three patients were recruited from a General Medicine floor of an urban hospital, aged 30-60, 66% female, with a range of medical conditions.

System Use. All patients used all system functions, averaging 17 interactions each with the agent during the 24 hours. All patients viewed their self-report pain levels, and one reported showing the chart to their doctor.

Quantitative Results on Relationship. Table 1 presents questionnaire responses. All patients reported feeling comfortable with the agent, were confident in its ability to help them, and felt that their relationship with the agent was important to them. Responses to other questions were generally positive, but mixed.

Table 1. Working Alliance Inventory Self-Report Scores for Each Patient



Qualitative Results. All patients reported that the system was very easy to use:

"I found it to be very easy… just by it being like a touch screen, you know it wasn't complicated at all. And with the questions, they're self-explanatory, so I didn't have a problem with it." (ID1).

When asked about how comfortable they were having the Hospital Buddy in their room, none of the patients reported any issues:

“I didn’t feel like somebody was watching me or anything like that. I like the buddy.” (ID2)

All 3 patients agreed that interacting with Hospital Buddy gave them something to do during their hospital stay in addition to keeping them company:

“[It] gave me something to do, other than just lay here.” (ID1)

“…I mean it kept me company, nothing was on TV…” (ID2)

System Functions. All 3 patients enjoyed the storytelling feature in Hospital Buddy. Two patients quoted some of the story parts during their interview, expressing their further interest in these stories.

All patients provided positive feedback on the reporting and event discussion functions as well:

“I thought how it kept the pain scale, and the schedule and the time, I thought that was awesome.” (ID2)

Companionship. All patients volunteered that the agent was effective at providing companionship during their hospital stay:

“I liked that, you know, she, you know, recognized my name and I like that she’s there to, you know, interact with…you know.” (ID1)

“The best thing about the system, like, you know, when you don’t have anyone here with you…it was actually nice to have her. I mean it kept me company.” (ID2)

“[It was] extremely comfortable, as a matter of fact, I relish it. I’m glad you came to me with this option, and have a chance to use it, to me, it helped me last night… the downtime, being lonely sometimes, this gives you something to do, something to hear.” (ID3)

Patient-Initiative. One patient volunteered that she appreciated the fact that the interactions were patient-initiated:

“It was nice, you know, to have the options to talk to her and she wasn’t bothersome, she wasn’t like: ‘talk to me! Talk to me!’ … she just kind of waited around until I talked to her. … Like I said, she was very warm and welcoming!” (ID2)

Suggestions for Improvement. Patients did think the agent could be more helpful if it had more medical capabilities and the ability to discuss their self-reported medical conditions in more detail:

“…when it was asking me how I was feeling, I wish it asked me whether or not I had a headache” (ID1).

“I wish it could answer some questions…medical questions, you know.” (ID3)

One patient mentioned that the Hospital Buddy could be used as a messaging system between them and their providers, so that they could get more sleep:

“I think it would have helped me to know that staff and doctors had knowledge of it, because then like…when I’m asleep, they could have easily come in and looked at whatever I had corresponded with the Buddy, so they didn’t have to bother me.”

3 Conclusions

Overall patient acceptance of and reaction to the Hospital Buddy was very positive, and it appeared to meet its primary objective of providing companionship. We are extending the Hospital Buddy with a suite of sensors—including acoustic sensors to detect medical device alarms, accelerometers to detect patient sleep/wake states, and long-range RFID sensors to detect and identify approaching providers— so that the agent is aware of events in the hospital room and can use these to initiate more intelligent and focused dialogue with patients.

Reference

1. 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)

Towards Assessing the Communication Responsiveness of People with Dementia

Yuko Nonaka[1], Yoichi Sakai[1], Kiyoshi Yasuda[2], and Yukiko Nakano[3]

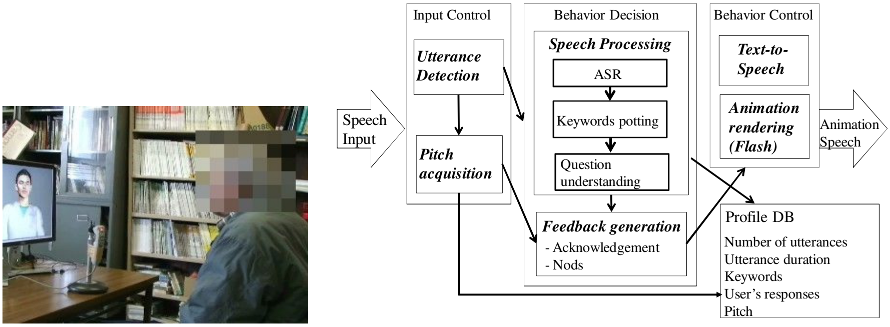
[1] Graduate School of Science and Technology, Seikei University, Japan
{dm126214,dm116217}@cc.seikei.ac.jp
[2] Chiba Rosai Hospital
fwkk5911@mb.infoweb.ne.jp
[3] Dept. of Computer and Information Science, Seikei University
y.nakano@st.seikei.ac.jp

1 Introduction

The number of elderly people is increasing. Some of them live alone, need physical help, and more seriously have cognitive impairment such as Dementia symptoms. To help such people, assistive technologies need to be developed. Pollak [2] proposed three assistive functions for elderly people with cognitive impairment, and assessing the elder's cognitive status is one of the issues she proposed. If the agent system can assess the patient's cognitive status through conversations, that will be a more natural way of measuring the cognitive status of the patient. Moreover, interpersonal communication is one of the most preferable daily activities for elderly people. As the first step towards assessment technologies, we developed a prototype of a listener agent, and collected conversations between the agent and people with Dementia. This paper reports the results of analyzing collected data for speech and head motion.

2 Listener Agent

Fig. 1 (a) shows a snapshot of communication between our listener agent and the user. To elicit responses from the user, the agent asks the user a set of questions one by



(a) Interaction snapshot (b) Architecture

Fig. 1. Listener agent

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 496–498, 2012.

one. We chose questions that are typically asked by doctors or nurses, such as inquiries regarding the patient's physical condition and meals, and more general topics, such as the patient's childhood memories and his/her locality. Fig. 1 (b) shows the system architecture of the agent. When the microphone input power exceeds a certain threshold, the pitch acquisition module calculates the pitch information for a given speech input. Moreover, some keywords are recognized using keyword spotting, utterance duration and the number of utterances are also measured and saved in the profile DB, which will be used in quantifying the communication responsiveness for a given interaction. This module also determines the agent's feedback actions based on previous studies in virtual agents [1] and analysis of Japanese corpus [3]. If there is no response from the user after a certain amount of time, the agent asks the next question and awaits the user's response.

3 Corpus Analysis and Conclusion

We analyzed behaviors of two male and eight female subjects with Dementia. By looking at the video data, two annotators categorized the user's response into two groups; high responsive (HR), and low responsive (LR). The judgments for 107 responses out of 162 were agreed between two annotators, and we only used the agreed data for the analysis. The average values for the following measures were compared between the two groups.

(1) Pause: Pause between the end of the agent's question and the start of the user's answer was analyzed. The average for HR was 0.98 sec and that for LR was 1.57 sec. The difference was statistically significant (t(95)=-2.82, p<0.05).

(2) Pitch: We calculated the average pitch for each user's response. The average of pitch for HR was 178.6 Hz and LR was 137.1 Hz (t(105)=2.99, p<0.01).

(3) Duration: Average utterance length was also compared. In HR, the average utterance duration was 4.77sec and that for LR was 1.60sec. The difference was statistically significant (t(105)=5.83, p<0.01).

