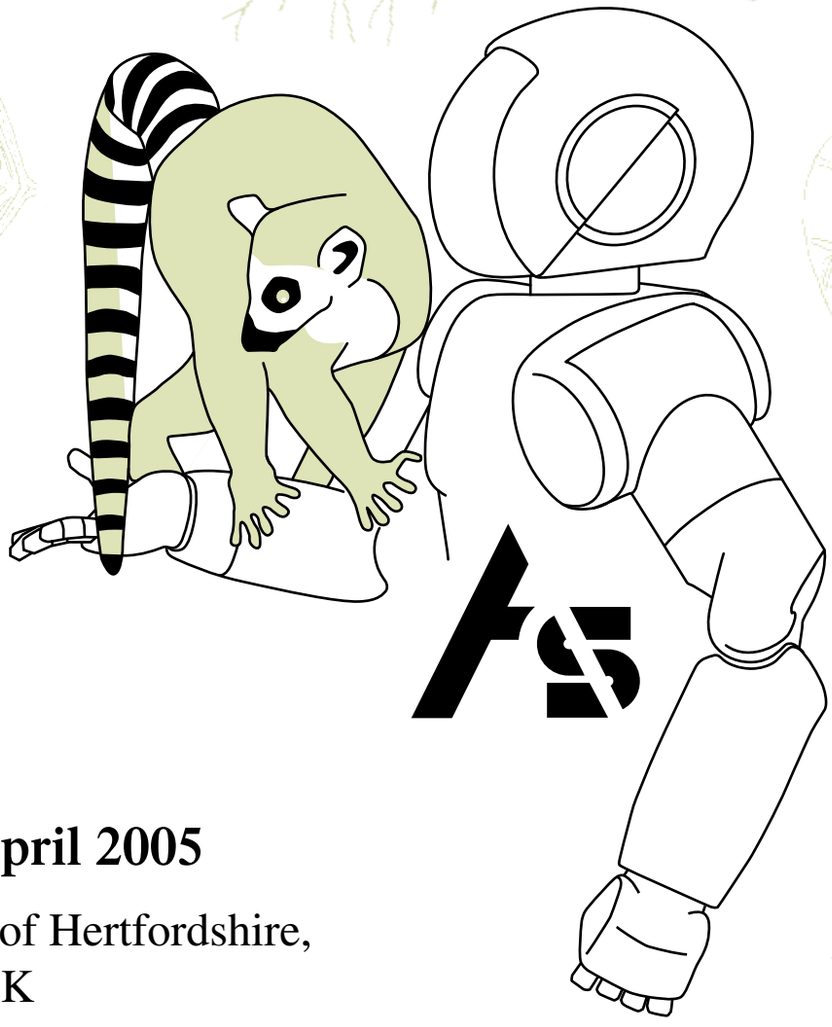
AISB'05: Social Intelligence and Interaction in Animals, Robots and Agents

Proceedings of the Symposium on Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation



12 - 15 April 2005

University of Hertfordshire,
Hatfield, UK

SSAISB 2005 Convention

AISB

UH

EPSRC
Engineering and Physical Sciences Research Council

AISB'05 Convention

Social Intelligence and Interaction in Animals, Robots and Agents

12-15 April 2005
University of Hertfordshire, Hatfield, UK

Proceedings of the Symposium on

Conversational Informatics for Supporting Social Intelligence and Interaction:

Situational and Environmental Information Enforcing Involvement in Conversation

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour
www.aisb.org.uk

Printed by

The University of Hertfordshire, Hatfield, AL10 9AB UK
www.herts.ac.uk

Cover Design by Sue Attwood

ISBN 1 902956 45 X

AISB'05 Hosted by

The Adaptive Systems Research Group
adapsys.feis.herts.ac.uk

The AISB'05 Convention is partially supported by:

The proceedings of the ten symposia in the AISB'05 Convention are available from SSAISB:

Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
1 902956 40 9

Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
1 902956 41 7

Third International Symposium on Imitation in Animals and Artifacts
1 902956 42 5

Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
1 902956 43 3

Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
1 902956 44 1

Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
1 902956 45 X

Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
1 902956 46 8

Normative Multi-Agent Systems
1 902956 47 6

Socially Inspired Computing Joint Symposium (Memetic theory in artificial systems & societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
1 902956 48 4

Virtual Social Agents Joint Symposium (Social presence cues for virtual humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)
1 902956 49 2

Table of Contents

The AISB'05 Convention
Social Intelligence and Interaction in Animals, Robots and Agents

Above all, the human animal is social. For an artificially intelligent system, how could it be otherwise?

We stated in our Call for Participation "The AISB'05 convention with the theme *Social Intelligence and Interaction in Animals, Robots and Agents* aims to facilitate the synthesis of new ideas, encourage new insights as well as novel applications, mediate new collaborations, and provide a context for lively and stimulating discussions in this exciting, truly interdisciplinary, and quickly growing research area that touches upon many deep issues regarding the nature of intelligence in human and other animals, and its potential application to robots and other artefacts".

Why is the theme of Social Intelligence and Interaction interesting to an Artificial Intelligence and Robotics community? We know that intelligence in humans and other animals has many facets and is expressed in a variety of ways in how the individual in its lifetime - or a population on an evolutionary timescale - deals with, adapts to, and co-evolves with the environment. Traditionally, social or emotional intelligence have been considered different from a more problem-solving, often called "rational", oriented view of human intelligence. However, more and more evidence from a variety of different research fields highlights the important role of social, emotional intelligence and interaction across all facets of intelligence in humans.

The Convention theme *Social Intelligence and Interaction in Animals, Robots and Agents* reflects a current trend towards increasingly interdisciplinary approaches that are pushing the boundaries of traditional science and are necessary in order to answer deep questions regarding the social nature of intelligence in humans and other animals, as well as to address the challenge of synthesizing computational agents or robotic artifacts that show aspects of biological social intelligence. Exciting new developments are emerging from collaborations among computer scientists, roboticists, psychologists, sociologists, cognitive scientists, primatologists, ethologists and researchers from other disciplines, e.g. leading to increasingly sophisticated simulation models of socially intelligent agents, or to a new generation of robots that are able to learn from and socially interact with each other or with people. Such interdisciplinary work advances our understanding of social intelligence in nature, and leads to new theories, models, architectures and designs in the domain of Artificial Intelligence and other sciences of the artificial.

New advancements in computer and robotic technology facilitate the emergence of multi-modal "natural" interfaces between computers or robots and people, including embodied conversational agents or robotic pets/assistants/companions that we are increasingly sharing our home and work space with. People tend to create certain relationships with such socially intelligent artifacts, and are even willing to accept them as helpers in healthcare, therapy or rehabilitation. Thus, socially intelligent artifacts are becoming part of our lives, including many desirable as well as possibly undesirable effects, and Artificial Intelligence and Cognitive Science research can play an important role in addressing many of the huge scientific challenges involved. Keeping an open mind towards other disciplines, embracing work from a variety of disciplines studying humans as well as non-human animals, might help us to create artifacts that might not only do their job, but that do their job right.

Thus, the convention hopes to provide a home for state-of-the-art research as well as a discussion forum for innovative ideas and approaches, pushing the frontiers of what is possible and/or desirable in this exciting, growing area.

The feedback to the initial Call for Symposia Proposals was overwhelming. Ten symposia were accepted (ranging from one-day to three-day events), organized by UK, European as well as international experts in the field of Social Intelligence and Interaction.

- Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
- Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
- Third International Symposium on Imitation in Animals and Artifacts
- Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
- Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
- Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
- Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
- Normative Multi-Agent Systems
- Socially Inspired Computing Joint Symposium (consisting of three themes: Memetic Theory in Artificial Systems & Societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
- Virtual Social Agents Joint Symposium (consisting of three themes: Social Presence Cues for Virtual Humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)

I would like to thank the symposium organizers for their efforts in helping to put together an excellent scientific programme.

In order to complement the programme, five speakers known for pioneering work relevant to the convention theme accepted invitations to present plenary lectures at the convention: Prof. Nigel Gilbert (University of Surrey, UK), Prof. Hiroshi Ishiguro (Osaka University, Japan), Dr. Alison Jolly (University of Sussex, UK), Prof. Luc Steels (VUB, Belgium and Sony, France), and Prof. Jacqueline Nadel (National Centre of Scientific Research, France).

A number of people and groups helped to make this convention possible. First, I would like to thank SSAISB for the opportunity to host the convention under the special theme of *Social Intelligence and Interaction in Animals, Robots and Agents*. The AISB'05 convention is supported in part by a UK EPSRC grant to Prof. Kerstin Dautenhahn and Prof. C. L. Nehaniv. Further support was provided by Prof. Jill Hewitt and the School of Computer Science, as well as the Adaptive Systems Research Group at University of Hertfordshire. I would like to thank the Convention's Vice Chair Prof. Chrystopher L. Nehaniv for his invaluable continuous support during the planning and organization of the convention. Many thanks to the local organizing committee including Dr. René te Boekhorst, Dr. Lola Cañamero and Dr. Daniel Polani. I would like to single out two people who took over major roles in the local organization: Firstly, Johanna Hunt, Research Assistant in the School of Computer Science, who efficiently dealt primarily with the registration process, the AISB'05 website, and the coordination of ten proceedings. The number of convention registrants as well as different symposia by far exceeded our expectations and made this a major effort. Secondly, Bob Guscott, Research Administrator in the Adaptive Systems Research Group, competently and with great enthusiasm dealt with arrangements ranging from room bookings, catering, the organization of the banquet, and many other important elements in the convention. Thanks to Sue Attwood for the beautiful frontcover design. Also, a number of student helpers supported the convention. A great team made this convention possible!

I wish all participants of the AISB’05 convention an enjoyable and very productive time. On returning home, I hope you will take with you some new ideas or inspirations regarding our common goal of understanding social intelligence, and synthesizing artificially intelligent robots and agents. Progress in the field depends on scientific exchange, dialogue and critical evaluations by our peers and the research community, including senior members as well as students who bring in fresh viewpoints. For social animals such as humans, the construction of scientific knowledge can't be otherwise.

Beppu, Japan.

Dedication:

I am very confident that the future will bring us increasingly many instances of socially intelligent agents. I am similarly confident that we will see more and more socially intelligent robots sharing our lives. However, I would like to dedicate this convention to those people who fight for the survival of socially intelligent animals and their fellow creatures. What would 'life as it could be' be without 'life as we know it'?

Kerstin Dautenhahn

Professor of Artificial Intelligence,
General Chair, AISB'05 Convention *Social Intelligence and Interaction in Animals, Robots and Agents*

University of Hertfordshire
College Lane
Hatfield, Herts, AL10 9AB
United Kingdom

Symposium Preface
Conversational Informatics for Supporting Social Intelligence & Interaction:
Situational and environmental information enforcing involvement in conversation

SYMPOSIUM OVERVIEW

Social intelligence is the ability to understand other social actors/agents and interact effectively with them. As social intelligence emerges through communication, communication ability is essential for accomplishing Social Intelligence and studying conversational exchanges as a social activity contributes to understanding Social Intelligence.

As a technology supporting Social Intelligence, this symposium discusses "Conversational Informatics": studies on human conversational behaviors as well as design/implementation of artifacts based on analyses of human conversations. For example, designing Embodied Conversational Agents (ECAs) based on a model of human communication behaviors is one of the main research approaches in Conversational Informatics. However, in previous ECAs, little has been studied on ECA's capability of engagement, such as eliciting users' spontaneous contribution to a conversation and initiation of it, and maintaining the conversation to accomplish a long interaction.

As essential aspects of communication that Conversational Informatics should support, this symposium focuses on "involvement" in a conversation and "situational information that makes conversants involved in a conversation". If conversational participants are not really involved in a conversation, information is not smoothly exchanged between them. In such a case, Social Intelligence does not emerge. Moreover, to make the conversation participants involved in a conversation, situational information is indispensable because communication is deeply linked and embedded in a situation.

Based on the motivation above, this symposium addresses issues on:
- identifying situational factors enforcing conversational involvement in human face-to-face communication, and investigating how conversational involvement contributes to accomplishing Social Intelligence.
- applying findings of an analysis of face-to-face communication to enhance artifact's ability of conversational involvement, specifically by recognizing/generating situational information, and facilitating Social Intelligence through human-agent interaction.
- developing technologies for creating a good amount of conversational contents by automatically analyzing situated conversations, and improving robustness of conversational systems using the rich real world contents.

INVITED SPEAKER

Dr. Candy Sidner from Mitsubishi Electric Research Laboratories (MERL) in Massachusetts, USA, will present a talk titled *“Engagement During Dialogues with Robots”*, addressing issues on communication robots.

PROGRAM COMMITTEE

Yukiko I. Nakano (RISTEX-JST, Japan) co-chair
Toyoaki Nishida (Kyoto University, Japan) co-chair

Elisabeth André (Universität Augsburg, Germany)
Timothy Bickmore (Boston University, USA)
Justine Cassell (Northwestern University, USA)
Satinder Gill (Middlesex University, UK)
Pat Healey (Queen Mary University of London, UK)
Jill Hewitt (University of Hertfordshire, UK)
Sadao Kurohashi (University of Tokyo, Japan)
Stacy Marsella (ISI, University of Southern California, USA)
Yuichi Nakamura (Kyoto University, Japan)
Igor Pandzic (University of Zagreb, Croatia)
Jeremy Pitt (Imperial College, UK)
Helmut Prendinger (Nippon Institute of Informatics, Japan)
Tomasz Rutkowski (Kyoto University, Japan)
Yasuyuki Sumi (Kyoto University, Japan)
David R. Traum (ICT, University of Southern California, USA)
Marilyn Walker (University of Sheffield, UK)

SYMPOSIUM WEBSITE

`http://kaiwa.ristex.jst.go.jp/AISB_CI/`

Conversational Agents, Humorous Act Construction, and Social Intelligence

Anton Nijholt
University of Twente, PO Box 217, 7500 AE Enschede
The Netherlands
anijholt@cs.utwente.nl

Abstract

Humans use humour to ease communication problems in human-human interaction and in a similar way humour can be used to solve communication problems that arise with human-computer interaction. We discuss the role of embodied conversational agents in human-computer interaction and we have observations on the generation of humorous acts and on the appropriateness of displaying them by embodied conversational agents in order to smoothen, when necessary, their interactions with a human partner. The humorous acts we consider are generated spontaneously. They are the product of an appraisal of the conversational situation and the possibility to generate a humorous act from the elements that make up this conversational situation, in particular the interaction history of the conversational partners.

1 Introduction

Embodied conversational agents have been introduced to play, among others, the role of conversational partner for the computer user. Rather than addressing the 'machine', the user addresses virtual agents that have particular capabilities and can be made responsible for certain tasks. The user may interact with embodied conversational agents to engage in an information service dialogue, a transaction dialogue, to solve a problem cooperatively, perform a task, or to engage in a virtual meeting. Multimodal emotion display and detection are among the research issues in this area of human-computer interaction. And so are investigations in the role of humour in human-computer interaction.

Humans use humour to ease communication problems in human-human interaction and in a similar way humour can be used to solve communication problems that arise with human-computer interaction. In Nijholt (2002) we discussed the role of humour for embodied conversational agents in the interface. It is a discussion on the possible role of humour support in the context of the design and implementation of embodied conversational agents. This paper is a revised version of Nijholt (2004). We discuss the role of embodied conversational agents in human-computer interaction and we have observations on the generation of humorous acts and on the appropriateness of displaying them by embodied conversational agents in order to smoothen, when necessary, their interactions with a human partner.