(4) Head nod: We implemented a simple head nod recognition method by looking at head position and rotation in y-axis. In HR, average frequency of head nods per response was 0.61, and that for LR was 0.34. We found a statistical trend for this difference (t(105)=1.86, p<0.1).

These results indicate that shorter pause, higher pitch, longer utterance, and more frequent head nods correlate with higher responsiveness, and suggest that these parameters are useful in assessing the responsiveness in conversation. We also proposed a method for giving responsiveness scores (1-4) for each response by integrating the speech and head motion data. This is the first step towards assessing cognitive status through interaction with agents. We need to improve our agent as a listener, and propose more accurate and reliable scoring method for responsiveness.

References

1. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S.C., Morales, M., van der Werf, R.J., Morency, L.-P.: Virtual Rapport. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P., et al. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 14–27. Springer, Heidelberg (2006)
2. Pollack, M.: Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairments. AI Magazine, 9–24 (2005)
3. Tsukahara, W., Ward, N.: Responding to Subtle, Fleeting Changes in the User's Internal State. In: CHI, 2001. ACM (2001)

A Conversational Agent for Social Support: Validation of Supportive Dialogue Sequences

Janneke M. van der Zwaan, Virginia Dignum, and Catholijn M. Jonker

Delft University of Technology

Abstract. Recently, we proposed a dialogue model for social support. To validate this model, we analyzed 23 real world chat conversations. After some adjustments, the dialogue sequence patterns specified in the model cover 87.4% of the data. Based on this result, we conclude that the dialogue model accurately describes comforting conversations. Next, the model will be incorporated into a comforting ECA.

1 Introduction

Social support or comforting refers to all communicative attempts to alleviate the emotional distress of another person [2]. In our research, we are exploring how and to what extent Embodied Conversational Agents (ECAs) can provide social support. Recently, we proposed a domain-independent dialogue model to provide social support in response to upsetting events such as bullying [4]. In particular, the model specifies dialogue sequences that allows an ECA to verbally express social support [3]. So far, it was unclear to what extent these sequences occur in actual comforting dialogues. To assess the validity of our model, we analyzed dialogue sequences in real comforting conversations about bullying.

2 Dialogue Model for Social Support

In the dialogue model, a comforting conversation consists of 5 phases (cf. the 5-phase model [1]): 1) Welcome; 2) Determine the user's situation; 3) Determine conversation objective (e.g., getting tips on how to deal with bullying); 4) Give advice; and 5) Round up. For communicating social support, phase 2 (Determine situation) and 4 (Give advice) are the most important phases. Therefore, only the sequences in these phases have been analyzed.

Conversation phases consist of one or more dialogue sequences. A dialogue sequence is a set of utterances (conversation turns) in which a request for information or the proactive sharing of a piece of information is completed by the dialogue partners. Phase 2 consists of a recurring pattern of the agent asking a question, the user answering that question and the agent acknowledging the answer. Optionally, the agent expresses support by giving a sympathetic remark, compliment or encouragement to the user (pattern QA). In phase 4, the agent asks the user his plans on how to deal with the situation; this topic is discussed

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 499–501, 2012.

using sequence pattern QA. Additionally, the agent proactively utters advice and the user confirms that advice (pattern Advice). Optionally, a piece of advice is followed by a list of instructions (pattern Teaching). More details about the sequence patterns can be found in [3].

3 Data Analysis

Twenty-three real world chat conversations about bullying were analyzed. The data was anonymized and consisted of the counselor's utterances and the *positions* of the user's utterances. After dividing the conversations into the five phases, phases 2 and 4 were divided into sequences of utterances. Next, patterns occurring in the sequences were extracted. We started with the sequences as specified in the dialogue model and adjusted or added patterns when needed.

Because the user utterances were unavailable in the data, the contents of a user turn are determined based on the response of the counselor. User utterances were included patterns only if the counselor explicitly responded to them.

4 Results

Analysis of the sequences in the corpus suggested that the sequence specifications were too strict. Therefore, two out of three sequences specified in the model have been adjusted to better fit the data. In total, the data contained 10 different sequence patterns. After adjustment, the patterns specified by the model cover 87.4% of the sequences. Due to space constraints, we only discuss how the sequence patterns were adjusted and present the most frequent new pattern.

In the QA pattern, the agent's acknowledgment was made optional, because counselors do not always respond to a user's answer. Additionally, users did not seem to confirm the counselor's advice, so the user confirmation in the Advice sequence was made optional as well. While the model specifies a direct way of giving advice (the counselor/agent telling the user what to do), the data contains more indirect styles for giving advice. After giving advice the counselor requests feedback, e.g. *'You can talk to a teacher. How about that?'* (C46) and optionally expresses support after a user responds to a feedback request.

The most important new sequence pattern found in the data can be characterized as the counselor responding to information the user proactively introduces during the conversation. This pattern RtU (Respond to User) accounts for 5.6% of the sequences that occur in phase 2 and 4.

5 Conclusion

Our analysis shows that there are many regularities in the data. While we had to adjust the proposed sequence patterns and found new patterns in the data, we conclude that the regularities in the data are captured to a large extent by our dialogue model. This proves the suitability of the model. Now the dialogue model is validated, it can be incorporated into a comforting ECA.

Acknowledgements. This work is funded by NWO under the Responsible Innovation (RI) program.

References

1. de Beyn, A.: In: gesprek met kinderen:de methodiek van de kindertelefoon. In: SWP (2003)
2. 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)
3. van der Zwaan, J.M., Dignum, V., Jonker, C.M.: A bdi dialogue agent for social support: Specification and evaluation method. In: Proceedings of the 3rd Workshop on Emotional and Empathic Agents @ AAMAS 2012 (2012)
4. van der Zwaan, J.M., Dignum, V., Jonker, C.M.: A conversation model enabling intelligent agents to give emotional support. In: Proceedings of the 25th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2012), Dalian, China (2012)

Rich Computational Model of Conflict for Virtual Characters

Reid Swanson and Arnav Jhala

Computational Cinematics Studio,
Center for Games and Playable Media,
University of California, Santa Cruz, CA - 95064
{reid,jhala}@soe.ucsc.edu

Abstract. Rich interactions with virtual characters in narrative-based environments can be enabled by providing characters with representation of parameters for reasoning about various types of conflict. This paper proposes a model of conflict that includes mechanics, context, and dynamics of conflict scenarios. This model extends and reconciles prior work on conflict management from various disciplines. This model complements task-oriented conflicts that are implemented in current agent architectures and seeks to motivate exploration to a new design space of possible conflict situations. This work is based on initial analysis of a corpus of conflict scenarios annotated with personality profiles and resolution strategies. This annotated corpus is made available to the community for further research on conflict.

Conflict is an essential part of our everyday interaction. However, a vast majority of intentionally designed conflict in virtual environments, such as video games, is task oriented, which is only one aspect of a rich set of interactions that are possible. In this paper, we describe a model of conflict that addresses several of these aspects that, if represented in virtual characters, will enable new interactive experiences.