2 Humour in Interpersonal Interaction

In interpersonal interactions humans use humour, humans smile and humans laugh. Humour can be spontaneous, but it can also serve a social role and be used deliberately. A smile can be the effect of appreciating a humorous event, but it can also be used to regulate the conversation. Laughs have been shown to be related to topic shifts in a conversation (Consalvo, 1989).

2.1 Conversations and Dialogues

People smile and laugh when humour is used. It is not necessarily because someone pursues the goal of being funny or is telling a joke, but because the conversational partners recognize the possibility to make a funny remark fully deliberately, fully spontaneously, or something in between, taking into account social (display) rules, and then make this remark.

Humans employ a wide range of humour in conversations. Humour support, or the reaction to humour is an important aspect of personal

interaction and the given support shows the understanding and appreciation of humour. In Hay (2001) it is pointed out that there are many different support strategies. Which strategy can be used in a certain situation is mainly determined by the context of the humorous event. Humour support may show our involvement in a discussion, our motivation to continue and how much we enjoy the conversation or interaction.

Sometimes, conversations have no particular aim, except the aim of providing enjoyment to the participants. The aim of the conversation is to have an enjoyable conversation and humour acts as a social facilitator. In Tannen (1984), for example, an analysis is given of the humorous occurrences in the conversations held at a Thanksgiving dinner. Different styles of humour for each of the dinner guests could be distinguished. All guests had humorous contributions. For some participants more than ten percent of their turns were ironic or humorous. Humour makes one's presence felt, was one of her conclusions.

Similarity in humour appreciation also supports interpersonal attraction (Cann *et al.*, 1997). This observation is of interest when later we discuss the use of embodied conversational agents in user interfaces. Sense of humour is generally considered a highly valued characteristic of self and others. Nearly everybody claims to have an average to above average sense of humour. Perceived similarity in humour appreciation can therefore be an important dimension when designing for interpersonal attraction.

Other studies show how similarity in attitudes is related to the development of a friendship relationship. The development of a friendship relationship requires time, but especially in the initiation phase similarities are exploited (Stronks *et al.*, 2001).

2.2 Benefits

As mentioned, humour helps to regulate a conversation and can help to establish some common ground between conversational partners. It makes a conversation enjoyable and it supports interpersonal attraction.

Many benefits have been mentioned regarding humour in the teaching or learning process and sometimes they have been made explicit in experiments. Humour contributes to motivation attention, promotion of comprehension and retention of information, a more pleasurable learning experience, a development of affective feelings toward content, fostering of creative thinking, reducing anxiety, etc. The role of humour during instruction has been discussed in several papers. Despite the many experiments, it seems to be hard to generalize from the experiments that are conducted (Ziv,1988).

Describing and explaining humour in small task-oriented meetings is the topic of a study conducted by Consalvo (1989). An interesting and unforeseen finding was the patterned occurrence of laughter associated with the different phases of the meeting. Others have reported similar findings for different phases in negotiations or problem solving.

3 Embodied Conversational agents

Embodied conversational agents (ECAs) have become a well-established research area. Embodied agents are agents that are visible in the interface as animated cartoon characters or animated objects resembling human beings. Experiments have shown that ECAs can increase the motivation of a student or a user interacting with the system.

Embodied agents are meant to act as conversational partners for computer users. An obvious question is whether they, despite available verbal and nonverbal communication capabilities, will be accepted as conversational partners. That is, can we replace one of the humans in a human-to-human interaction by an embodied conversational agent without being able to observe important changes in the interaction behaviour of the remaining human? Can we model human communication characteristics in an embodied conversational agent that guarantee or improve natural interaction between artificial agent and human partner? Obviously, whether something is an improvement or more natural depends very much on the context of the interaction, but being able to model such characteristics allows a designer of an interface containing embodied agents to make decisions about desired interactions.

In the research on the 'computers are social actors' (CASA) paradigm (Reeves & Nass 1996) it has been convincingly demonstrated that people interact with computers as they were social actors. Due to the way we can let a computer interact, people may find the computer polite, dominant, extrovert, introvert, or whatever attitudes or personality (traits) we can display in a computer. Moreover, they react to these attitudes and traits as if a human being displayed them.

From the many CASA experiments we may extrapolate that humour, because of its role in human-human interaction, can play an important role in human-computer interactions. This has been confirmed with some specially designed experiments (Morkes et al. 2000) to examine the effects of humour in task-oriented computer-

mediated communication and in human-computer interaction.

4 Generation and Appropriateness

4.1 Introduction

In the previous sections we discussed the role of humour in human-human interaction and a possible role of humour in human-ECA interaction. Obviously, there are many types of humour and it is certainly not the case that every type of humour is suited for any occasion during any type of interaction. Telling a joke among friends may lead to amusement, while the same joke among strangers will yield misunderstanding or be considered as abuse. Therefore, an assessment of the appropriateness of the situation for telling a joke or making a humorous remark is necessary in all situations.

Appropriateness does not mean that every conversational participant has to be in a jokey mood for a humorous remark. Rather, it means that the remark or joke can play a role in the interaction process, whether it is deliberately aimed at achieving this goal, whether there is a mutually accepted moment for relaxing and playing or whether it is somewhere in between on this continuum. Clearly, it is also the 'quality' of the humorous remark that makes it appropriate in a particular situation. Here, 'quality' does not only refer to the contents of the remark, which may be based on a clever observation or ingenious wordplay, but in particular on an assessment whether or not to produce the humorous utterance. Just to make things more complicated, in some situations the possibility and the urge to make a humorous remark may overrule almost any social rule on how to behave.

In what follows we will talk about Humorous Acts (HA's). In telephone conversations a HA is a speech utterance. Apart form the content of what is being said, the speaker can only use intonation and timing in order to generate or support the humorous act.

In face-to-face conversations a humorous act can include, be supported or even made possible, by non-verbal cues. Moreover, references can be made, implicitly or explicitly, to the environment that is perceivable for the partners in the conversation. This situation also occurs when conversational partners know where each of them is looking at.

We emphasize again that participants in a discussion may, more or less deliberately, use humour as a tool to reach certain goals. A goal may be to smooth the interaction and improve mutual understanding. In that case a HA can generate and can be aimed at generating feelings of common attitudes and empathy, creating a bond between speaker and hearer. Whatever the aim is, conversational participants need to be able to compose elements of the context in order to generate a HA and they need to assess the current context (including their aims) in order to determine the appropriateness of generating a HA. This includes a situation where the assumed quality of the HA overrules conventions concerning cooperation during a goal-oriented dialogue.

We emphasize the spontaneous character of HA construction during conversational humour. The opportunity is there and although the generation is intended, it is also unpredictable and irreproducible. Nevertheless, it can be aimed at entertaining, to show skill in HA construction or to obtain a cooperative atmosphere. HA creation can occur when the opportunity to create a HA and a humorous urge to display the result temporarily overrules Gricean interaction principles concerning truth of the contribution, completeness of the contribution, or relevance of the contribution for the current conversation.

Generation (and interpretation) of HA's during a dialogue or conversation has hardly been studied. There is not really a definition, but the notion of conversational humour has been introduced in the literature (Attardo 1996).

4.2 Staging ECA Humour Generation

In Human-Computer Interaction one of the partners has to be designed and implemented. While on the one hand we need to understand as good as possible the models underlying human communication behaviour, this also gives us the freedom to make our own decisions concerning communication behaviour of the ECA, taking into account the particular role it is expected to play. From a design point of view, everything is allowed to make an ECA believable. In ECA design, rather than adhere to a guideline that says "try to be as realistic as possible", the more important guideline is "try to create an agent that permits the audience's suspension of disbelief."

When looking at embodied conversational agents we need to distinguish four modes of humour interpretation and generation. We mention these modes, but it should be understood that we are far from being able to provide the necessary appropriate models that allow them to display these skills. On the other hand, we don't always need agents that are perfect, as long as they are believable in their application. The first two modes concern the skills of the ECA:

- The ECA should be able to generate HA's. How should it construct and display the HA? When is it appropriate to do so? Apart from the verbal utterance to be used, it should consider intonation, body posture, facial expression and gaze, all in accordance with the HA. The ECA should have a notion of the effect and the quality of the HA in order to have it accompanied with nonverbal cues. Moreover, when in a subsequent utterance its human partner makes a reference to the HA, it should be able to interpret this reference in order to continue the conversation.
- The ECA should be able to recognize and understand the HA's generated by its human conversational partner. Apart from understanding from a linguistic or artificial intelligence point of view, this also requires showing recognition (e.g., for acknowledgement) and comprehension by generating appropriate feedback, including nonverbal behaviour (facial expression, gaze, gestures and body posture).

These are the two ECA points of view. Symmetrically, we have two modes concerning the skills of the human conversational partner. Generally, we may assume that humans have at least the skills mentioned above for ECAs.

- The human conversational partner should be able to generate HA's and accompanying signals for the ECA. Obviously, the human partner may adapt to the skills and personality of the particular ECA, as will be done when having a conversation with an other human.
- The human conversational partner should recognize, acknowledge and understand HA generation by the ECA, including accompanying nonverbal signals. Obviously, the ECA may have different ideas about acts being humorous than its particular conversational partner.

Our aim is to make ECA's more social by investigating the possibility to have them generate humorous acts. Two observations are in order. Firstly, when we talk about the generation of a HA and corresponding nonverbal communication behaviour of an ECA we should take into account an assessment of the appropriateness of generating this particular HA. This includes an assessment of the appreciation of the HA by the human conversational partner and therefore it includes some modelling of the interpretation of HA's by human conversational partners. That is, a model for generation of HA's requires a model of interpretation and appreciation of HA's. This is not really different from discourse modelling in general. An ECA needs to make predictions of what is going to happen next. Predictions help to interpret a next dialogue act or, more generally, a successor of a humorous act.

A second observation also deals with what is happening after introducing a HA in a conversation. What is its impact on the conversation and the next dialogue acts from a humour point of view? This introduces the issue of humour support, that is, apart from acknowledging, will the conversational partner support and further contribute to the humorous communication mood.

Finally, as a third observation, we need to consider whether HA generation by a computer or by an ECA gives rise to HA's that are essentially different and maybe more easily generated or accepted than human-generated HA's. An ECA may have less background and be less erudite, but it may have encyclopaedic knowledge of computers or a particular application. In addition, a computer or an ECA can become easily the focus of humour of a human conversational partner. Being attacked because of imperfect behaviour can be anticipated and the use of self-deprecating humour can be elaborated in the design of an ECA.

4.3 Appropriateness of HA Generation

Humour is about breaking rules, e.g. violating politeness conventions or, more generally, violating Gricean rules of cooperation. In creating humorous utterances during an interaction people hint, presuppose, understate, overstate, use irony, tautology, ambiguity, etc. (Brown and Levinson, 1978), i.e., all kinds of matters that do not follow Grice's Maxims. Nevertheless, humorous utterances can be constructive, that is, support the dialogue, and there can be a mutual understanding and cooperation during the construction of a HA. The HA's we would like to consider are, contrary to canned jokes that often lack contextual ties, woven into the discourse.

For HA construction, we need to zoom in on two aspects of constructing humorous remarks:

- recognition of the appropriateness of generating a humorous utterance by having an appraisal of the events that took place in the context of the interaction; dialogue history, goals of the dialogue partners (including the dialogue system), the task domain and particular characteristics of the dialogue partners have to be taken into account; and
- using contextual information, in particular words, concepts and phrases from the dialogue and domain knowledge that is available in networks and databases, to generate an

appropriate humorous utterance, i.e., a remark that fits in the context and that nevertheless is considered to be funny, is able to evoke a smile or a laugh, or that maybe is a starting point to construct a funny sequence of remarks in the dialogue.

It is certainly not the case that we can look at both aspects independently. With some exceptions, we may assume that, as should be clear from human-human interaction, HA's can play a useful and entertaining role at almost every moment during a dialogue or conversation. Obviously, some common ground, some sharing of goals or experiences during the first part of the interaction is useful, but it is also the quality of the generated HA that determines whether the situation is appropriate to generate this act. We cannot simply assess the situation and decide that now is the time for a humorous act. When we talk about the possibility to generate a HA and assume a positive evaluation of the quality of the HA given the context and the state of the dialogue context, then we are also talking about appropriateness.

4.4 Generation of HA's: An Example

Below we present an example of constructing a humorous act using linguistic and domain knowledge. The example is meant to be representative for our approach, not for its particular characteristics. It is an example of deliberately misunderstanding, an act that can often be employed in a conversation when some ambiguity in words, phrases or events is present, in order to generate a HA. Consider the text used in a Dilbert cartoon where a new "Strategic Diversification Fund" is explained in a dialogue between the Adviser and Dilbert:

Adviser: "Our lawyers put your money in little bags, then we have trained dogs bury them around town."

How to continue from this utterance? Obviously, we are dealing with a situation that is meant to create a joke, but nevertheless, all the elements of a non-constructed situation are there. What are these dogs doing? Burying lawyers or bags? So, a continuation could be:

Dilbert: "Do they bury the bags or the lawyers?"

Surely, this Dilbert remark is funny enough, although, from a natural language processing point of view it can be considered as a clarifying question, without any attempt to be funny. There is an ambiguity, that is, the system needs to recognize that generally dogs don't bury lawyers and therefore 'them' is more likely to refer to bags than to lawyers. Dogs can bury bags, dogs don't bury lawyers.

We need to be able to design an algorithm that is able to generate this question at this particular moment in the dialogue. However, the system should nevertheless know that certain solutions to this question are not funny at all. It can take the most likely solution, from a common sense point of view, but certainly this is not enough for our purposes. We need to introduce algorithms for anaphora resolution that decide to take a wrong but humorous solution, rather than that they take solutions that are the most likely correct ones. Obviously, then there is the question when this incorrect solution leads to a funny remark. When looking at previous language and humour research we can start with results that tell us about word and word meaning relations.

The example is certainly not complete in illustrating the full range of research aspects we need to tackle. In the cartoon we have a linguistic ambiguity, it can be resolved using common-sense knowledge and advanced methods for reference resolution, and we choose not to resolve it that way because we recognize that a less obvious solution can be used to construct a humorous continuation of the dialogue. In order to recognize this less obvious solution we need to include it on a stack of solutions, where in general the order of elements on

Figure 1: Strategic diversification Fund

the stack is determined by the increasing possibility to relate it to features of the antecedent in the history of the dialogue and the real world (Lappin and Leass, 1994). However, in this case, rather than following the order of the stack from the top to the bottom, we need to make a shortcut from elements, probably near the bottom of the stack, to nodes in a network containing semantic information that allows to reason about possibly humorous relationships between words and concepts.

4.5 Discussion

Although we have not seen humour research devoted to erroneous anaphora resolution, the approaches in computational humour research in general are not that different from what we saw in the example presented here. The approaches are part of the incongruity-resolution theory of humour. This theory assumes situations – either deliberately created or spontaneously observed – where there is a conflict between what is expected and what actually occurs. Ambiguity plays a crucial role. Phonological ambiguity, for example in certain riddles, syntactic ambiguity, semantic ambiguity of words, or events that can be given different interpretations by observers. Due to the different interpretations that are possible, resolution of the ambiguity may be unexpected, especially when one is led to assume a 'regular' context and only at the last moment it turns out that an other context allowing an other interpretation was present as well. These surprise disambiguations are not necessarily humorous. Developing criteria to generate humorous surprise disambiguations only is one of the challenges of humour theory. Attempts have been made, but they are rather preliminary. Pun generation is one example (Binsted & Ritchie, 1997), acronym generation (Stock & Strapparava, 2003) an other. In both cases we have controlled circumstances. These circumstances allow the use of WordNet and WordNet extensions and reasoning over these networks, for example, to obtain a meaning that does not fit the context or is in semantic opposition of what is expected in the context. No well-developed theory is available, but we see a slow increase in the development of tools and resources that make it possible to experiment with reasoning about words and meanings in semantic networks, with syllable and word substitutions that maintain properties of sound, rhyme or rhythm and with some higher-level knowledge concepts that allow higher-level ambiguities.

5 Tools, corpora, future research

5.1 Introduction

When discussing humour research for ECAs and their future development it is useful to distinguish between methods, tools and resources for verbal HA generation and methods and tools that may be called to help in order to have ECA's generate and display HA's using non-verbal communication acts. Graphics, animation and speech synthesis technology make it possible to have ECAs that display smiles, laughs and other signs of appreciation of the interaction. Multimodal and affective mark-up languages need to be extended in order to include the multimodal presentation of humorous acts in ECA behaviour.

5.2 Corpora, Annotations, Markup

Corpora are needed in order to study the creation of HA's in dialogues and naturally occurring conversations, including conversations that make references to common knowledge, task and domain knowledge, conversation history and the two- or three-dimensional visualized context of the conversation. With visualized context we mean the ECA and its environment (e.g., a reception desk, a lounge, posters in the environments, a particular training environment, other ECAs, including users and visitors, et cetera).

Corpora of conversations have been collected, but until now this collecting has hardly or not all been done from the point of view of humour or emotion research.[1] Consequently, hardly any experiments can be reported that have been performed using a corpus containing data that can be explored from the point of view of humour research. Hence, there is no attention to analysis, annotation, training, recognition or generation from a humour research and humour application point of view.

During a conversation or dialogue, having a particular HA or joke schema, an ECA can detect the appropriate moment to generate a particular type of joke or HA and it can use the average three-dimensional head movements to display the joke using verbal and nonverbal humour features.

[1] There exist corpora of jokes and, more interestingly for our purposes, there are corpora of conversations and dialogues between humans and computer services (e.g., travel and flight information). It will be interesting to look at corpora that are being collected and studied in the context of the European FP6 Integrated Project AMI (Augmented Multi-party Interaction) on meetings and the European FP6 Network of Excellence HUMAINE (Human-Machine Interaction Network on Emotion).

Average nonverbal communication behaviour as described in the previous paragraphs can be adapted by adding personality and emotional characteristics features. See (Ball and Breese, 2000), linking emotions and personality to nonverbal behaviour using Bayesian Networks. In (Allbeck and Badler, 2002), the emphasis is on adapting the gestures of ECA to its personality and gestures features.

5.3 Future Research Approaches

In the line of research on autonomous (intelligent and emotional) agents we need an ECA to understand why the events that take place generate enjoyment by its conversational partner and why it should display enjoyment because of its partner's appreciation of a HA. That is, models are needed that allow generation, prediction, detection and interpretation of humorous events. What events need to be distinguished, how does the ECA perceive them, and how does it integrate them at a semantic and pragmatic level of understanding of what is going on? There are two approaches to this question when we look at state-of-the-art research. One approach deals with speech and dialogue act prediction. What is going to happen next, given the history and the context of the dialogue? Can an ECA predict the next dialogue act by its conversational partner or can it compute the next dialogue act that is expected by its (i.e., the ECA's) conversational partner? Previous and possibly future dialogue acts are events that need to be 'appraised'.

In earlier research we used Bayesian Networks in order to predict dialogue acts. While this approach is unconvential from the usual point of view of event appraisal, it is an accepted approach in dialogue modelling research that has been implemented in a number of dialogue systems. Some attempts have been made to introduce multimodal dialogue acts. It seems to be useful to introduce more refined dialogue acts that take into account the willingness of a conversational partner to construct a humorous utterance and that take into account the possibility to give interpretations to (parts of) previous utterances that may lead to humorous acts. Obviously, in order to be able to do so we need corpora of natural conversations that allows us to design, train and test algorithms and strategies. Holistic user-state modelling, as advocated in the German SmartKom project (http://www.smartkom.org/), is a possible way to obtain data from which recognition algorithms can be designed.

Clearly, with such an approach we enter the area of emotion research. One of its viewpoints is that of appraisal theory, the evaluation of events and situations followed by categorizing arising affective states. Some of the theories that emerged from this viewpoint have been designed with computation in mind: designing a computational model to elicit and display emotions in a particular situation. A mature theory for calculating cognitive aspects of emotions is the OCC model, a framework of 22 distinct emotion types. A revised version of this model, presented in the context of believable ECA design was given in Ortony (2001). Can we make a step from event appraisal theories for deciding an appropriate emotion to appraisal theories for deciding the appropriateness of constructing a humorous act? As mentioned, issues that should be taken into account are the ability to construct a HA using elements of the discourse and the appropriateness of generating a HA in the particular context. In human-computer interaction applications some (mostly, stripped-down) versions of the model have been used.

It seems also useful to review existing theories and observations concerning the appraisal of (humorous) situations (available as events, in conversations, in verbal descriptions or stories) in terms of possible agent models that include explicit modules for beliefs, desires, intentions and emotions. Believes, desires and intentions (goals) define the cognitive state of an agent. Because of perceptive events state changes take place. From the humour modelling point of view agent models of states and state changes need to include reasoning mechanisms about situations where there is the feeling that on the one hand the situation is normal, while at the same time there is a violation of a certain commitment of the agent about how things ought to be. From a humour point of view, relevant cognitive states should allow detection of surprise, incongruity and reconstruction of incongruity using reasoning mechanisms.

6 Conclusions

This paper touches upon the state of the art of conversational agents, humour modelling and affective computing. We made clear that it is useful to introduce characteristics of human-human interaction in agent-human interaction, including the generation of humour and the display of appreciation of humour. We introduced the notion of a humorous act in a conversation. No algorithms for constructing humorous acts or for deciding when to generate an act were given. Rather we discussed the issues involved and we presented examples.

Acknowledgements

Research reported in this paper was supported by the European Future Emerging Technologies assessment project HAHAcronym (IST-2000-

30039), a joint project of ITC-IRST (Trento) and the University of Twente.

References

J. Allbeck & N. Badler. Toward Representing Agent Behaviours Modified by Personality and Emotion. Workshop *Embodied conversational agents - let's specify and evaluate them! AAMAS 2002*, Bologna, Italy.

S. Attardo. Humour. In J. Verschueren et al. (eds.), *Handbook of Pragmatics*. Amsterdam: John Benjamins Publishing Company, 1-17.

G. Ball & J. Breese. Emotion and personality in a conversational agent. In: J. Cassell et al. (eds.), Chapter 7 in: *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

K. Binsted. Using humour to make natural language interfaces more friendly. In Proc. *AI, ALife, and Entertainment* Workshop, Intern. J. Conf. on AI, 1995.

K. Binsted & G. Ritchie. Computational rules for punning riddles. *Humour: Intern. J. of Humour Research* 10 (1), 1997.

P. Brown & S.C. Levinson. Universals in language usage: Politeness phenomena. In Goody, E.N. (ed.), *Questions and Politeness*. Cambridge: Cambridge University Press, 1978, 56-289.

A. Cann, L.G. Calhoun & J.S. Banks. On the role of humour appreciation in interpersonal attraction: It's no joking matter. *Humour: Intern. J. of Humour Research* 10 (1), 1997, 77-89.

C.M. Consalvo. Humour in management: no laughing matter. *Humour: Intern. J. of Humour Research* 2 (3), 1989, 285-297.

J. Hay. The pragmatics of humour support. *Humour: Intern. J. of Humour Research* 14 (1), 2001, 55-82.

S. Lappin & H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 1994, 535-561.

J. Morkes et al. Effects of Humour in Task-Oriented HCI and Computer-Mediated Communication. *Human-Computer Interaction* 14 (4), 2000, 395-435.

A. Nijholt. Embodied Agents: A New Impetus to Humour Research. In: *The April Fools Day Workshop on Computational Humour*, O. Stock, C. Strapparava & A. Nijholt (eds.), In: Proc. Twente Workshop on Language Technology 20, Trento, Italy, 2002, 101-111.

A. Nijholt. Observations on humour act construction. In Trappl, R. (ed.), 17th *European Meeting on Cybernetics and Systems Research*, 2004, Vienna, 91-96.

A. Ortony. On making believable emotional agents believable. In Trappl, R. & P. Petta (eds.), *Emotions in humans and artifacts*. Cambridge, MIT, 2001.

B. Reeves & C. Nass. *The Media Equation: how people treat computers, televisions and new media like real people and places*. Cambridge: Cambridge University Press, 1996.

B. Stronks et al. Designing for friendship. In: Marriott, A. et al. (eds.), Proc. *ECA's - let's specify and evaluate them!* Bologna, 2002, 91-97.

O. Stock & C. Strapparava. An Experiment in Automated Humorous Output Production. Proceedings of the *International Conference on Intelligent User Interfaces (IUI)*, Miami, Florida, 2003, 300-302.

O. Stock & C. Strapparava. Getting serious about the development of computational humour. International Joint Conference on Artificial Intelligence (IJCAI) 2003, 59-64.

D. Tannen. *Conversational Style. Analyzing Talk among Friends*. Westport, Ablex Publishing, 1984.

A. Ziv. Teaching and learning with humour. J. of *Experimental Education* 57 (1), 1988, 5-15.

DigitalBlush: Towards a self-conscious community

Asimina Vasalou, Jeremy Pitt

Intelligent Systems and Networks Group
Electrical and Electronic Engineering Department
Imperial College London
Exhibition Road
LONDON SW7 2BT
{a.vasalou, j.pitt}@imperial.ac.uk

Abstract. In human societies, behavior is often mediated by a set of rules which if broken, have practical (i.e. tarred reputation) as well as emotional consequences (i.e. feeling shame for one's action). It is fear for both sanctions that leads to norm internalization. Trust and reputation mechanisms have focused on the practical consequences, neglecting to account for the weight of emotions. In this article, we present shame and embarrassment as a possible behavior-controlling mechanism working on an emotional level and strengthening current trust and reputation systems. Through a three line inquiry, we discuss our future work joining into a single objective: to develop a self-conscious community whose members may engage in controlled and even more importantly, prosocial behaviors towards other members.

1 Introduction

Increasingly, a substantial proportion of individuals engage in social activities by means of computer-mediation (e.g. chat, email, message boards). However, among other reasons, the lack of social context cues has put a strain on communication resulting in an accumulative number of dishonest and hostile behaviors (Collins, 1992). Technological applications (i.e. trust and reputation mechanisms) have managed to partly solve this phenomenon but have not been entirely successful (e.g. eBay, Amazon). Evidence to this point, is their constant, technical adaptation aimed towards managing the increasing number of evolutionary behaviors.

In this article, we take an emotion-driven approach towards solving this problem from the perspective of conversational informatics. We motivate human actors or agents to engage in fluid conversational behaviors and contribute to the objectives of this symposium through a fundamental exploration of self-evaluative emotions and the influence they exert on human behavior. As a consequence, we propose two strengthening extensions to existing approaches. First, we consider self-evaluative emotions in their social facility. The inclusion of social cues is considered a means towards controlling human behaviors and even supporting prosocial gestures in human-human networked interactions. Second, we consider self-evaluative emotions in their pacifying role, promoting forgiveness with the aid of situational and historical factors, thus encouraging more fulfilling interactions between human actors.

This article is structured as follows. In section two, we discuss the motivations behind our work, denominated 'DigitalBlush' (Pitt, 2004). We continue on in section 3 to formally define the self-evaluative emotions of shame and embarrassment. Section 4 presents the computational platform we are currently developing in which our theories will reify. In section 5 we discuss our future work on three key topics: (1) The makings of a self-conscious community where self-evaluative emotions may be experienced and human behaviors may consequently change; (2) The technical development of outlets which will promote prosocial gestures amongst members (e.g. politeness, forgiveness); (3) Evaluation techniques for testing the presence of self-evaluative emotions and their behavioral consequences in the DigitalBlush setting. Finally, we conclude this research outline in section 6 with a summary of future directions.

2 Motivations

Computer-mediated human-human interactions have been recognized to promote behavior that departs from the social dynamics of face-to-face communication. Uncontrolled, harmful behaviors have overrun communities of social function, ones driven by economic motives, online gaming groups and even more surprisingly, health and emotional support groups. As a response, many types of tools have been developed, all meant to recover and enhance one's sense of trust. For example, social communities use implicit and explicit peer-based networks where one engages in communication with another upon inferred or direct recommendation of a close friend (Jensen, Davis &

Farnham, 2002). Economic communities such as e-Bay and Amazon among others have deployed reputation mechanisms driven by buyers' impressions of their transaction with a particular seller. The historical compilation of the seller's success rate serves as a trust informant where sellers with high reputations are more trustworthy and reliable (Resnick, Zeckhauser, Friedman, & Kuwabara, 2000). Emotional support forums use human mediators who reliably check new members' backgrounds to make sure their motives are honest (Grady, 1998). But despite these efforts, malevolent actions have not been eradicated. Instead, members have persisted and found new ways to trick the system (Grady, 1998; Kong, 2000).

Online anonymity is regarded to be an important contributor to this problem (Johnson, 1997; Friedman, & Resnick, 1999). But maintaining an anonymous presence is particularly important in many domains. For example, an alcohol addict seeking advice on medical matters is more likely to engage in an online community which respects his/her privacy. In contrast, anonymity provides the means for deceit as anyone can take on the persona of an alcohol addict, elicit undue sympathy or even worst emotionally attack genuine members.

This ethical dilemma calls for new solutions. Johnson (1997) has presented the socially-driven human mechanism of 'internalizing norms of behavior' as a possible solution to this problem. In human societies, behavior is often mediated by a set of rules which if broken, have practical (i.e. tarred reputation) as well as emotional consequences (i.e. feeling shame for one's action). It is fear for both sanctions that leads to norm internalization. Current efforts have focused on the practical consequences, neglecting to account for the weight of emotions.

In psychology, two emotions emerge as positive informants for human behavior that often guide interactions and prevent harmful offences. The emotions in reference are shame and embarrassment. Shame has been described as a behavior controller that points to its host the standards of propriety and behaviors needed to operate in its social group (e.g. Scheff, 1988). In similar ways, shame is considered in some instances an "affective style" guiding human processing information, self-evaluation and self-regulatory behavior across time and situations (Magai & McFadden, 1995). The most severe consequences of shame point that it is only natural it would discretely hinder certain behaviors. Consequences that follow shame for ones' friends, family and acquaintances are subsequent drops in their self-esteem and respect, rejection, disappointment, and ridicule (Buss, 1980). In contrast to shame, embarrassment has a softer function as a guide for social etiquette and public conduct; For example we dress appropriately for a formal occasion and extend efforts when politely addressing a superior officer. In short, we place importance in our self presentation and strive for positive evaluations from the ones around us (Miller, 1996).