Recently there has been increasing interest in games that feature rich social interactions as core gameplay mechanics, particularly educational games for teaching conflict resolution skills. These games can become powerful tools that enable players to explore a wide range of social interactions within virtual environments. There are many different ways in which conflict resolution is learned, including peer mediation, films, and role-playing sessions. Most of the traditional pedagogical strategies that teach about conflict are non-interactive, leading to fixed examples that are communicated to students as *narratives in third-person* with a fixed point-of-view. [1]

Virtual characters are a good sandboxes for interactively exploring various types of conflict scenarios [2]. They provide players with the vocabulary for talking about conflict and can show them the consequences of their decisions in a controlled environment without real-life repercussions. Virtual characters can also provide feedback that teaches students to choose resolution strategies more rationally through better expectations of the possible outcomes. For researchers, virtual environments also provide an experimental tool to test relevant theoretical hypotheses.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 502–504, 2012.

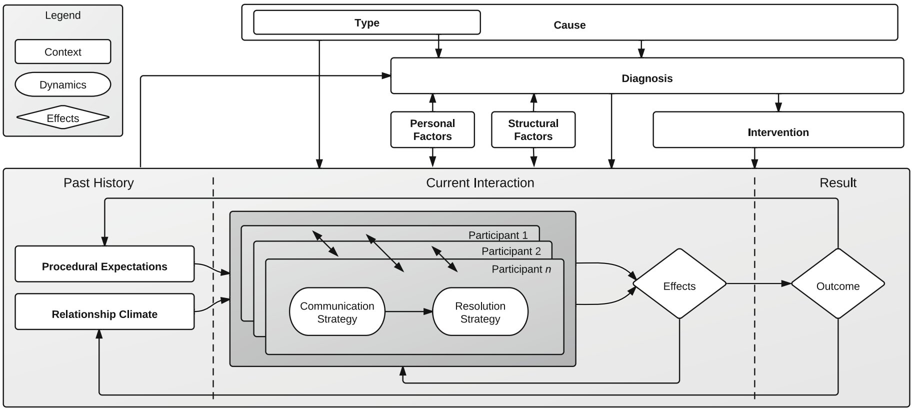


Fig. 1. A model of inter-personal and group conflict

A computational process model of conflict (Figure 1) is proposed that combines and extends prior work in psychology, sociology and game studies. A taxonomy of types of conflict that can be represented and reasoned over with this model, and their representation in terms of procedural elements and proceed to elaborate on the representation of causes, influences, and strategies is presented. This enables a study of types of conflict that can be explicitly designed into a game with a moderate level of control and predictability in how these affect gameplay and the overall experience of players. This is particularly useful in serious games that involve pedagogical goals that often conflict with or take cognitive load away from gameplay goals. Finally, to hint at the applicability of the model, we provide the design of a simple 4-turn game and analyze it with respect to our process model. Our game scenarios are based on data collected from crowd sourcing on Amazon's Mechanical Turk.

To hint at the applicability of this work, we describe a simple social game that illustrates how our model can be used specifically for game design. The setting of the game is a schoolyard where the children have just been let out of class for recess[1]. Unfortunately, several children distrust each other, spread rumors, have conflicting goals and have opposing values. The goal of the player is to interact with each of the children in order to find an optimal social solution that ensures the maximum number of children enjoy their recess before they have to back to class. Each interaction with an NPC is a series of short (e.g., four turn) minigames that are strung together to from a complete experience over the duration of the entire game. The mechanics of this game scenario follows the process model from Figure 1. At each turn of the game, virtual characters go through the process of *intent recognition*, *goal selection*, *prioritization of actions*, and *response* by utilizing a reasoning process that involves the various parameters described in this paper.

[1] `http://games.soe.ucsc.edu/project/siren` and `http://promweek.soe.ucsc.edu`

References

1. Cheong, Y., Khaled, R., Grappiolo, C., Campos, J., Martinho, C., Paiva, A., Yannakakis, G.: A computational approach towarlution for serious games. In: Foundations of Digital Games, Bordea, France (2011)
2. Smith, J.H.: The games economists play: Implications of economic the study of computer games. Game Studies 6 (2006)

A Model for Embodied Cognition in Autonomous Agents

Marco Vala, Tiago Ribeiro, and Ana Paiva

INESC-ID and Instituto Superior Técnico, Technical University of Lisbon,
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal
{marco.vala,ana.paiva}@inesc-id.pt,
tiago.ribeiro@gaips.inesc-id.pt

Abstract. The traditional efforts to mimic basic human behavior in embodied agents use an approach that draws a clear line between the "mind" and the "body" of those agents. However, recent findings in neuroscience show that our bodies have an active role in what we call "intelligence". We studied several processes related with our body and propose a model to integrate them in generic embodied agents. The model was validated in a small case study with the NAO robot.

Keywords: Embodied Agents; Body Model; Physiological Approach.

1 Motivation and Related Work

The concept of embodied agent has been widely explored in both robotics [7] and computer science [4]. The traditional computational approach for such agents uses a dualist perspective: a central "mind" receives sensory information and performs actions through the "body" (sensors and effectors) in a continuous sense-reason-act loop. However, using a centralized decision-making process has some implications. The mind has to cope with different levels of control and abstraction at the same time, which range from lower-level sensors and effectors to higher-level cognitive tasks.

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 the higher-level cognitive tasks. And although one may argue that most physiological processes are "hidden" and will have a limited impact on embodied agents, we think that they have an important role in the the generation of subconscious behavior which, to some extent, shapes the conscious mind [6].

Some previous work that explored physiological architectures includes an hormonal approach to model motivations and emotions in behavior selection [3], and an action-selection mechanism grounded in ethology [2]. More recently, Lim et al. [9] developed an approach to add physiological aspects to agents with high-level emotional planning and storytelling capabilities. We follow the latter work and propose a model of embodied cognition to be used in generic agents with different forms of embodiment.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 505–507, 2012.

2 Embodied Cognition

The main idea behind Embodied Cognition is to enrich the aforementioned sense-reason-act loop with an explicit model of the body. The model defines: a physiological space [1], which represents the current state and condition of the body; a set of *internal sensors*, which monitor the body's physiological condition (interoception [5]) and gather feedback from the effectors (proprioception [8]); a set of *internal effectors*, which can execute changes within the body; and an implicit memory [10], which stores procedural memories (sequences of actions which are executed in certain conditions). These mechanisms are part of a secondary control loop, a "subconscious mind", that runs in parallel with the "conscious mind" (Figure 1).

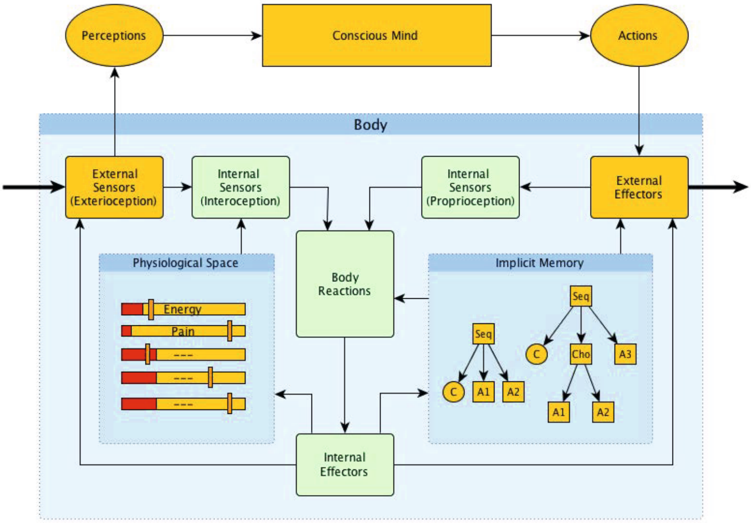


Fig. 1. Generic Architecture of an Agent with the Model of Embodied Cognition

The model was implemented in a small case study with the NAO robot. The "subconscious mind" creates a parallel execution layer that frees the "conscious mind" from body-related tasks, like background behaviors or instinctive reactions. The "conscious" layer is in control, but the "subconscious" can always step up to cope with unbalanced situations in the body. The "subconscious" also filters sensory data and adapts motor commands, thus creating an indirection that fosters the definition of generic bodily behaviors to be shared across different types of embodiment (both robotic and virtual). Therefore, the next step will be to explore the reusability of the model in different bodies. We believe it will support a faster development of complex behavior in embodied agents as well as richer interaction possibilities.

Acknowledgements. This work was supported by the EU FP7 ICT-215554 project LIREC (LIving with Robots and intEractive Companions). This research was supported by EU 7th Framework Program (FP7/2007-2013) under grant agreement no. 215554, and FCT (INESC-ID multiannual funding) through the PIDDAC Program funds.

References

[1] Bernard, C.: Lectures on the phenomena of life common to animals and plants. Lectures on the Phenomena of Life Common to Animals and Plants, vol. 1, Thomas (1974)

[2] Blumberg, B.M.: Old Tricks, New Dogs: Ethology and Interactive Creatures. Ph.D. thesis, Massachusetts Institute of Technology (1997)

[3] Canamero, D.: A hormonal model of emotions for behavior control. VUB AILab Memo 2006 (1997)

[4] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in conversational interfaces: Rea. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: The CHI is the Limit, CHI 1999 pp. 520–527. ACM, New York (1999)

[5] Craig, A.D.: How do you feel? interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience 3(8), 655–666 (2002)

[6] Damasio, A.: Self Comes to Mind: Constructing the Conscious Brain. Random House (2011)

[7] 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)

[8] Jones, L.A.: Motor illusions: What do they reveal about proprioception? Psychological Bulletin 103(1), 72–86 (1988)

[9] Lim, M.Y., Dias, J., Aylett, R., Paiva, A.: Intelligent NPCs for Educational Role Play Game. In: Dignum, F., Bradshaw, J., Silverman, B., van Doesburg, W. (eds.) Agents for Games and Simulations. LNCS, vol. 5920, pp. 107–118. Springer, Heidelberg (2009)

[10] Schacter, D.L.: Implicit memory: History and current status. Journal of Experimental Psychology. Learning Memory and Cognition 13(3), 501–518 (1987)

Evaluation of an Affective Model: COR-E

Sabrina Campano, Etienne de Sevin, Vincent Corruble, and Nicolas Sabouret

Laboratoire d'Informatique de Paris 6,
4, place Jussieu, 75005 Paris, France
sabrina.campano@lip6.fr

Abstract. In this paper, we present an evaluation of the affective model COR-E. This model intends to produce behaviors judged as emotional and believable ones. Emotions are seen as an emergent phenomenon, they are not encoded in the model. Our results show that COR-E effectively produce intended behaviors, thanks to its various characteristics.

Keywords: affect, emotion, believability, behavior, virtual agent.

1 Introduction

Most existing computational affective models rely on a number of numerical emotion variables that must be manually parametrized so as to outline believable affective responses and behaviors [5,3,2]. However, finding the correct value of these parameters and the influence of each one on the general model is a significant challenge. Other approaches, such as Pfeifer's work [6] or the MicroPsi model [1], aim at obtaining emotional behaviors without using emotion variables. Emotions are considered as an emergent phenomenon. The model COR-E presented in this paper enters this category of model. COR-E (COR-Engine) is based on the psychological theory of COnservation of Resources (COR) [4]. The central tenet of the theory is that people strive to obtain, retain, and protect resources. The concept of resource refers to many types of subjective items : social ones such as self-esteem or caring for others, material ones such as a car, or physiological ones such as energy. COR-E intends to keep an architecture with a small number of parameters, while allowing the simulation of a wide variety of affective behaviors, including social ones. COR-E is based on the principle that an agent tries to protect and acquire resources that it values when they are respectively threatened or desired. The general architecture of the model is shown in figure 1.

2 Evaluation

In order to evaluate COR-E, we recorded videos clips of agents simulated by the model, in the scenario of a waiting line. Then we asked some human participants to answer an online questionnaire about these videos. The evaluation had two main objectives: (i) to determine whether agents' behaviors simulated by

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 508–510, 2012.

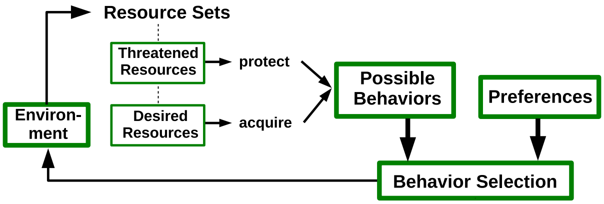


Fig. 1. General Architecture of COR-E - Resources Sets determine possible behaviors ; behavior selection depends on possible behaviors and agent's preferences ; environment is updated with behavior's effects ; resource sets are updated according to the environment.

COR-E are considered as believable and emotional by human observers (general hypothesis $H1$); (ii) to validate the impact of the main characteristics of COR-E's architecture: acquisitive and protective behaviors ($H2$), preferences ($H3$), and the use of *reputation* psychological resources ($H4$).

113 participants contributed to this study. According to our results, the hypotheses about COR-E were all validated. COR-E allows the simulation of believable emotional behaviors (general hypothesis $H1$) thanks to its characteristics ($H2$, $H3$, $H4$). Acquisitive and protective behaviors, preferences, and the psychological resource of "Reputation" seem necessary to produce such behaviors.

A large majority of participants recognized the behaviors produced by COR-E as related to agents' emotions (71.68 % to 92.04 %). These good results may be due in part to the use of the textual indication "protest", used by an agent in order to react to an intrusion in the waiting line. Indeed, this term can be psychologically associated with the emotion of anger, thus facilitating the recognition of that emotion among participants. It would be interesting to know if the same results will be obtained without textual indications.

Behaviors produced by COR-E model were rated as believable (mean from 5.08 to 6.01 on a scale from 1 to 7). It is possible that the score on believability was lowered because of a bias related to the interpretation of the term "believable". As emotional behaviors tend to occur rarely, they might be judged as less believable. Another possible bias that could have lowered the score on believability may be related to agents' moves. These elements indicate that the believability of agents' behaviors could be further improved.

3 Conclusions and Future Work

We presented in this paper an evaluation aimed at assessing whether the behaviors produced by COR-E are judged as believable and emotional. In a further evaluation, we plan to assess whether human observers recognize appropriate emotional states in agents according to a given context, and also whether they recognize agents' intentions. This will allow to check whether the internal state of an agent is well understood, according to the observation of its current behavior in the environment. This is an essential factor in agent believability [7].

References

1. Dörner, D., Gerdes, J., Mayer, M., Misra, S.: A simulation of cognitive and emotional effects of overcrowding. In: Proceedings of the Seventh International Conference on Cognitive Modeling, pp. 92–98 (2006)
2. Elliott, C.D.: The Affective Reasoner: A process model of emotions in a multi-agent system (1992)
3. Gebhard, P.: ALMA: a layered model of affect. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 29–36. ACM (2005)
4. Hobfoll, S.E.: Conservation of resources. American Psychologist 44(3), 513–524 (1989)
5. Marsella, S.C., Gratch, J.: EMA: A process model of appraisal dynamics. Cognitive Systems Research 10(1), 70–90 (2009)
6. Pfeifer, R.: The"Fungus Eater"Approach to Emotion: A View from Artificial Intelligence. Cognitive Studies 1, 42–57 (1994)
7. Sengers, P.: Designing comprehensible agents. In: International Joint Conference on Artificial Intelligence, vol. 16, pp. 1227–1232. Lawrence Erlbaum Associates Ltd. (1999)

Evaluation of the Uncanny Valley in CG Characters

Vanderson Dill, Laura Mattos Flach, Rafael Hocevar, Christian Lykawka,
Soraia R. Musse, and Márcio Sarroglia Pinho

Faculdade de Informática,
Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
vanderson.dill@acad.pucrs.br, soraia.musse@pucrs.br

Abstract. This article revisits the uncanny valley subject in order to evaluate its effects on people's perception of Computed Graphics (CG) characters from movies and games. We analyzed the ”uncanny” rates given by the users and as result we obtained a graph quite similar to the original on proposed by Mori [1].