The inclusion of self-evaluative emotions in online interactions presents at least two important challenges. First, there is a need to determine whether human behavior may change as a consequence of this inclusion. Second, it is essential to go beyond this behavior controlling function to consider the prosocial derivatives of self-evaluative emotions (e.g. forgiveness).

In this article we discuss the previous two dimensions and their hypothetical application in a distance learning community, used as a running example. In this community, a trust and reputation mechanism will encourage peer-to-peer performance ratings as a way to collectively inform the decision-maker (i.e. professor) of the student's overall performance. The proposals offered here, will work on top of this mechanism towards attaining two end-objectives: (1) to transmit social cues making the experience of shame/embarrassment during a given violation possible and ultimately *leading towards behavior internalization* (2) to facilitate forgiveness where warranted as a reversal and prosocial mechanism ultimately *nurturing generous gestures amongst community members in replacement of malignant ones*.

We begin our work on the subject by first providing a formal definition of shame and embarrassment.

3 Shame and embarrassment, a formal definition

The self-conscious, self-evaluative emotions of shame and embarrassment emerge after the age of two. Their late onset is explained by their causation as they result from a direct comparison of *one's behaviour* to a *set of rules or standards*. This appraisal requires an advanced cognitive capacity which becomes available with age (Lewis, Sullivan, Stanger, & Weiss, 1989). Although both emotions share common grounds, their functions are to some extent divergent.

Shame is a severe painful experience which is overwhelmingly self-focused (Lewis, 1971). It is a direct result of the *perception* that one's behavior constitutes a rule, standard or goal violation. The violation in reference may be personal, cultural or

social in nature (Lewis, 2003). Shame can result in extremely harsh self-evaluations, in many cases leading to a withdrawal from society in an attempt to hide and avoid further humiliation (Buss, 1980; Ferguson & Crowley, 1997). Shameful experiences, although they may occur in social environments, do not depend on their actual presence.

In contrast to shame, which is generated through infractions of one's own evaluative standards, embarrassment is caused by fear of undesired evaluations from others. It is accompanied by the feeling of being observed and by an excessive awareness of one's social stature. Embarrassment in a sense results from one's efforts to project a desirable public image and one's failure to sustain it (Mansted & Semin, 1981; Miller 1996; Semin & Mansted, 1981). Unlike shame, the presence of others actively stimulates embarrassment.

There is no unanimous agreement on the relation between shame and embarrassment. Some emotion researchers claim the two emotions to be the same (Izard, 1977; Tomkins, 1963) while others believe them to be qualitatively distinct and self-standing (Babcock & Sabini, 1990; Lewis, 1992). We consider shame to derive from self-directed evaluations against one's own personal, social or cultural standards and embarrassment to result from self-directed evaluations against social standards. Given that ones' standards are subjective, it is possible for one person to feel embarrassment while another to feel shame as a result of the same violation. This subjective factor leads us to believe that the two emotions work interchangeably and should be considered together.

For our purposes and use, the flow of emotion experience for shame and embarrassment is simplified in the visualization below (see figure 1).

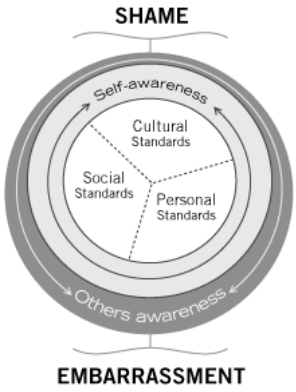


Figure 1: Shame and embarrassment visualization that depicts the emotional cause, a standard violation, while experiencing self and others awareness. Others awareness is shown stronger during the experience of embarrassment

Via this representation, we account for a number of factors that support or induce the two emotions.

- Both emotions are a consequence of a cultural and/or personal and/or social standard violation
- Self-awareness is requisite for the two emotions to materialize
- In embarrassing situations the awareness of others is a strong predictor while in shame it plays a less important role.

We briefly return to our distance learning scenario to draw a parallel to the previous three points. If a student delivers a low quality assignment to his peers (constituting a social or personal *standard violation*), he/she is censured with bad ratings. In a community, where one has an acute sense of self and others (*self and others awareness*), such a social punishment should result into one of two emotions: shame, if maintaining quality work constitutes a personal standard, or embarrassment, due to the strong evaluation of fellow students present.

4 A communicative platform

In considering the makings of a platform for DigitalBlush, *socio-cognitive grids* as termed by Pitt and Artikis (2000) allow us a degree of freedom. Their definition considers both resources of networked computing and human participants as constituent parts of a single, unified grid. Therefore under the definition of *socio-cognitive grids*, we propose a DigitalBlush community brought together in a distance learning setting. In this setting, participants of the community (i.e. students and professors) will connect with other members via their personal agents termed 'moral agents'. Participants will conduct in team activities such as project assignments and will carry certain responsibilities towards their fellow members. Among other possibilities, in the context of this community a dishonest exchange of information (i.e. not delivering a promised assignment) will constitute a violation. Following successful assignment completions or rule violations, peers will rate each other quantitatively with performance ratings and qualitatively with expressive content.

Towards developing our two proposals, namely an emotional behavior controlling mechanism and one that encourages prosocial activity, a number of exchanges will take place in our platform (a more detailed discussion follows in 5.1 and 5.2):

- Human participants of a networked community will connect to each other synchronously and asynchronously via their moral agents.

- Moral agents will represent their human counterparts in the agent community.
- Moral agents will not be entirely autonomous. Their role will be one of communicative and moral facilitation. Specifically,
    - A moral agent will carry social cues from its human participant to other participants and in reverse (see 5.1)
    - A moral agent will evaluate the success of its interactions with others and will relate its moral judgment to its human counterpart (see 5.2)

In figure 2, we visually depict our platform, tying human to agent constituents.

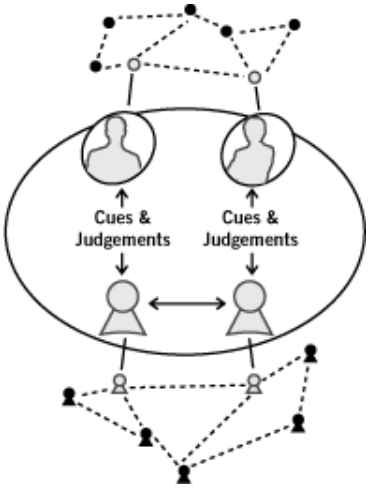


Figure 2: Human to human agent-facilitated interaction within the distance learning community

Revisiting the running example, the community exchanges discussed aim to activate two emotional reactions (further developed in section 5). (1) Social cues from self and others intend to fortify a sense of online awareness, where a student's bad conduct should be followed by a sense of shame or embarrassment; (2) Following a member's undesired behavior, the agent will facilitate reversal of the violation (when appropriate) by encouraging offender and victim to extend prosocial gestures to one another.

5 Future work: three key topics

In the section to follow we reveal our future work on three key topics. First, we consider the role of social cues on the experience of self-evaluative emotions and we examine consequential behaviors. We then talk about the fallible human nature and the importance of supporting the coping mechanism of forgiveness. We conclude with a proposal for emotional evaluation which will put our theories to the test.

5.1 The importance of social cues in sustaining self-evaluative emotions

Shame and embarrassment are involuntarily expressed through the face, posture and speech (Buss, 1980; Lewis & Ramsey, 2002). In the presence of others, efforts to conceal the emotion such as covering the face are often made (Buss, 1980). This feeling of conspicuousness and exposure is not surprising given the threat self-evaluative emotions pose to one's presentation and perception of self. The acute awareness brought on by face-to-face contact is communicated through a problem-solving experiment examining group communication preferences (i.e. phone, face-to-face, CMC). Participants of the study were found to be more at ease communicating by phone. According to their accounts, face-to-face contact brought on social pressures (Connell & Mendelsohn, 2001). Thus, the emotional experience begins and continues evolving around a strong awareness of one's self.

A second factor, awareness of others, is vital for embarrassment to occur. Others can be physically present as onlookers. Their presence may also be imaginary and their arrival anticipated, posing a continuous threat (Miller, 1996). This effect is vividly demonstrated through Dahl, Manchanda & Argo's (2001) work examining embarrassment during sensitive purchases. Subjects making purchases in a room alone experienced embarrassment due to the anticipated presence of others (i.e. someone walking in on them at any moment).

The interfacing of self and others during embarrassing or shameful encounters has been repeatedly investigated in psychology (Apsler, 1975; Costa, Dinsbach, Manstead & Bitti, 2001; MacDonald & Davies, 1983). But social cues channeled from one person to another (and in reverse) affect human behavior *online* as much as they are known to do so offline. We briefly illustrate this point with two examples.

Absence of self-awareness and others-awareness. A teenage girl joined an online emotional support message board pretending to be an expecting-mother. Other group members related to her and extended their support. But despite the honesty of others, the fraud continued on for months (Grady, 1998).

An awareness of self and others. During a problem-solving task, participants who were observed by video or application-sharing, in contrast to those who were not, demonstrated impeded cognitive performance. According to their reports, there was an acute sense of being watched

which resulted to fear of others' judgment (Brander & Mark, 2001).

The previous examples are informative as they point to some of the strengths and weaknesses of self and others awareness in online settings. But for our purposes they are limited in scope. To our knowledge, online social cues have not been examined in relation to self-evaluative emotions. Hence, we concentrate on the impact of identity and social cues during shameful or embarrassing instances online. To be more specific we are concerned with the following points:

- The strength of one's online identity (e.g. onymous or anonymous) in relation to the degrees of perceived self-awareness. Behavioral consequences will be examined through the forms of address one uses towards another member
- In the presence of a number of cues transmitted from the 'self' (e.g. video), we will measure self-awareness, the strength of self-evaluative emotions and resulting attitudes
- In the presence of 'others' (e.g. social proxy), we will measure one's sense of public self-awareness, the strength of self-evaluative emotions and resulting attitudes
- Finally we will explore the possibility of intentional embarrassment (i.e. embarrassing someone following their violation) in relation to self-awareness

Mainly, our future work on social cues attempts to strengthen self and others presence in order to cultivate the appropriate environment for an emotional response. We expect to find stronger emotional reactions when more aware of oneself and others present, in the long-term leading to more internalized behaviors. In a broader sense, our overarching aim is to reinvent the social stimulants of shame and embarrassment as manifested online and to explore their possible consequences alongside the practical implications as managed by trust and reputation mechanisms. Our findings will be used to inform the design of a self-conscious community, to be built on top of our communicative platform (see figure 2).

5.2 Moral agents as prosocial facilitators

The appeasement theory considers the *relationship restoration* between the victim and the transgressor, to qualify the very existence of embarrassment in the human species. Shame and embarrassment are followed by identifiable external signals (i.e. the blush) whose function is to appease and pacify observers or victims of violations. Hence, the emotion display plays an important role in exposing the transgressor's violation acknowledgement which as a result may prompt sympathy or *forgiveness* from others during more serious transgressions and amusement during milder ones (Keltner & Buswell, 1997). As a consequence, in DigitalBlush, the interplay between one's acknowledgment (supported through our work in 5.1) and the possibility of another's forgiveness is of paramount importance as it follows the natural cycle of emotion-reaction.

Moreover, there are a number of other incentives to support this suggestion. Unlike in human societies where forbidden actions are coupled with legal repercussions, reputation systems fulfill a socially-oriented duty by alerting the community's members of one's good standing. The decision to engage in collaborative efforts with another member is chiefly placed in the hands of each individual. *In human-human interactions, a violation of standards is unavoidable but not unforgivable.* Therefore, excluding forgiveness eradicates a significant moral-mechanism while attaching more value to cognitive ones. This as a result inhibits the prosocial effects that could follow an emotion-driven mechanism such as forgiveness. For example, reversal of one's harmful action with a good deed offers a possible pathway to forgiveness (Buss, 1980). In some instances, issuing forgiveness alone is known to stimulate feelings of gratitude from the transgressor leading to voluntary reparative actions. In contrast to that, punishment of low-intent acts (i.e. accidentally breaking something) may result in anger and low compliancy behaviors (Kelln & Ellard, 1999).

5.2.1 Motivating human forgiveness

Forgiveness from others depends on a complex of many factors. In sum, severity of the violation, frequency of past acts, one's intent, efforts to reverse the harm done, prior commitment with the transgressor and empathy felt for them, encompass some of the motivations interacting together towards forgiveness.

In a more detailed analysis, the severity of the current act is at first assessed. Harsher judgments result from more severe violations (Boon & Sulsky, 1997; Buss, 1980). Furthermore, a historical trail of one's past behaviors is compared against the current violation. Together, frequency and severity of past acts impact one's inclination to forgive (Buss, 1980). Apparent intent leads towards more negative attributions (Mansted & Semin, 1981) therefore suggesting that low intent actions lead to more positive attributions. Additionally, in human societies an immoral act may be reversed with a good deed (Buss, 1980). Prior familiarity and a relationship of commitment with the transgressor also increase the likelihood of forgiveness (McCullough, Rachal, Sandage, Worthington,

Brown & Hight, 1998). Good friends or successful business partners rely on a richer and mutually-rewarding history that fosters a propensity towards forgiveness. Finally, empathy, one's emotional response towards another's affect (Gruen & Mendelsohn, 1986) is regarded as a mediator appeasing the victim and facilitating forgiveness (McCullough, Worthington & Rachal, 1997).

5.2.2 Toward a formal model

We are building a model of forgiveness that integrates with a trust and reputation mechanism (e.g. Neville & Pitt, 2003) where students rate one another on a number of metric scales. The accumulation of student ratings overtime will actively support the trust mechanism and will also be used to facilitate forgiveness (see figure 2). More specifically, we consider our agents to have a 'moral' function because of their prosocial disposition following a given violation. In this role, a moral agent conducts an objective judgment and determines whether forgiveness is appropriate. It then conveys and justifies its evaluation to its human counterpart who will ultimately decide whether forgiveness will be issued. We are concerned with the following objectives:

- Propose a dynamic model for forgiveness that is domain independent (i.e. social or economical)
- Develop an operational model that will effectively integrate with the trust and reputation mechanism
- Design a facilitation tool which will communicate the agent's forgiveness decision and its judgment process to its human counterpart
- Determine whether the inclusion of *moral agents* in DigitalBlush will alter human behavior by fostering forgiveness from the victim and stimulating prosocial gestures from the transgressor.

5.3 Evaluating DigitalBlush

In designing an emotion-driven platform, we face the challenge of evaluating its success. This section considers the think-aloud protocol as a possible evaluation solution.

Ericsson and Simon (1993) presented the think aloud protocol as a viable method that accurately reveals the cognitive processes supporting problem-solving. Since then, among other applications, usability testing has leveraged the think aloud protocol to provide insight on users' cognitive workings during their interaction with an interface. In a typical usability session, human subjects are instructed to verbalize their thought process while conducting a certain task. Therefore, think-aloud accounts in the domain of usability testing, provide the reasons behind users' decisions or behaviors. Interestingly, the think aloud protocol in its original form did not suffice and was readapted to fit the special requirements of usability testing (Boren & Ramey, 2000).

Departing from its cognitive facility, we propose an emotional extension of the think aloud protocol that will serve as an evaluation tool revealing information about a user's emotional state. We present two facts to support this proposal. First, vocal expression is characterized by identifiable changes in acoustic cues that map to a number of emotions with above chance accuracy. In fact, the possibility of correlating vocal expression to an emotional state has been the driving force behind many vocal evaluation efforts (Scherer, 2003). Second, voice appears to be spontaneous in expression. For example, voice pitch has been found to persist despite efforts to suppress expression and to deceive others (Ekman, Friesen & Scherer, 1976).

In our short term objectives, we envision the emotional think-aloud extension to evolve around three points:

- ***User training***. Designing useful and effective instructions to train users in the extended protocol administration
- ***Individual differences.*** Determining the cultural universality of vocal emotional expressivity during the application of the extended think aloud
- ***Decoding.*** Achieving inter-rater agreement and reliability (always in terms of user-rater culture).

Our long term efforts will focus on speech analysis tools that will support evaluators during their assessments.

6 Summary

In this research outline, we have presented our vision of DigitalBlush as a platform where self-conscious emotions are experienced, control human behavior and foster prosocial gestures amongst members. In summary, our approach towards attaining the full-fledged objective is threefold. First, we suggest the inclusion of social cues from the self and others to support increasing self-awareness and others-awareness, as a result leading to norm internalization. Second, it is our belief that social cues transmitting one's emotion acknowledgement along with a moral agent's facilitation can foster the prosocial coping mechanism of forgiveness. Third, we intend to develop a viable emotional extension of the think

aloud protocol, aimed towards evaluating our work in the DigitalBlush platform.

As we end this article, we briefly mention the User-Centered Design (UCD) discipline and its transparent role throughout our work. Although emotion research poses a different set of requirements, we still find certain user-centered advocacies transferable and valuable. More specifically, our work is driven by the following two principles:

- ***Targeting the right group of users:*** DigitalBlush will be built around a distance learning scenario pertinent to student users. Our experiments will rely on this user base
- ***Involve users in the design process***: During our three research inquiries we will conduct exploratory studies with student users. The results yielded will be used to guide our design choices.

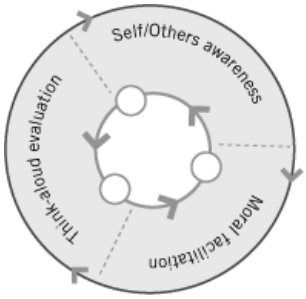


Figure 3: A cyclic approach to DigitalBlush where each research segment helps inform the next one

By means of the UCD principle, DigitalBlush is visualized into three distinct 'parts' (see figure 3) each evolving and developing around its self-contained needs. Driven by the same cause, all three are inter-related in several ways making knowledge-transfer possible in support of others.

In closing, DigitalBlush driven by the workings of self-evaluative emotions in human societies aims to enhance conversations between constituents of a community in several ways. Social cues inhibit harmful behaviors by integrating underlying components found in human-human conversations. Even more importantly, social cues and moral judgments facilitate forgiveness by encouraging members to extend prosocial gestures to one another. We envision these two points to add to current limitations of trust and reputation mechanisms by supporting 'internalization of norms of behavior' while at the same time departing from their behavior-controlling role to create fluid, altruistic and fulfilling interactions.

Acknowledgements

We gratefully acknowledge Winand Dittrich and Tanja Bänziger for their insightful comments on several parts of this article. This research was funded by the HUMAINE IST Framework VI Network of Excellence.

References

Apsler, R. Effects of Embarrassment on Behavior Toward Others. *Journal of Personality and Social Psychology*, 32(1): 145-153, 1975.

Babcock, M.K., & Sabini, J. On differentiating embarrassment from shame. *European Journal of Social Psychology*, 20: 151-169, 1990.

Boon, S., & Sulsky, L. Attributions of Blame and Forgiveness in Romantic Relationships: A Policy-capturing Study. *Journal of Social Behavior and Personality*, 12: 19-26, 1997.

Boren, M.T. & Ramey, J. Thinking Aloud: Reconciling Theory and Practice. *IEEE Trans. Prof. Comm.*, 43: 261-278, 2000.

Bradner, E. and Mark, G. Social Presence with Video and Application Sharing. In proceedings of the *ACM International Conference on Supporting Group Work* (GROUP '01). Boulder, Colorado, 154-161, 2001.

Buss, A.H. *Self-consciousness and social anxiety*. San Francisco. Freeman, 1980.

Connell, J.B. & Mendelsohn, G.A. Effects of Communication Medium on Interpersonal Perceptions: Don't Hang Up on the Telephone Yet! *Communications of the ACM*, 40 (1): 117-124, 2001.

Collins, M. *Flaming: The relationship Between Social Context Cues and Uninhibited Verbal Behaviour in Computer-mediated Communication.* [On-line]. Available: http://www.emoderators.com/papers/flames.html, 1992.

Costa, M., Dinsbach, W., Manstead, A.S.R., Bitti, P.E.R. Social presence, embarrassment, and nonverbal behavior.

Journal of Nonverbal Behavior, 25: 225-240, 2001.

Dahl, D. W., Manchanda, R.V. & Argo, J. J. Embarrassment in Consumer Purchase: The Roles of Social Presence and Purchase Familiarity. *Journal of Consumer Research*, 28(3): 473-481, 2001.

Ekman, P., Friesen, W. V. & Scherer, K. R. Body movement and voice pitch in deceptive interaction. *Semiotica*, 16: 23-27, 1976.

Ericsson, K. A. & Simon, H. A. *Protocol analysis; Verbal reports as data*. Cambridge, MA: Bradford books/MIT Press, 1993.

Ferguson, T.J. & Crowley, S.L. Gender differences in the organization of guilt and shame. *Sex Roles: A Journal of Research*, 37*:* 19-44, 1997.

Friedman, E. & Resnick, P. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy*, 10(2): 173-199, 1999.

Grady, D. Faking pain and suffering on the Internet. *The New York Times*, 1998.

Gruen, R. J. & Mendelsohn, G. Emotional responses to affective displays in others: The distinction between empathy and sympathy. *Journal of Personality & Social Psychology*, 51: 609-614, 1986.

Izard, C. *Human Emotions.* New York: Plenum Press, 1977.

Jensen, C., Davis, J. & Farnham, S. Finding others online: Reputation systems for social online spaces. In proceedings of the *CHI 2002*, Minneapolis, 447-454, 2002.

Johnson, D. Ethics Online: Shaping social behavior online takes more than new laws and modified edicts. *Communications of the ACM*, 40(1): 60-65, 1997.

Kelln, B.R.C., & Ellard, J.H. An equity theory analysis of the impact of forgiveness and retribution on transgressor compliance. *Personality and Social Psychology Bulletin*, 25: 864-872, 1999.

Keltner, D., & Buswell, B. N. Embarrassment: Its distinct form and appeasement functions. *Psychological Bulletin*, 122: 250-270, 1997.

Kong, D. Internet auction fraud increases. *USA Today*, 2000.

Lewis, H.B. *Shame and guilt in neurosis*. New York: International Universities Press, 1971.

Lewis, M. & Ramsey, D. Cortisol Response to Embarrassment and Shame. *Child Development*, 73: 1034-1045, 2002.

Lewis, M. The Role of the Self in Shame. *Social Research*, 70(4): 1181-1204, 2003.

Lewis, M. *Shame. The exposed self.* New York: The Free Press, 1992.

Lewis, M., Sullivan, M.W., Stanger, C. & Weiss, M. (1989). Self Development and self-conscious emotions. *Child Development*, 60:146-156, 1989.

MacDonald, L.M., & Davies, M.F. Effects of being observed by a friend or stranger on felt embarrassment and attributions of embarrassment. *Journal of Psychology*, 113: 171-174, 1982.

Magai, C. & McFadden, S.H. *The role of emotions in social and personality development*. New York: Plenum, 1995.

Manstead, A.S.R., & Semin, G.R. Social transgression, social perspectives, and social emotionality. *Motivation and Emotion*, 5: 249-261, 1981.

McCullough, M. E., Rachal, K.C., Sandage, S. J., Worthington, E. L., Brown, S. W. & Hight, T. L. Interpersonal Forgiving in Close Relationships: II. Theoretical Elaboration and Measurement. *Journal of Personality and Social Psychology*, 75: 1586-1603, 1998.

McCullough, M.E., Worthington, Jr., E.L. & Rachal, K.C. Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology*, 73(2): 321-336, 1997.

Miller, R.E. *Embarrassment. Poise and peril in everyday life*. New York: Guilford Press, 1996.

Neville, B., Jeremy P. A Computational Framework for Social Agents in Agent Mediated E-commerce. *ESAW* , 376-391, 2003.

Pitt J, Artikis A. Socio-cognitive grids: a partial ALFEBIITE perspective. In proceedings of the *First international workshop on socio-cognitive grids*. Santorini, Greece, 2003.

Pitt, J. Digital blush: towards shame and embarrassment in multi-agent information trading applications. *Cognition, Technology and Work*, 6(1): 23-36, 2004.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. Reputation Systems. *Communications of the ACM*, 2000.

Scheff, E.J. The shame-rage spiral: A case study of an interminable quarrel. In H.B. Lewis (Ed.), *The role of shame in symptom formation*. Hillsdale, NJ: Erlbaum, 1987.

Scherer, K. R. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40: 227-256, 2003.

Semin, G.R. & Manstead, A.S.R. The beholder beheld: A study of social emotionality. *European Journal of Social Psychology*, 11: 253-265, 1981.

Tomkins, S.S. *Affect, imagery, and consciousness: Vol. 2. The negative affects.* New York: Springer, 1963.

Mining Social Networks in Message Boards

Naohiro Matsumura*
*Graduate School of Economics, Osaka University
1-7 Machikaneyama, Toyonaka, Osaka, 560-0043, Japan
matumura@econ.osaka-u.ac.jp

David E. Goldberg†
†Illinois Genetic Algorithms Laboratory
University of Illinois at Urbana-Champaign
117 Transportation Building, 104 S. Mathews Avenue, Urbana, IL 61801
deg@uiuc.edu

Xavier Llorà†
‡Illinois Genetic Algorithms Laboratory
University of Illinois at Urbana-Champaign
117 Transportation Building, 104 S. Mathews Avenue, Urbana, IL 61801
xllora@uiuc.edu

Abstract

In the paper, we first present an approach to extract social networks from message boards on the Internet. Then we show structural features of 3,000 social networks extracted from 3,000 message boards from 15 categories in Yahoo!Japan Message Boards to prove the relationships between the features and the categories. After we classify social networks into three types (interactive communication, distributed expertise communication and soapbox communication), we suggest an approach for mining social networks to identify the types of communication, the roles of individuals, and important ties, all of which can be used to redesign the means communication as well as understand the state of communication.

1 Introduction

With the advent of popular social networking services on the Internet such as Friendster [1] and Orkut [2], social networks are again coming into the limelight with respect to this new communication platform. A social network shows the relationships between individuals in a group or organization where we can observe their social activities. For example, individuals who share a serious problem might all join a discussion on ways to solve the problem, while in another case only a few knowledgeable individuals might give information when individuals seek to find out more about specific topics.

Social networks have been studied for decades. Milgram discovered what we now call 'small-world phenomena' which show that everyone is connected to each other through a short chain in their social networks (Milgram 1967). Rogers classified each person into five types according to their stage regarding the adoption of new ideas (Rogers 1995). Granovetter insisted that weak ties spanning local relationship boundaries contribute to the diffusion of information (Granovetter 1973). Freeman proposed centrality measures to identify the importance of individuals and ties in a social network (Freeman 1978). Scott used social networks to identify gaps in information flow within an organization to find ways to get work done more effectively (Scott 1992). Krackhardt studied the importance of informal networks in organizations and revealed the effect of these networks on the accomplishment of tasks (Krackhardt and Hanson 93).

Social networks in online communication have been studied as well. Ohsawa et al. classified communities into six types according to the structural features of the word co-occurrence structure of communications (Ohsawa et al. 2002). Tyler et al. analyzed e-mail logs within an organization to identify communities of practice – informal collaborative networks – and leaders within the communities (Tyler et al. 2003). Matsumura et al. revealed the ef-

[1] http://www.friendster.com/
[2] http://www.orkut.com/

fect of anonymity and ASCII art on communication through message boards (Matsumura et al. 2004).

In addition, much research on social networks has been done in past decades and many properties of these networks are now well known. However, the causality between social networks and communication types are still veiled in mystery. In this paper, we aim at revealing some of the mystery by mining social networks in message boards to understand the state of communication and obtain new ideas regarding communication redesign.

The remainder of this paper proceeds as follows. In Section 2, we present an approach to extract social networks from message boards on the Internet. Then we introduce five indices that can be used to measure the structural features of social networks in Section 3. We describe three types of communication based on the classification of 3,000 social networks in Section 4. In Section 5, we suggest an approach for mining social networks which enables us to identify the types of communication, the roles of individuals, and important ties. Our conclusions and directions for future work are given in Section 6.

2 Extracting Social Networks

In a social network based upon online communication, the distance between individuals does not mean 'geographical distance' because each person lives in a virtual world. Instead, distance can be considered 'psychological distance' and this can be measured by the "influence" wielded among the members of the network.

Consider the situation where an individual p has a great deal of influence on an individual q. In this case, we can consider three types of relationship.

Case 1. p is close to q.

Case 2. q is close to p.

Case 3. p and q are close to each other.

Cases 1 and 2 show uni-directional relationships, and Case 3 shows a bi-directional relationship. Most previous studies of social networks employed undirectional social networks, i.e., Case 3, because of the simplicity of analysis. However, the relationship between individuals is not symmetric because of their activities and social situations (Wallace 1999). In addition, the difference of the distances between two individuals could be a key to understanding the relationships since it shows the communication gap between them. For this reason, we treat a social network as an asymmetric network where vertices denote individuals and directed links denote the flows of influence. In this paper, we do not need to distinguish Case 1 from Case 2 (e.g., the direction of the relationship distance) because our approach to measuring the communication gap, described later, produces the same results regardless of the direction.

We measure the influnce by using the IDM (Influence Diffusion Model) algorithm in which the influence between a pair of individuals is measured as the sum of propagating terms among them via messages (Matsumura 2003). Here, let a message chain be a series of messages connected by post-reply relationships, and the influence of a message x on a message y (x precedes y) in the same message chain be $i_{x \to y}$. Then, $i_{x \to y}$ is defined as

$$i_{x \to y} = |w_x \cap \cdots \cap w_y|, \tag{1}$$

where w_x and w_y are the set of terms in x and y, respectively, and $|w_x \cap \cdots \cap w_y|$ is the number of terms propagating from x to y via other messages. If x and y are not in the same message chain, we define $i_{x \to y}$ as 0 because the terms in x and y are used in a different context and there is no influence between them.

Based on the influence between messages, we next measure the influence of an individual p on an individual q as the total influence of p's messages on other's messages through q's messages replying to p's messages. Let the set of p's messages be α, the set of q's messages replying to any of α be β, and the message chains starting from a message z be ξ_z. The influence from p onto q, $j_{p \to q}$, is then defined as

$$j_{p \to q} = \sum_{x \in \alpha} \sum_{z \in \beta} \sum_{y \in \xi_z} i_{x \to y}. \tag{2}$$

Here we see the influence of p on q as q's contribution toward the spread of p's messages. The influence of each individual is also measurable using $j_{p \to q}$. Let the influence of p be k_p, and all other individuals be γ. Then, k_p is defined as

$$k_p = \sum_{q \in \gamma} j_{p \to q}. \tag{3}$$

As an example of measuring the influence, let us use the simple message chain shown in Figure 1 where Anne posted Message 1, Bobby posted Message 2 as a reply to Message 1, and Cathy posted Message 3 and Message 4 as replies to Message 2 and Message 1, respectively. In the figure, solid arrows show the replies to previous messages, and dotted arrows show the flows of influence. Here, the influence between a pair of individuals is as follows.