1 Introduction

This paper studies the effects of the uncanny valley caused by CG characters, first presented by Mori in 1970 [1] for robots. According to his work, robots should not be made too similar to real humans because such robots can fall into the ”uncanny valley”, where too high degree of human realism evokes an unpleasant impression in the viewer. The increasing exposure of virtual characters to the general public motivated our group to test the original hypothesis using more up-to-dated samples from movies and games.

Several studies have tested Mori's theory, like Hanson [2], MacDorman [3,4], Seyama & Nagayama [5] and McDonnell & Breidt [6]. MacDorman *et al.* [7], explores the perception of different features in the face of CG characters. Our study is an evaluation on how people perceive CG characters that are present in modern digital media. We try to answer: Does the uncanny valley exists in this type of CG characters? Moreover, does adding movement to this characters changes the shape of the uncanny valley curve, like Mori [1] suggested?

To answer these questions, we proposed an evaluation methodology based on a questionnaire with subjects. First, we selected ten representative characters from recent and well known movies and games. In order to evaluate the public's empathy with the chosen characters, we conducted a two-staged questionnaire. In the first stage, we presented still images and in the second, videos with the performance of the same characters in the same scenes.

2 Obtained Results

The obtained data from the questionnaires is shown in Figure 1. It follows the original chart proposed by Mori [1], where the horizontal axis represents the human likeness and the vertical axis indicates the familiarity level. Figure 1 illustrates two curves plotted in the chart, one for still images and one for videos. It is noticeable that samples C.11, C.16, C.1 and C.14 suggest the same pattern of a negative emotional response as in result of an ”almost human but unnatural” character representation.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 511–513, 2012.

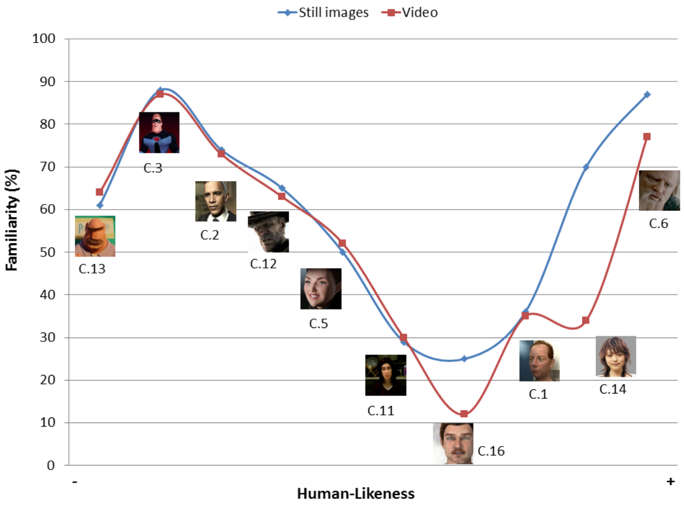


Fig. 1. Chart displaying the curves obtained from the collected data

The uncanny valley hypothesis investigates the effects of moving versus still characters. In Figure 1, even with a limited number of characters, it was possible to detect a similar behavior to the original chart [1], more noticeable in characters C.16 and C.14 that exaggerated the valley in the video's curve. The analysis of both questionnaires shows that the curves of still and moving characters are coherent to the original.

The questionnaire also asked to the participants which parts of the face the participant felt more discomfort, if it was the case. Analysing the answers, we could highlight two specific parts of the face: eyes and mouth.

3 Final Considerations

It was possible to validate the uncanny valley's persistence using a contemporary scenario: CG characters and animations are a rich environment to observe the phenomena. Future works may be able to revisit the uncanny valley subject under a similar methodology, in order to distinguish better the effects of the technological evolution over the population against the human instincts. This study didn't distinguish between gender, age, neither familiarity with animated movies nor computer games, which is also an interesting direction to be evaluated more deeply.

References

1. Mori, M.: Bukimi no tani (the uncanny valley). Energy, 33–35 (1970)
2. Hanson, D.: Exploring the aesthetic range for humanoid robots. In: Proceedings of the CogSci Workshop Towards Social Mechanisms of Android Science, pp. 206–216 (2006)

3. MacDorman, K.F.: Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. In: ICCSCogSci 2006 Long Symposium Toward Social Mechanisms of Android Science, Vancouver, pp. 26–29 (2006)
4. MacDorman, K.F., Ishiguro, H.: The uncanny advantage of using androids in cognitive and social science research. Interaction Studies 7(3), 297–337 (2006)
5. Seyama, J., Nagayama, R.S.: The uncanny valley: Effect of realism on the impression of artificial human faces. In: Presence: Teleoperators and Virtual Environments, pp. 337–351 (2007)
6. McDonnell, R., Breidt, M.: Face reality: investigating the uncanny valley for virtual faces. In: ACM SIGGRAPH ASIA, Sketches, SA 2010, pp. 41:1–41:2. ACM, New York (2010)
7. MacDorman, K.F., Green, R.D., Ho, C.C., Koch, C.T.: Too real for comfort? uncanny responses to computer generated faces. Comput. Hum. Behav. 25, 695–710 (2009)

Full-Body Gesture Interaction with Improvisational Narrative Agents

Andreya Piplica, Christopher DeLeon, and Brian Magerko

Digital Media, Georgia Institute of Technology
apiplica@gmail.com, chris@deleonic.com, magerko@gatech.edu

Abstract. Interactive narrative research strives to allow humans and intelligent agents to *co-create* narratives in real-time as equal contributors. While intelligent agents can interact with humans through expressive embodied representations, typical interactive narrative interfaces provide humans no way to reciprocate with embodied communication of their own. This disparity in interaction capabilities has been informally discussed as the *human puppet problem*. Improvisational theatre (improv) provides a real-world analogue to embodied co-creative interactive narrative experiences. This paper presents an overview of a system for combining improvisational acting with full-body motions to support human-AI co-creation of interactive narratives. The human begins an improvised narrative with an AI improviser while an intelligent avatar mediates interaction and gives the human an embodied presence in the scene.

Keywords: improvisational or dramatic interaction, conversational and storytelling agents, postures and gestures.

1 Overview

This paper presents a system for combining improv acting with full-body motions to support the co-creation of interactive narratives with an AI agent and a human interactor. Our system translates a human interactor's movements into actions that an AI improviser can respond to within the context of an improvised story. We have constructed a framework for human interaction with an AI improviser for beginning an improvised narrative with interaction mediated through an intelligent user avatar [1]. This approach uses human gestural input to contribute part, but not all, of an intelligent avatar's behavior. While joint human-AI agents have been employed in interactive narrative [2, 3], this is the first system to do so with full-body gestures. Our framework is derived from the interactions involved in the improv games *Three Line Scene*, which focuses actors in establishing the initial elements of a story within three lines (e.g. what characters are in the scene, where they are, and what they are doing together), and *Moving Bodies*, which allows audience members to control the bodies of human improvisers while the improvisers interpret those inputs and perform discourse acts.

We have built a system based on our socio-cognitive studies of improvisational actors [4] where a human interactor and an AI can improvise a pantomimed three-line

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 514–516, 2012.

scene. An intelligent avatar uses motion data from a Microsoft Kinect sensor to represent the interactor's motions in the same virtual space as the AI improviser. This gives the human interactor an embodied presence in the scene and shows how the Kinect senses their motion. This feedback can help the user understand the avatar's interpretations and adjust their movements to accommodate the sensor's limitations. The avatar reasons about the human's intentions for the scene and creates its own mental model. This mental model is used to inform the avatar about potential discourse acts to select. While the intelligent avatar, inspired by *Moving Bodies*, does provide a means of joint AI / human control of a character, our current implementation omits dialogue in favor of specifically studying gestural interaction.

The user stands before a large screen – like a television or projector screen – where the AI improviser and the intelligent avatar are displayed on a stage. The user faces the screen while standing approximately four to ten feet away from a Microsoft Kinect below the screen (i.e. within the Kinect's sensor range). The intelligent avatar and the AI improviser are shown as two-dimensional animated characters on a virtual stage. These simplified visual characters map 3D motion data from the Kinect directly onto the characters' 2D animations. The system does not currently support animated facial expressions, though such animations may be supported in future iterations.

The user performs a motion to begin a three-line scene, such as putting one fist on top of the other and moving their hands from side to side. The Kinect sends the sensed motion data to the intelligent avatar and the AI improviser, who each interpret the motion as an action [1]. The avatar displays the user's motion while it reasons about what character and joint activity the user may be portraying, which in turn can inform select of discourse acts. The AI improviser then reasons about the user's character and joint activity, as well as its own character. The AI improviser and the intelligent avatar draw on the same reasoning processes to understand how the human interactor contributes to the scene. They utilize background knowledge about a specific domain to make inferences about the platform. They incorporate the human's motions into their reasoning based on joint position data from the Microsoft Kinect sensor. Additionally, the AI improviser reasons about how to contribute presentations to the scene with its own motions.

The inputted motion data from the Microsoft Kinect consists of 3D coordinates for joint and limb positions, which we convert to 2D and use to animate the on-screen characters. The 2D coordinates are evaluated as "signals," which are hand-authored sets of joint angles and positions. Signals can be *simple* (joint angles and positions at one time) or *temporal* (joint angles and positions varying across time). A neutral stance is defined as standing still with feet together and hands at one's sides. Actions are defined as sets of positive and negative signals. If the Kinect data satisfies all positive signals for an action, it becomes a candidate. However, if the data triggers any negative signals for that action, it is removed as a candidate. When the user's turn ends, the intelligent agent and AI improviser select an action from the candidate interpretations based on their mental model of the scene [1]. After the AI improviser interprets the user's motion and reasons about how it contributes to the scene, the AI improviser must select an action to present and a motion to present it with based on its own perception of the ambiguous knowledge communicated in the scene.

References

1. O'Neill, B., Piplica, A., Fuller, D., Magerko, B.: A Knowledge-Based Framework for the Collaborative Improvisation of Scene Introductions. In: the Proceedings of the 4th International Conference on Interactive Digital Storytelling, Vancouver, Canada (2011)
2. Hayes-Roth, B., Sincoff, E., Brownston, L., Huard, R., Lent, B.: Directed Improvisation. Technical Report KSL-94-61. Stanford University, Palo Alto, CA (1994)
3. Brisson, A., Dias, J., Paiva, A.: From Chinese Shadows to Interactive Shadows: Building a storytelling application with autonomous shadows. In: Proc. Workshop on Agent-Based Systems for Human Learning and Entertainment (ABSHLE), AAMAS 2007. ACM Press, New York (2007)
4. Magerko, B., Manzoul, W., Riedl, M., Baumer, A., Fuller, D., Luther, K., Pearce, C.: An Empirical Study of Cognition and Theatrical Improvisation. In: The Proceedings of ACM Conference on Creativity and Cognition, Berkeley, CA (2009)

Understanding How Well *You* Understood – Context-Sensitive Interpretation of Multimodal User Feedback

Hendrik Buschmeier and Stefan Kopp

Sociable Agents Group - CITEC and Faculty of Technology, Bielefeld University,
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme,skopp}@uni-bielefeld.de

1 Introduction

Human interlocutors continuously show behaviour indicative of their perception, understanding, acceptance and agreement of and with the other's utterances [1,4]. Such evidence can be provided in the form of verbal-vocal feedback signals, head gestures, facial expressions or gaze and often interacts with the current dialogue context. As feedback signals are able to express subtle differences in meaning, we hypothesise that they tend to reflect their producer's mental state quite accurately.

To be cooperative and human-like dialogue partners, virtual conversational agents should be able to interpret their user's evidence of understanding and to react appropriately to it by adapting to their needs [2]. We present a Bayesian network model for context-sensitive interpretation of listener feedback for such an 'attentive speaker agent', which takes the user's multimodal behaviour (verbal-vocal feedback, head-gestures, gaze) as well as its own utterance and knowledge of the dialogue domain into account to form a model of the user's mental state.

2 Bayesian Model of the Listener

In previous work [2], we adopted the concept of 'listener state' [5] for a model that an attentive speaker agent *attributes* to its user, i.e., a representation that emulates the user's listener state. Here we present an implementation of 'attributed listener state' (ALS) by modelling it probabilistically as a Bayesian network. This (1) allows us to manage the uncertainties inherent in the mapping between feedback signal and meaning; (2) gives us a natural and robust mechanism of interpreting feedback in its dialogue context; and (3) enables inference and learning within a well understood formalism.

The Bayesian network (Figure 1a) models the notions of contact, perception, understanding, acceptance and agreement with one random variable each, so that the values of C, P, U, AC and AG are to be interpreted in terms of degrees of belief. Assuming discrete variables for simplicity, strengths are modelled via their states: *low, medium* and *high*. The influences between ALS-variables are modelled after Allwood's hierarchy of feedback functions [1], e.g., if understanding is assumed, perception and contact can be assumed as well; a lack of perception, on the other hand, usually implies that understanding cannot be assumed either. Apart from influencing each other, the ALS-variables are influenced by the dialogue context and the user's multimodal feedback behaviour, which we model here, exemplarily, in the form of six influencing variables.

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 517–519, 2012.

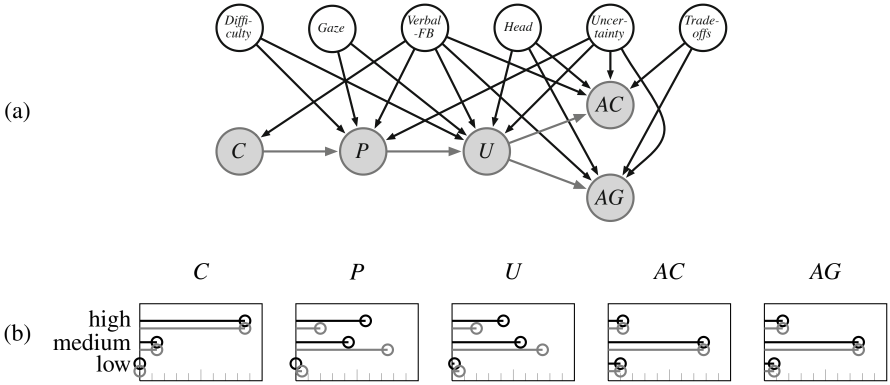


Figure 1. (a) Structure of the Bayesian model of the listener. The attributed listener state, drawn in shades of grey, consists of five random variables *C*, *P*, *U*, *AC* and *AG*. These are influenced by variables representing the dialogue context and the user's behaviour (drawn with black lines). (b) Plot of the example belief states. The *x*-axes show the probabilities of each variable's state. Black lines show the values for the first, grey lines values for the second variant of the example.