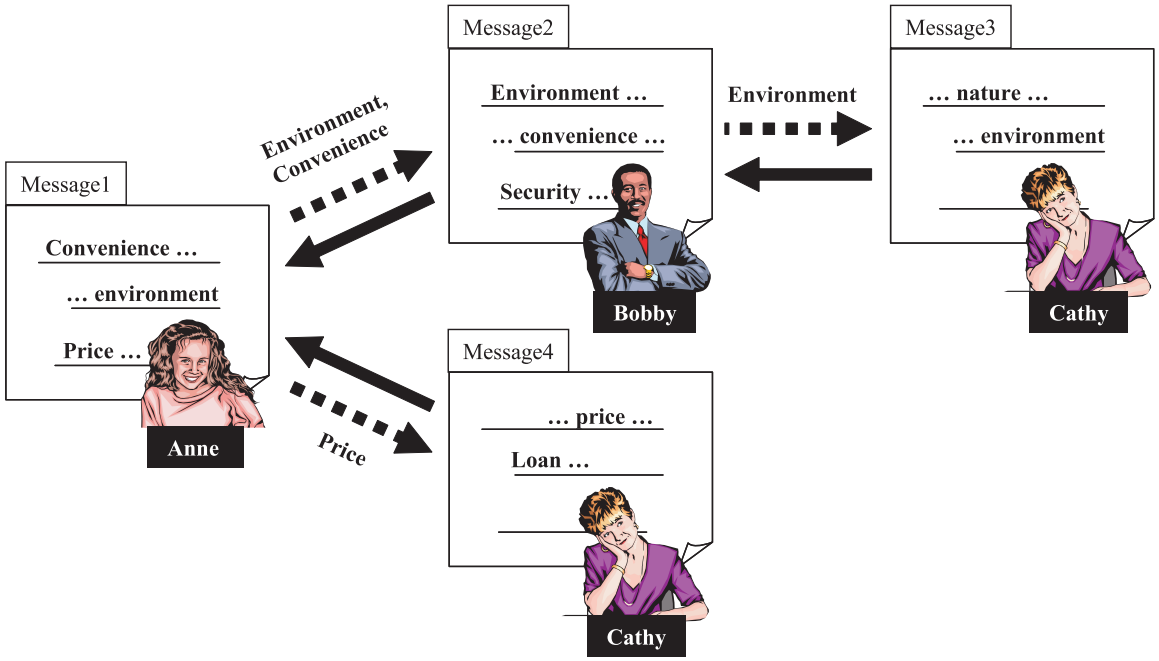


Figure 1: A message chain of five messages sent by three individuals.

- The influence of Anne on Bobby is 3 (i.e., $j_{Anne \to Bobby} = 3$), because two terms ("Environment" and "Convenience") were propagated from Anne to Bobby, and one term ("Environment") was propagated from Anne to Cathy via Bobby.
- The influence of Anne on Cathy is 1 (i.e., $i_{Anne \to Cathy} = 1$), because one term ("Price") was propagated from Anne to Cathy.
- The influence of Bobby on Cathy is 1 (i.e., $i_{Bobby \to Cathy} = 1$), because one term ("Environment") was propagated from Bobby to Cathy.
- The influence of Bobby on Anne and of Cathy on Anne is 0 (i.e., $i_{Bobby \to Anne} = 0$ and $i_{Cathy \to Anne} = 0$), because no term was propagated to Anne from either Bobby or Cathy.

Note that we ignore the influence of Anne on Cathy, even though a term "Environment" was propagated from Anne to Cathy via Bobby, because we want to measure direct influence between individuals. Instead, we consider the indirect influence of Anne on Cathy via Bobby as the contribution of Bobby, and add it to the influence of Anne on Bobby.

By mapping the influence between individuals, we can obtain a social network showing influence as in Figure 2 where their relationships are shown as directional links and the influence between them. From the figure, we can understand the influential relationships between individuals, and guess their roles in the communication — e.g., as opinion leaders and followers (Rogers 1995) — from the influence. For example, Anne would be an opinion leader because she was the source of the most influence on others. Bobby would be a mediator because he was a recipient of influence and transmitted some of it to Cathy. Cathy would be a follower because she only received influence from others.

The influence between individuals also shows the distance between them with respect to contextual similarity since the influence indicates the degree of their shared interest represented as terms. The influence and contextual distance between individuals are inversely related; i.e., the greater the influence, the shorter the distance. Here, let us define the length of a link (i.e., distance) as follows.

Definition 1 (The distance of a link) *The distance from an individual p to an individual q, $d_{p \to q}$, is defined as the value inversely proportionate to the influence from p to q; i.e., $d_{p \to q} = 1/j_{p \to q}$.*

The distance is between 0 and 1 when the influence is more than 0. However, the distance cannot be measured by the above definition if the influence is 0. In that case, we define the distance $n - 1$ (n is the number of individuals participating in communication) as the case of the weakest relationships; i.e.,

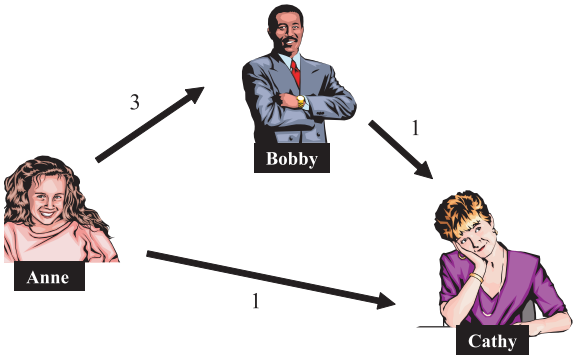


Figure 2: A social network showing the influence from Figure 1.

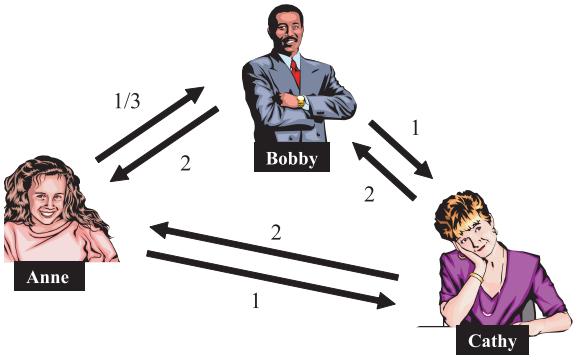


Figure 3: A social network showing the distance from Figure 1.

the diameter of a social network where all individuals are connected linearly with maximum distance. In this way, an asymmetric social network with distance is extracted from message chains as shown in Figure 3.

3 Communication Gaps in Social Networks

Based on the forward and backward distances between two individuals in a social network, we can consider three types of relationship between them.

Type 1. Two-sided communication: Two individuals actively exchange their ideas with each other. In this case, they are equivalent in communication, and their closeness can be identified by the short distances between them.

Type 2. One-sided communication: One individual's idea is received by an individual, while another individual's idea is not. In this case, the two senders are not equivalent in communication, as can be identified from the distance between them since the forward distances and backward distances will differ from each other.

Type 3. Sparse communication: Two individuals rarely exchange ideas with each other. In this case, they are not involved in communication, and their relationship can be identified from the long distances between them.

These three types, in other words, correspond to various "communication gaps" between two individuals. By measuring the communication gaps for all pairs of individuals in a social network, we can understand the state of their communication. For example, a small gap indicates active communication, while a large gap suggests inactive communication.

The distances can also be measured for two individuals who are not directly connected, but are indirectly connected via other individuals because they can exchange their ideas through others. Interestingly, the shortest path between individuals is often not an existing direct path, but an indirect path. In this paper, we consider the distance of the shortest path as the distance between two individuals since it reflects the real channel of their communication.

Here, let the distance of the shortest path from an individual p to another individual q be $d_{p\to q}$ ($d_{p\to q} = 0$ if p is equal to q) and the set of all individuals in a social network be γ. We then propose three indices, G_{diff}, G_{max} and G_{min}, that measure the collective communication gap in the social network from different points of view.

- G_{diff} measures the communication gap by accumulating the differences in the distances of the shortest paths between all pairs of individuals. G_{diff} is defined as

$$G_{diff} = \frac{1}{2}\sum_{p\in\gamma}\sum_{q\in\gamma}|d_{p\to q} - d_{q\to p}|, \qquad (4)$$

  where $|d_{p\to q} - d_{q\to p}|$ is the absolute value of $(d_{p\to q} - d_{q\to p})$.

- G_{max} measures the communication gap by accumulating the longer distances of the shortest paths between all pairs of individuals as a bottleneck hindering communication. G_{max} is defined as

$$G_{max} = \frac{1}{2}\sum_{p\in\gamma}\sum_{q\in\gamma}\max(d_{p\to q}, d_{q\to p}), \qquad (5)$$

  where $\max(d_{p\to q}, d_{q\to p})$ returns the maximum value from $\{d_{p\to q}, d_{q\to p}\}$.

- G_{min} measures the communication gap by accumulating the shorter distances of the shortest paths between all pairs of individuals. G_{min} is defined as

$$G_{min} = \frac{1}{2} \sum_{p \in \gamma} \sum_{q \in \gamma} \min(d_{p \to q}, d_{q \to p}), \quad (6)$$

where $\min(d_{p \to q}, d_{q \to p})$ returns the minimum value from $\{d_{p \to q}, d_{q \to p}\}$.

We can also consider another approach to identifying the state of communication by using only the distances between individuals instead of the differences between distances since the distance itself shows another aspect of communication gaps. The approach is known as "closeness centrality" where individuals nearby are more like to give/receive information more quickly than others (Freeman 1978). Based on closeness centrality, we propose two more indices for measuring the collective communication gap, C_{diff} and C_{dist}, as follows.

- C_{diff} measures the communication gap by accumulating the differences in the closeness centralities of individuals. C_{diff} is defined as

$$C_{diff} = \sum_{p \in \gamma} |c_p^{in} - c_p^{out}|, \quad (7)$$

where c_p^{in} means the inward closeness centrality that shows the sum of distances of the shortest paths from all other individuals to p, and c_p^{out} means the outward closeness centrality that shows the sum of distances of the shortest paths from p to all other individuals.

- C_{dist} measures the communication gap by accumulating the closeness centralities of individuals. C_{dist} is defined as

$$C_{dist} = \sum_{p \in \gamma} c_p^{out}. \quad (8)$$

Note that C_{dist} doesn't change even if we use c_p^{in} instead of c_p^{out}.

4 Three Types of Communication

To determine the features of the five indices proposed in Section 3, we analyzed 3,000 social networks. The analysis procedure was as follows.

Step 1. We downloaded 3,000 message boards from 15 categories of Yahoo!Japan Message Boards. To equalize the number of messages for each message board, we selected message boards having more than 300 messages and downloaded the first 300 messages. Then, we removed stop words (words except for noun and verb words) from all the messages to accurately measure content-derived influence. In this way, we prepared 3,000 message boards with each having 300 messages.

Step 2. We extracted a social network from each message board using the approach described in Section 2. To equalize the number of individuals in a social network, we constructed a social network with the 10 most influential individuals identified by Equation (3). We thus obtained 3,000 social networks, each consisting of 10 individuals.

Step 3. We measured G_{diff}, G_{max}, G_{min}, C_{dist}, and C_{diff} for the 3,000 extracted social networks. To equalize the range of the indices, we normalized each index by dividing it by its theoretical maximum. The normalized indices, G'_{diff}, G'_{max}, G'_{min}, C'_{diff}, and C'_{dist}, were

$$G'_{diff} = \frac{G_{diff}}{{}_nC_2(n-1)} \quad (9)$$

$$G'_{max} = \frac{G_{max}}{{}_nC_2(n-1)} \quad (10)$$

$$G'_{min} = \frac{G_{min}}{{}_nC_2(n-1)} \quad (11)$$

$$C'_{diff} = \frac{C_{diff}}{n(n-1)^2} \quad (12)$$

$$C'_{dist} = \frac{C_{dist}}{n(n-1)^2} \quad (13)$$

Step 4. The average of each index was calculated for each category.

The values of the five indices for the fifteen categories are shown in Table 1. Here, to investigate the relationships between the indices and categories, we applied hierarchical cluster analysis to the data. This analysis merges clusters based on the mean Euclidean distance between the elements of each cluster. A tree-like diagram, called a dendrogram, is then constructed as shown in Figure 4. From this figure, we can find three major clusters, each corresponding to a type of communication. We named the clusters as follows.

Table 1: The average of normalized five indices for 15 categories measured from 3,000 message boards in Yahoo!Japan Message Boards.

Categories	C'_{dist}	C'_{diff}	G'_{diff}	G'_{max}	G'_{min}
Family & Home	0.032	0.028	0.032	0.048	0.017
Health & Wellness	0.065	0.065	0.073	0.100	0.029
Arts	0.068	0.070	0.081	0.108	0.029
Science	0.072	0.068	0.078	0.110	0.034
Cultures & Community	0.085	0.061	0.071	0.120	0.051
Romance & Relationships	0.081	0.079	0.089	0.125	0.038
Hobbies & Crafts	0.095	0.092	0.106	0.146	0.043
Regional	0.161	0.120	0.142	0.230	0.093
Entertainment	0.151	0.129	0.157	0.228	0.075
Government & Politics	0.217	0.167	0.197	0.313	0.120
Business & Finance	0.241	0.160	0.184	0.331	0.150
Schools & Education	0.253	0.161	0.195	0.349	0.158
Recreation & Sports	0.239	0.208	0.253	0.362	0.116
Computers & Internet	0.447	0.221	0.272	0.579	0.315
Current Events	0.455	0.220	0.271	0.588	0.322

Interactive Communication: This cluster includes seven categories (“Arts”, “Sciences”, “Health & Wellness”, “Culture & Community”, “Romance & Relationships”, “Hobbies & Crafts”, and “Family & Home”), the indices of which are considerably smaller than those of other categories. The topics in these categories are common and familiar to many individuals who share these interests. As a consequence, individuals are naturally involved in the communication, and actively exchange their ideas with others.

Distributed Expertise Communication: This cluster includes six categories (“Regional”, “Entertainment”, “Business & Finance”, “Schools & Education”, “Government & Politics”, and “Recreation & Sports”), the indices of which are generally higher than those of the interactive communication categories. As the topics in these categories are somewhat specific and disputable, experienced or knowledgeable individuals contribute most to the communication.

Soapbox Communication: This cluster includes two categories (“Computers & Internet” and “Current Events”), the indices of which are higher than those of the above two clusters. The topics in these categories are mainly current affairs or topical news, and the communication is one-way from informers to audiences (or lurkers).

From the above results, we can say that there are roughly three types of communication in Yahoo!Japan Message Boards. If these types are common properties in other social networks, it will be possible to identify the state of communication by measuring the five indices.

We expected that the five indices would reveal different aspects of communication, however the Pearson correlation coefficients between them were over 0.9. This meant that these indices were not statistically distinct. In other words, we should be able to identify the types of communication by using only one index instead of all five. In the following, we show some approaches to mining social networks based on G'_{max} because G'_{max} proved to be the most discriminative index for identifying clusters.

5 Mining Social Networks

Mining of social networks is done to identify the types of communication, understand the roles of individuals, and explore remedies for communication gaps in social networks. In this section, we present case studies of mining social networks based on communication types and G_{max} explained before.

5.1 Communication Types

We prepared two types of message log, log 1 and log 2, each of which was extracted from a mes-

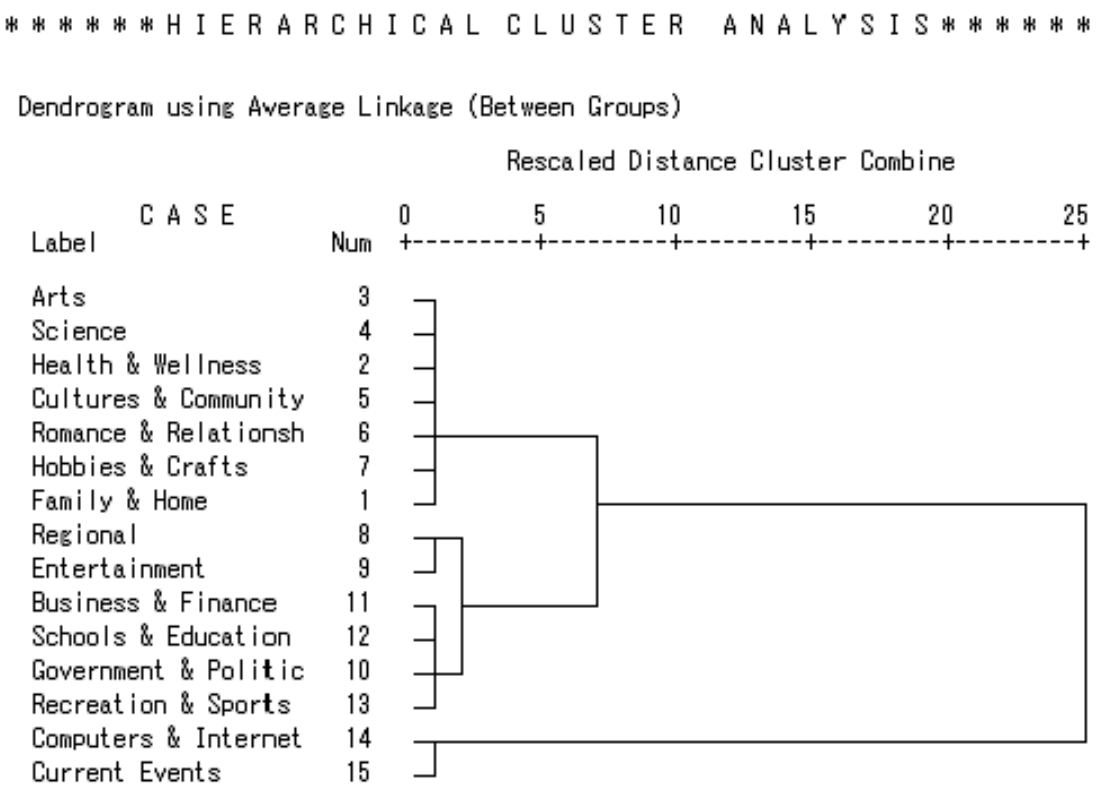


Figure 4: A dendrogram produced through hierarchical cluster analysis.

sage board on DISCUS[3] (Goldberg 2003). Log 1 consisted of 72 messages posted by five individuals (one Japanese researcher, one Spanish researcher, and three Japanese businesspersons) who discussed "cell phones and women" in English. Log 2 consisted of 102 messages posted by six individuals (two Japanese researchers, one Spanish researcher, one Japanese businessperson, one Chinese student, and one Taiwanese student) who discussed "purchase of a house" in English. Note that the real names of individuals have been replaced with fictional names to protect their privacy.

Figures 5 and 6 are social networks showing the distance extracted from log 1 and log 2, respectively, using the approach described in Section 2. Once we obtained the social networks, G'_{max} was measured as 0.069 from Figure 5 and 0.364 from of Figure 6. Comparing G'_{max} with the clustering results in Section 4, we can identify the types of communication of log 1 and log 2 as "interactive communication" and "distributed expertise", respectively.

The topic in log 1 was familiar to the individuals because they use cell phones everyday. On the other hand, the purchase of a house is a big event in life, and many individuals have no experience of making such a purchase. Therefore, the communication in log 2 was controlled by a few experienced individuals. Thus, the communication types of log 1 and log 2 seem to have properly reflected the types of real communication.

5.2 Roles of Individuals

As shown by the definition of G_{max} in Equation (5), G_{max} is measured by summing each individual's communication gaps with respect to other individuals. That is, G_{max} for each individual is easily measurable by translating the definition of G_{max}. Let the communication gap of an individual p be g_{max}. We can then define g_{max} as

$$g_{max} = \frac{1}{2} \sum_{q \in \gamma} \max(d_{p \to q}, d_{q \to p}). \qquad (14)$$

Normalized g_{max} is measured by dividing by the theoretical maximum. Let the normalized g_{max} be

[3] http://www-discus.ge.uiuc.edu/

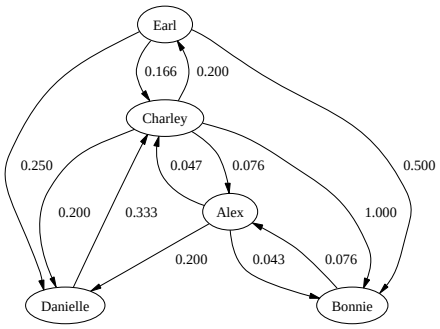


Figure 5: The social network with distance extracted from log1. $G_{max} = 0.069$.

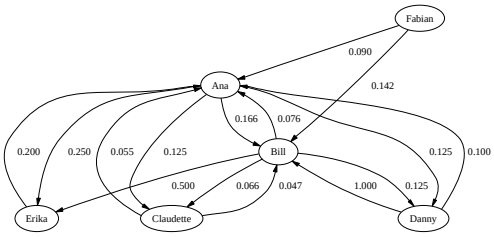


Figure 6: The social network with distance extracted from log 2. $G_{max} = 0.364$.

Table 2: Three most important links in Figure 6.

Link	l_{max}	G''_{max}	G'_{max}
Erika → Ana	0.574	0.938	0.364
Danny → Ana	0.270	0.634	0.364
Ana → Erika	0.226	0.590	0.364

g'_{max}. Then, g'_{max} is defined as

$$g'_{max} = \frac{g_{max}}{\frac{1}{2}n(n-1)}. \quad (15)$$

As the G_{max} of Figure 6 is high, let us measure each individual's g'_{max} to reveal the source of the communication gap. As shown in Table 3, Fabian has the highest g'_{max}. From this, we can understand that there are communication gaps around Fabian.

The roles of individuals are also identified from their relationships with others. If removing a link raises G'_{max}, the link is considered to help reduce the communication gap and is therefore important. That is, we can measure the importance of each link by comparing the G'_{max} of the original social network with that link to G'_{max} of a social network without the link. Let the importance of a link be l_{max}. We then define l_{max} as

$$l_{max} = G''_{max} - G'_{max}, \quad (16)$$

Table 3: g'_{max} of each individual in Figure 6.

	Ana	Bill	Claudette	Danny	Erika	Fabian
g'_{max}	0.038	0.039	0.038	0.040	0.042	0.167

where G''_{max} is the G'_{max} measured from a social network without the link. From the top three l_{max} values in Figure 6 listed in Table 2, we can see that a link from Erika to Anna is the most important.

During interviews with the individuals in Figure 6, we found that they considered Fabian creative, but strong-willed to the degree that nobody could counter his ideas. As a result, the communication around him did not go well. They also agreed with the results regarding the important links in Table 2 because these were the links actively used to exchange ideas during the period under study.

5.3 Remedies for Communication Gaps

Once we can obtain information about who are the causes of communication gaps and which links are most important for communication, we can prepare a remedy to improve communication. For example, the following three approaches could be the candidates of the remedies, apart from the feasibility.

- Persuade inactive individuals to contribute to the communication, or persuade strong individuals to listen to others' ideas. While this approach is straightforward, the effectiveness of such persuasion depends on many factors which are beyond the scope of this paper.
- Remove inactive or strong individuals from the communication. This approach is easy and has an immediate effect, but it is not constructive since we would lose the potential contributions of the excluded individuals.
- Add another individual who can communicate with inactive or strong individuals to bridge communication gaps. To find such individuals, the link importance can be used.

We can also simulate the importance of virtual links which make the communication gap smaller. For example, if there was a link from Anne to Fabian, G'_{max} would dynamically drop from 0.364 to 0.069 as shown in Figure 7.

Needless to say, there approaches above should be planned carefully before putting into practice. Having an inverview with individuals for surveying the

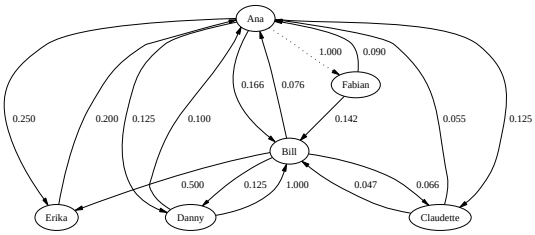


Figure 7: A link from Ana to Fabian causes G'_{max} to dynamically drop from 0.364 to 0.069.

effect of a remedy would further enhance the possibility of the success of mining social networks.

6 Conclusion

We have described an approach to extracting social networks from a message board. We then revealed three types of communication from point of communication gaps, and suggested some approaches to mining social networks.

Human beings are social creatures, and we could not survive without cooperating with others. It also suggests that understanding how relationships are created and functioned is essential to make our lives happier and richer. The range of our social networks is rapidly expanding than before as the Internet accelerates our communication via E-mail, Chat, Video conference system etc. Reflecting such a situation, managing and utilizing social networks will be an increasingly important part of our lives. We hope this study will contribute to the realization of a better way of life through human relationships.

Acknowledgements

We thank the Chance Discovery Consortium for their financial support. We are also extremely grateful to Prof. Yukio Ohsawa, Hiroshi Tamura, Yuichi Washida, and Masataka Yoshikawa, and all the members of IlliGAL for their insightful discussions.

References

[Freeman 1978] Freeman, L.C. (1978) Centrality in Social Networks, *Social Networks*, **1**, 215–39.

[Goldberg 2003] David E. Goldberg, Michael Welge, and Xavier Llorà (2003) DISCUS: Distributed Innovation and Scalable Collaboration In Uncertaion Settings, IlliGAL Report No. 2003017

[Granovetter 1973] Granovetter, M. (1973) The Strength of Weak Ties, *American Journal of Sociology*, **78**, 1360–1380.

[Krackhardt and Hanson 93] David Krackhardt and Jeffery R. Hanson (1993) Informal Networks: The Company behind the Chart, *Harvard Business Review*, July-August 1993, 104–111.

[Matsumura 2003] Naohiro Matsumura (2003) Topic Diffusion in a Community, Yukio Ohsawa and Peter McBurney (Eds.), *Chance Discovery*, Springer Verlag, 84–97.

[Matsumura et al. 2004] Naohiro Matsumura, Asako Miura, Yasuyuki Shibanai, Yukio Ohsawa, and Toyoaki Nishida (2004) The Dynamism of Nichannel, *Journal of AI & Society*, Springer (in press)

[Milgram 1967] Milgram, S. (1967) The Small World Problem, *Psychol. Today*, **1**, 61–67.

[Ohsawa et al. 2002] Yukio Ohsawa, Hirotaka Soma, Yutaka Matsuo, Naohiro Matsumura, and Masaki Usui (2002) Featuring Web Communities Based on Word Co-occurrence Structure of Communications, Proc. of WWW2002, 736–742.

[Rogers 1995] Rogers, E.M. (1995) *Diffusion of Innovations* (Fourth Edition), New York, Free Press.

[Scott 1992] Scott, W.R. (1992) *Organizations: Rational, Natural, and Open Systems*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[Tyler et al. 2003] Tyler, J.R., Wilkinson, D.M. and Huberman, B.A. (2003) Email as spectroscopy: Automated Discovery of Community Structure within Organizations. (Preprint http://www.lanl.gov/arXiv:cond-mat/0303264)

[Wallace 1999] Wallace, Patricia (1999) The Psychology of the Internet, Cambridge University Press.

Engagement During Dialogues with Robots

Candace L. Sidner, Christopher Lee [*†]
[†] Mitsubishi Electric Research Laboratories
Cambridge, MA 02139
{sidner,lee} @merl.com

Cory Kidd[†]
[†]Media Lab, MIT
Cambridge, MA 02139
corky @media.mit.edu

Abstract

This paper reports on our research on developing the ability for robots to engage with humans in a collaborative conversation. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Many of these interactions are dialogues and we focus on dialogues in which the robot is a host to the human in a physical environment. The paper reports on human-human engagement and its application to a robot that collaborates with a human on a demonstration of equipment.

1 Introduction

One goal for interaction between people and robots centers on conversation about tasks that a person and a robot can undertake together. Not only does this goal require linguistic knowledge about the operation of conversation, and real world knowledge of how to perform tasks jointly, but the robot must also interpret and produce behaviors that convey the intention to start the interaction, maintain it or to bring it to a close. We call such behaviors *engagement behaviors*. Our research concerns the process by which a robot can undertake such behaviors and respond to those performed by people.

Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Engagement is supported by the use of conversation (that is, spoken linguistic behavior), ability to collaborate on a task (that is, collaborative behavior), and gestural behavior that conveys connection between the participants. While it might seem that conversational utterances alone are enough to convey connectedness (as is the case on the telephone), gestural behavior in face-to-face conversation provides significant evidence of connection between the participants.

Conversational gestures generally concern gaze at/away from the conversational partner, pointing behaviors, (bodily) addressing the conversational participant and other persons/objects in the environment, and various hand signs, all with appropriate synchronization with the conversational, collaborative behavior. These gestures are culturally determined, but every culture has some set of behaviors to accomplish the engagement task. These gestures sometimes also have the dual role of providing sensory input (to the eyes and ears) as well as telling conversational participants about their interaction. Our research focuses on how gestures tell participants about their interaction, but we also must address the matter of sensory input as well.

Conversation, collaboration on activities, and gestures together provide interaction participants with ongoing updates of their attention and interest in a face-to-face interaction. Attention and interest tell each participant that the other is not only following what is happening (i.e. grounding), but intends to continue the interaction at the present time.

Not only must a robot produce engagement behaviors in collaborating with a human conversational partner (hereafter CP), but also it must interpret similar behaviors from its CP. Proper gestures by the robot and correct interpretation of human gestures dramatically affect the success of interaction. Inappropriate behaviors can cause humans and robots to misinterpret each other's intentions. For example, a robot might look away for an extended period of time from the human, a signal to the human that it wishes to disengage from the conversation and could thereby terminate the collaboration unnecessarily. Incorrect recognition of the human's behaviors can lead the robot to press on with an interaction in which the human no longer wants to participate.

2 Learning from Human Behavior

To determine gestures, we have developed a set of rules for engagement in the interaction. These rules are gathered from the linguistic and psycholinguistic literature (for example, Kendon (1967)) as well as from 3.5 hours of videotape of a human host guiding a human visitor on tour of laboratory artifacts. These gestures reflect US standard cultural rules for US speakers. For other cultures, a different set of rules must be investigated.

Our initial set of gestures were quite simple, and applied to a hosting activities, that is, the collaborative activity in which an agent provides guidance in the form of information, entertainment, education or other services in the user's environment and may also request that the user undertake actions to support the fulfillment of those services. Initially, human-robot conversations consisted of the robot and visitor greeting each other and discussing a project in the laboratory. However, in hosting conversations, robots and people must discuss and interact with objects as well as each other.