3 Worked Example

To demonstrate that the model behaves reasonably, we have tested its performance in an example situation in a calendar management domain. Figure 1b shows the belief state of the ALS-variables *C*, *P*, *U*, *AC* and *AG* under a certain assignment of (some of) the variables that represent the user's behaviour and the dialogue context. Model parameters (i.e., the conditional probability tables) were generated from structured representations [3]. In the example situation, the agent produces the utterance 'Your supervisor would like to meet you for lunch on Thursday at 12 PM', which is of *medium* difficulty as it might be a bit surprising but does not involve any complex structures or lexical items. The agent further knows that there will be no conflict with any other appointment of the user (*Trade-offs* = *low*). The user verbally signals understanding (*Verbal-FB* = *u*) but does not move her head (*Head* = *none*). In the first variant (black lines), she gazes at the correct target slot (*Gaze* = *on-target*), in the second variant (grey lines) at an incorrect target slot in the calendar (*Gaze* = *off-target*). As shown in Figure 1b, the belief state makes reasonable predictions of the mental state that the user might be in. The probability mass of the variables *P* and *U* is distributed mostly between *medium* and *high* in both variants. This makes sense, as the utterance was not too difficult and the user verbally signalled understanding. In the first variant, however, where the user looks at the correct target, the probability mass is shifted towards *high* for both variables *P* and *U*, whereas in the second variant, where the user is not looking at the correct target slot, the probability mass is shifted more towards *medium*. Since gazing at the correct target is not a strong indicator of acceptance or agreement, there is only a minimal difference between the two variants for the variables *AC* and *AG*.

More information on the model's causal structure and inner details along with its perfomance in further example situations is presented in [3].

Acknowledgements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence in Cognitive Interaction Technology (CITEC).

References

1. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9, 1–26 (1992)
2. Buschmeier, H., Kopp, S.: Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback. In: Vilhjálmsson, H.H., Kopp, S., Marsella, S., Thórisson, K.R. (eds.) IVA 2011. LNCS, vol. 6895, pp. 169–182. Springer, Heidelberg (2011)
3. Buschmeier, H., Kopp, S.: Using a Bayesian model of the listener to unveil the dialogue information state. In: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue, Paris, France (to appear)
4. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
5. 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)

Toward a Computational Model for the Automatic Generation of Character Personality in Interactive Narrative

Julio César Bahamón and R. Michael Young

North Carolina State University, Raleigh NC 27695, USA
jcbahamo@ncsu.edu, young@csc.ncsu.edu

Abstract. This paper introduces an approach for the incorporation of interesting and compelling characters in automatically generated interactive narrative. The approach is based on the development of a computational model that enables virtual characters to have distinct and well-defined personalities. In this model, character personality is founded on the hypothesis that choices that lead to actions can be used in interactive narrative to significantly influence a character's perceived personality.

Keywords: interactive narrative, planning, artificial intelligence

1 Introduction

In the area of Interactive Narrative (IN), the ability to generate character behavior that adjusts in response to user actions or changing story conditions has not been fully addressed by existing research. Models have been developed to direct character interactions [1], compose stories using predefined character models [2,3], generate dialog based on personality traits [4], and control facial and physical gestures to express emotion [5,6,7]; however, none of these approaches focus specifically on controlling character behavior over time to elicit the perception of a distinct and well-defined personality.

Previous approaches have primarily addressed character personality by considering behavior at a very fine-grained level of representation. That is, these approaches focus on a character's immediate reaction to stimuli. In contrast, our work focuses on the story as a whole and in particular the use of choices made by characters as a means to express their personality. We propose the creation of a model focused on physical actions and the role that these play over the course of a narrative in the construction of the mental model that the audience forms when experiencing it.

2 Character Personality Based on Choice

Considering narrative structure, specifically plot-points where branching occurs [8], we can intuitively expect the presence of actions that follow a choice. We

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 520–522, 2012.

posit that the link between action and choice can be harnessed in an interactive narrative system, such as a video game, to facilitate the expression of personality in virtual agents. We base this idea on research in behavioral psychology that has found correlation between people's actions and their personality traits [9,10]. Our hypothesis is that audiences can form opinions about a character's personality traits based on the choices that the character makes during the course of a story and the causal chain of events that contextualize such choices.

3 Modeling Choice in Planning-Based Interactive Narrative

Our model aims at enabling the representation of a general subset of personality traits with enough detail to elicit a predictable cognitive response from the audience. We represent traits using a taxonomy based on the Big Five personality structure as defined by Goldberg [11]. The process for action selection uses a declarative approach in which character attributes are considered in conjunction with the story context to choose actions that best represent a character's personality.

The implementation of the model extends a partial-order planning algorithm (e.g., POP [12]) to address narrative generation, similar to previous approaches developed by Young and his colleagues [13,14,15,16]. The key objective is to modify the process to ensure that choice is treated as a first-class object, i.e., the story structure and contents promote the existence of choices and make their existence evident to the audience. To accomplish this, actions are added to the plan after evaluating the context in which they are executed and examining the set of possible story plans. Context includes character and story attributes such as goals, beliefs, moral traits, relationships, story events, previous choices, and action effects. The result of the evaluation is a list of viable actions, ranked based on how closely they represent a character's personality traits. For example, an agreeable character is more likely to obtain money by working than by robbing a bank.

4 Conclusion

Results from this research will be applicable to systems used to create IN due to the reduction of authorial burden and increased creative freedom that may be provided. A narrative generation system based on our model could produce multiple different stories based on simple changes to character personality traits. For example, if the personality of the *hero* of a story is modified from conscientious to non-conscientious the resulting story could be markedly different but the work of the author would only involve changing one of the character's properties.

We are currently developing the algorithm for action selection and the process used to rank and place actions in the story plan. The model is expected to scale to complex domains and generalize to IN applications that include training simulations, activity visualizations, instructional video generation, and games.

References

1. Riedl, M.O., Stern, A.: Failing Believably: Toward Drama Management with Autonomous Actors in Interactive Narratives. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 195–206. Springer, Heidelberg (2006)
2. Mosher, R., Magerko, B.: Personality Templates and Social Hierarchies Using Stereotypes. In: Göbel, S., Malkewitz, R., Iurgel, I. (eds.) TIDSE 2006. LNCS, vol. 4326, pp. 207–218. Springer, Heidelberg (2006)
3. Lebowitz, M.: Creating characters in a story-telling universe. Poetics 13(3), 171–194 (1984)
4. Mairesse, F., Walker, M.: PERSONAGE: Personality generation for dialogue. In: Annual Meeting-Association For Computational Linguistics, pp. 496–503 (2007)
5. Doce, T., Dias, J., Prada, R., Paiva, A.: Creating individual agents through personality traits. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 257–264. Springer, Heidelberg (2010)
6. André, E., Klesen, M., Gebhard, P., Allen, S., Rist, T.: Integrating models of personality and emotions into lifelike characters. Affective interactions, 150–165 (2000)
7. Loyall, A.B.: Believable agents: building interactive personalities. PhD thesis, Carnegie Mellon University (1997)
8. Barthes, R., Duisit, L.: An Introduction to the Structural Analysis of Narrative. New Literary History 6(2), 237–272 (1975)
9. Mehl, M.R., Gosling, S.D., Pennebaker, J.W.: Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. Journal of Personality and Social Psychology 90(5), 862–877 (2006)
10. Funder, D.C., Sneed, C.D.: Behavioral manifestations of personality: An ecological approach to judgmental accuracy. Journal of Personality and Social Psychology 64(3), 479–490 (1993)
11. Goldberg, L.R.: An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology 59(6), 1216 (1990)
12. Weld, D.S.: An introduction to least commitment planning. AI Magazine 15(4), 27–61 (1994)
13. Young, R.M.: Notes on the use of plan structures in the creation of interactive plot. In: AAAI Fall Symp. on Narrative Intelligence, pp. 164–167 (1999)
14. Riedl, M.O., Young, R.M.: Character-focused narrative generation for execution in virtual worlds. In: Virtual Storytelling. Using Virtual Reality Technologies for Storytelling, pp. 47–56 (2003)
15. Riedl, M.O., Young, R.M.: Narrative Planning: Balancing Plot and Character. Journal of Artificial Intelligence Research, 164–167 (2010)
16. Harris, J., Young, R.M.: Proactive Mediation in Plan-Based Narrative Environments. In: Panayiotopoulos, T., Gratch, J., Aylett, R.S., Ballin, D., Olivier, P., Rist, T. (eds.) IVA 2005. LNCS (LNAI), vol. 3661, pp. 292–304. Springer, Heidelberg (2005)

Efficient Cultural Models of Verbal Behavior for Communicative Agents

Alicia Sagae[1], Jerry R. Hobbs[2], Suzanne Wertheim[1], Michael H. Agar[3],
Emily Ho[1], and W. Lewis Johnson[1]

[1] Alelo Inc
Los Angeles, CA, USA
{asagae,swertheim,eho,ljohnson}@alelo.com
[2] USC Information Sciences Institute,
Los Angeles, CA, USA
hobbs@isi.edu
[3] Ethknoworks,
Santa Fe, NM, USA
magar@umd.edu

Abstract. This paper presents compositional models of culture for conversational agents that are embedded in a training system for cross-cultural competency. Our models are implemented as ontologies of statements in common logic, with culture-specific and culture-general components. We compare the compositional framework to a finite-state system, in terms of development effort, number of reused and new objects, and flexibility and accuracy of resulting conversational simulations.

Keywords: Training & Simulation, Conversational Agents, Common Logic.

1 Background

This paper addresses modeling needs for conversational agents that appear as dialog partners for learners in a widely-used line of language and culture training systems [1]. In such a system, fine-grained control over agent behavior can be achieved using custom-developed scripts where every utterance performed by the agent has been authored for a particular context: one turn in one dialog, where the language, culture, and task at hand are all fixed. However, this approach is non-compositional, in the sense that the cultural knowledge from a given script cannot be extracted and reused in a new script with a different language and task. In contrast, stand-alone models of culture can be trained on samples of behavior, either from members of the community being modeled [2] or from culture-general samples [3]. However this prevents system designers from guaranteeing the appearance or timing of any given behavior, which is a disadvantage in a pedagogical context. In the current paper, we address these issues with a hierarchical rule-driven model of culture. Our system produces flexible, model-driven dialogs with conversational agents that use both culture-general and culture-specific rules. We achieve the novel capability to swap cultural models, in the form of

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 523–525, 2012.

rule sets, in and out of a social simulation to reveal pedagogically relevant differences at the level of verbal behavior (utterances, gestures) and intention (communicative act), captured in *dialogs*. In this paper we evaluate the gains in efficiency that our new architecture provides.

This work builds on related research on models of personality and emotion [4], agent goals [5], and task-based models of conversation [1]. We apply techniques from these systems to the task of generating culturally appropriate verbal behavior.

2 Experiments

The baseline architecture we use for agent conversations is based on finite state automata (FSAs) that encode conversational branches at the level of communicative act [6]. The experimental modular architecture encodes agent behavior in a group of unsequenced, context-dependent rules. Culture-general rules, such as "engage counterparts with respect" are inherited and combined with culture-specific rules such as "in Afghan culture, questions about female family members is disrespectful." Together they comprise a logical commonsense model of culture that focuses on microsocial interaction [7-8]. The content of the model was developed using ethnographic methods based on the work of Agar and Wertheim [9-10].

We compare authoring efficiency under the baseline and model-based conditions. One subject, who was not previously trained in either condition, authored six dialogs in both conditions with control for ordering effects. The first result is that instantiating a given scenario in a new culture is simpler, in terms of file changes and build process, using the model-based system. Our second result is that the time required to author the first dialog in each condition was comparable, but time to author subsequent dialogs falls faster in the modular condition than in the baseline. The third result is that the number of new objects required for the first dialog is larger in the modular condition, but like the time to author, it falls more steeply as more dialogs are authored. Although the number of subjects is small, these results indicate promising trends in the scalability and authoring efficiency of the modular system. In future work, we would like to investigate the tradeoff between efficiency and word-level accuracy with respect to a given dialog.

Acknowledgements. The authors would like to acknowledge the Office of Naval Research for their support of this work under contract NOOO14-09-C-0613.

References

1. Johnson, W.L., Valente, A.: Tactical Language and Culture Training Systems: Using AI to teach foreign languages and culture. AI Magazine 30(2), 72–83 (2009)
2. Lipi, A.A., Nakano, Y., Rehm, M.: A Socio-Cultural Model Based on Empirical Data of Cultural and Social Relationships. In: Ishida, T. (ed.) Culture And Computing. LNCS, vol. 6259, pp. 71–84. Springer, Heidelberg (2010)

3. Georgila, K., Traum, D.: Learning Culture-Specific Dialogue Models from Non Culture-Specific Data. In: Stephanidis, C. (ed.) HCII 2011 and UAHCI 2011, Part II. LNCS, vol. 6766, pp. 440–449. Springer, Heidelberg (2011)
4. Marsella, S., Gratch, J.: EMA: A Model of Emotional Dynamics. Journal of Cognitive Systems Research 10(1), 70–90 (2009)
5. Traum, D., 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)
6. Sagae, A., Johnson, W.L., Valente, A.: Conversational Agents in Language and Culture Training. In: Perez-Marin, D., Pascual-Nieto, I. (eds.) Conversational Agents and Natural Language Interaction: Techniques and Effective Practices, pp. 358–377. IGI Global, Madrid (2011)
7. Hobbs, J.R., Gordon, A.: Goals in a Formal Theory of Commonsense Psychology. In: Galton, A., Mizoguchi, R. (eds.) Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010), pp. 59–72. IOS Press, Amsterdam (2010)
8. Hobbs, J.R., Sagae, A.: Toward a Commonsense Theory of Microsociology: Interpersonal Relationships. In: Proceedings of the 10th Symposium on Logical Formalizations of Commonsense Reasoning. AAAI Spring Symposium Series, Stanford, California, March 21-23 (2011)
9. Agar, M.H.: The Professional Stranger: An Informal Introduction to Ethnography. Elsevier Science & Technology Books (1996)
10. Wertheim, S., Agar, M.: Culture that Works. In: Proceedings of the Second Conference on Cross-Cultural Decision Making, San Francisco, CA, July 23-25 (2012)

Author Index