As we have learned from careful study of the videotapes we have collected (see Sidner et al, 2003), people do not always track the speaking CP, not only because they have conflicting goals (e.g. they must attend to objects they manipulate), but also because they can use the voice channel to indicate that they are following information even when they do not track the CP. They also simply fail to track the speaking CP sometimes without the CP attempting to direct them back to tracking. Our results differ from those of Nakano et al (2003), perhaps because of the detailed instruction giving between the participants in Nakano's experiments.

Experience from this data has resulted in the *principle of conversational tracking*: participants in a collaborative conversation track the other's face during the conversation in balance with the requirement to look away to: (1) participate in actions relevant to the collaboration, or (2) multi-task activities unrelated to the collaboration at hand, such as scanning the surrounding scene for interest, avoidance of damaging encounters, or personal activities. To explore interactions with such gestures, our robot acts as a host to a human visitor participating in a demo in a laboratory. The use of the Collagen™ system (Rich et al, 2001) to model conversation and collaboration permits the interaction to be more general and easily changed than techniques such as (Fong et al, 2001). One such conversation taken from a conversation log is shown in Appendix 1; it does not show the robot's or the human's gestures. There are many alternatives paths in the conversation that cannot be provided in a short space. The conversation concerns an invention, called IGlassware (a kind of electronic cup sitting on a table), that the robot and visitor demonstrate together. As the reader will notice, the robot's conversation are robot controlled in large part because when a more mixed initiative style is used, participants tend to produce many types of utterances, and speech recogntion becomes to unreliable for successful conversation.

Figure 1: Mel, the penguin robot

The robot is a penguin (see Figure 1) with a humanoid face (eyes facing forward and a beak that opens and closes), which we hypothesize is essential to allow human participants to assume familiarity with what the robot will at least say. We have not attempted yet to test this hypothesis as doing so would require experimenting with other non-humanoid models, which we are not equipped to do. The robot is a 7 DOF stationary robot. Details of the robot's sensory devices and the architecture it uses can be found in (Sidner et al, 2004b).

The penguin robot has been provided with gestural rules so that it can undertake the hosting conversations discussed previously. The robot has gestures for greeting a visitor, looking at the visitor and others during the demo, looking at the IGlass cup and table when pointing to it or discussing it, for ending the interaction, and for tracking the visitor when the visitor is speaking. The robot also interrupts its intended conversation about the demo, when the visitor does not take a turn at the expected point in the interaction. Failing to take a turn is an indication of the desire to disengage, and the robot queries the visitor about his/her desire to continue. Continuing lack of response or an answer indicating desire to end the demo will lead to a closing sequence on the robot's part.

Our robot attempts to keep its face on the human visitor when it speaks and listens except for when it looks at the cup and table. It does not expect that the human will look at it at all times for two reasons: (1) as our video data show, people do not do so with human partners; and (2) humans may pay less attention to the robot just because it is a robot and not a human partner. However, the robot does expect that the human partner will look at objects it

points out, and when the human does not do so, it prompts the human for a response.

3 Evaluating Human-Robot Interactions

Evaluating a robot's interactions is a non-trivial undertaking. In separate work (Sidner et al, 2004b), we have begun to explore both the success of the robot's behavior as well as the matter of what measures to use in order to accomplish such evaluations. We have evaluated 37 subjects in two conditions of interaction, one in which the robot has all the gestures we have been able to program (moving), and a second (talking) condition where the only movement is that of the robot's beak (after the robot locates the participant and locks onto the location of the participant's face, which it holds for the remainder of the interaction).

One of our challenges in that work was to decide how to measure the impact of the robot's behavior on the interaction. We used a questionaire given to participants after the demo with the robot to gather information about their liking of the robot, involvement in the demo, appropriateness of movements and predictabilty of robot behavior. However, we also studied the participant's behaviors from video data collected during the experiment. To further measure participant's engagement, we used interaction time, amount of mutual gaze, talk directed to the robot, overall looking back to the robot, and for two pointing behaviors, how closely in time the participant tracked the robot's pointing.

Does this robot's engagement gestural behavior have an impact on the human partner? The answer is a qualified yes. While details can be found in (Sidner et al, 2004b), in summary, a majority of participants in both conditions were found to turn their gaze to the robot whenever they took a turn in the conversation, an indication that the robot was real enough to be worthy of conversation. Furthermore, participants in the moving condition looked back at the robot significantly more whenever they were attending to the demonstration in front of them. The participants with the moving robot also responded to the robot's change of gaze to the table somewhat more than the other subjects.

Another gesture that is common in conversation is nodding, which serves at least the purpose of backchanneling and grounding (Clark, 1996). In collaboration with researchers at MIT, we are using the Watson system and a set of HMM algorithms to interpret hod nods from human participants (Lee et al, 2004).

Figure 2: Human participant with Mel

Most of our experiments with human participants (41 so far) have largely only provided us with further training data for the HMMs. As we have discovered, human head nodding is distinctive in conversation for being a very small motion (as little as 3 degrees), and one that is also very idosyncratic for different people. Our plan is to improve the recognition to the point that people's nodding will be recognized. In our first study (discussed above), we discovered that people naturally nod at the robot: 55% of the participants in the moving condition did so, while 45% in the talker condition, even though the participants had no reason to do believe the robot recognized this behavior. Our subsequent studies (where participants were told that the robot could recognize nods) show an even higher incidence of head nods as backchannels and accompanying "yes" answers to questions. We are currently using that data to explore new means of interpreting head nods in conversational contexts (Morency, this workshop).

4 Related Research

While other researchers in robotics have explored aspects of gesture (for example Breazeal (2001) and Kanda et al, 2002), none of them have attempted to model human-robot interaction to the degree that involves the numerous aspects of engagement and collaborative conversation that we have set out above. Recent work by Breazeal et al (2004) is exploring teaching a robot a physical task that can be performed collaboratively once learned. A robot developed at Carnegie Mellon University serves as a museum guide (Burgard et al, 1998) and navigates well while avoiding humans, but interacts with users via a 2D talking head with minimal engagement and conversational abilities. Most similar in spirit to work reported here is the Armar II robot (Dillman et al, 2004). Armar II is speech enabled, has some dialogue capabilities, and has abilities to track gestures and people. However, the Armar II work is focused on teaching the robot new tasks (with programming by demonstration techniques), while our work

has been focused on improving the interaction capabilities needed to hold conversations and undertake tasks. Recent work on head nodding and head tilts has been applied as part of two-person and oen robot conversation (Fujie et al, 2004).

Work on 2D conversational agents originally inspired our work (Cassell et al, 200, Johnson et al, 2000). Recent work by Traum and Rickel (2002) have explored issues in giving attention to agents in an interaction as part of an agent architecture in 2D virtual worlds.

5 Future Directions

Work has just been completed in mobilizing our previously stationary robot. In addition to continuing our observations of human-human data and better recognition of nodding, we intend to use mobilization for two problems: (1) the process of undertaking the engagement of another, in particular, to have to robot move toward a group of people or an individual, "catch their eye" and begin a conversation, (2) the use of body stance by the robot in interactions. Body stance is a means to indicate the primary focus of engagement. When body stance is required to be used to face an object other than the speaker (as for example, when washing the dishes), the individual must still turn either its body or its head back to the interlocutor at points in the interaction. Our intention is to understand how people do this and apply it to the robot's behavior.

6 Summary

Human-robot interaction offers new challenges in the use of interaction behaviors. This paper has discussed a type of human robot interaction aimed at hosting activities, and the need for engagement models to govern the connective relations between human and robot. The paper also discusses our investigations in developing and evaluating engagement rules and evaluations of human participants in interactions with a robot on a hosting problem.

Appendix 1: Sample conversation for the IGlassware demo

Agent says "Hello, I'm Mel, a robotic penguin."
User says "Hi Mel."
Agent says "What's your name?"
User says "Kathy."
Agent says "Kathy, I'd like to show you a demo. OK?"
User says "Ok."
Agent says "It would be really nice to know what kind of progress people are making in their dining. So Paul Dietz created a new product called I Glassware for this."
Agent says "By the way, Kathy, do you know Paul?"
User says "Yes."
Agent says "Then you probably know how creative he is."
Agent says "I Glassware stands for instrumented glassware!"
Agent says "Kathy, right there is the I Glassware cup."
User looks at cup.
Agent says "And near it, is the table readout."
User looks at readout.
Agent says "The word re-fill on the readout means that it is empty."
User says "Ok."
Agent says "All right, now we'll see how to use the cup."
User says "Alright."
Agent says "First you should pour enough water from the pitcher into the glass to fill it up. Then make sure the glass is on the IGlassware table."
User pours water into the glass from the pitcher. Then put the glass on the table.
Agent says "Good."
Agent says "Notice, the bar on the readout shows that the glass is full."
User says "Ok."
Agent says "Okay, pour the water back into the pitcher."
User says "Ok."
Agent says "I'm waiting for a reading from the table for the glass."
User pours the water from the glass back into the pitcher.
Agent says "Good."
Agent says "See, it registers needing a re-fill!"
User says "Great."
Agent says "Would you like me to explain how this works?"
User says "Yes."
Agent says "The copper in the glass transmits to the readout display by inductance with the surface of the table. The readout then displays the information coming to the table."
User says "Ok."
Agent says "Do you know how this technology might be used in restaurants?"
User says "Yes."
Agent says "Okay."
Agent says "Well, Kathy, that's about all. Go see Paul Dietz, for more about I Glassware. So long!"
User says "Good bye."
Agent performs ShutDown.

6 Acknowledgements

The authors wish to acknowledge the work of Charles Rich and Neal Lesh on aspects of Collagen and Mel, and Max Makeev for mobilizing Mel.

References

C. Breazeal. Affective interaction between humans and robots. *Proceedings of the 2001 European Confer-*

ence on Artificial Life (ECAL2001). Prague, Czech Republic, (2001).

C. Breazeal, G. Hoffman and A. Lockerd. Teaching and Working with Robots as a Collaboration. *The Third International Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004*, pp.1028-1035, ACM Press, 2004.

W. Burgard, A.B. Cremes, D. Fox, D. Haehnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun. The Interactive Museum Tour Guide Robot. *Proceedings of American Association of Artificial Intelligence Conference 1998*, 11-18, AAAI Press, Menlo Park, CA, 1998.

J. Cassell, J. Sullivan, S. Prevost and E. Churchill. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

H. H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

R. Dillman, R. Becher and P. Steinhaus. ARMAR II-- A learning and Cooperative Multimodal Humanoid Robot System. *International Journal of Humanoid Robotics*, Vol 1:1, pp 14155, 2004.

T. Fong, C. Thorpe, C. Baur. Collaboration, Dialogue and Human-Robot Interaction, *10th International Symposium of Robotics Research*, Lorne, Victoria, Australia, November, 2001.

S. Fujie, Y. Ejirir, K. Nakajima, Y. Matsusaka, and T. Kobayashi. A Conversation Robot Using Head Gesture Recognition as Para-Linguistic Information. *Proceedings of RO-MAN, the 13th IEEE International Workshop on Robot and Human Interactive Communication*, 2004.

W.L. Johnson, J. W. Rickel, J. W. and J.C. Lester. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education*, 11: 47-78, 2000.

T. Kanda, H. Ishiguro, M. Imai. T. Ono, and K. Mase. A constructive approach for developing interactive humanoid robots. *Proceedings of IROS 2002*, IEEE Press, NY, 2002.

A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26: 22-63, 1967.

C. Lee, N. Lesh, C. Sidner, L. Morency, A. Kapoor, T.Darrell. Nodding in Conversations with a Robot. *Proceedings of the ACM International Conference Human Factors in Computing Systems*, ACM Press, 2004.

L.P. Morency, C. Sidner and T. Darrell. Towards Context Based Vision Feedback Recognition for Embodied Agents, (this workshop) 2004.

Y. Nakano, G. Reinstein, T. and J. Cassell. Towards a Model of Face-to-Face Grounding. *Proceedings of the 41st meeting of the Association for Computational Linguistics,* pp. 553-561, ACL Press, 2003.

C. Rich, C.L. Sidner, and N. Lesh. COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction. *AI Magazine, Special Issue on Intelligent User Interfaces*, AAAI Press, Menlo Park, CA, Vol. 22: 4: 15-25, 2001.

C. Sidner, C.Lee, C.Kidd, N. Lesh. Explorations in Engagement for Humans and Robots. *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2004)*, IEEE Press, 2004b.

C.L. Sidner, C.D. Kidd, C.H. Lee, and N. Lesh. Where to look: A study of human-robot engagement. *ACM International Conference on Intelligent User Interfaces (IUI)*, ACM, pp. 78—84, January 2004b.

C.L. Sidner, C.H. Lee, and N. Lesh. Engagement when looking: behaviors for robots when collaborating with people. *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue* (I.~Kruiff-Korbayova and C.Kosny, eds.), University of Saarland, 2003, pp. 123--130.

D. Traum and J. Rickel. Embodied Agents for Multi-party Dialogue in Immersive Virtual World. *Proceedings of the International Joint Conference on Autonomous Agents and Multi-agent Systems*, 2002, pp. 766-773, 2002.

A Two-layered Approach to Make Human-Robot Interaction Social and Robust

Yong XU [*] Takashi TAJIMA[†] Makoto HATAKEYAMA[†]
Yasuyuki SUMI[*] Toyoaki NISHIDA[*]
[*]Graduate School of Informatics, Kyoto University
606-8501 Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
xuyong@ii.ist.i.kyoto-u.ac.jp
[†]Graduate School of Info. Sci. and Tec. , the University of Tokyo
113-8656 Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
tajima@kc.t.u-tokyo.ac.jp
hatake@kc.t.u-tokyo.ac.jp sumi@i.kyoto-u.ac.jp nishida@i.kyoto-u.ac.jp

Abstract

The capacity of involvement and engagement plays an important role in making a robot social and robust. In order to reinforce the capacity of robot in human-robot interaction, we proposed a two-layered approach. In the upper layer, social interaction is flexibly controlled by Bayesian Net using social interaction patterns. In the lower layer, the robustness of the system can be improved by detecting repetitive and rhythmic gestures. To verify our approach, we applied and tested different methods on separate problems respectively. We use the application of waiter robot as the proof of concept of our approach. We assert that it is a promising solution to the problem of enforcing involvement and engagement in human-robot interaction.

1 Introduction

The field of robotics is changing at an unprecedented pace and varying from traditional industrial robot to entertainment robot. According to the survey of United Nations, the robotics can be grouped into three major categories: industrial robotics, professional service robotics, and personal service robotics. Personal service robots have the highest expected growth rate. They are estimated to grow from 176,500 in 2001 to 2,021,000 in 2005(UN, 2002). Many personal service robots assist people directly in domestic and institutional settings. More and more robots can interact with people without special skills or training to operate the robot. Obviously the human-robot interaction will become more and more important in recent future.

Bartneck defined social robot in (Christoph Bartneck, 2004) as "an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioural norms expected by the people with whom the robot is intended to interact." Communication and interaction with human is a critical point in this definition. Being social is bound to understanding and, in some cases, mimicking human activity, the surrounding society and culture, which shapes social values, norms and standards. If the robot can understand and follow the social norms when it interacts with people, it may positively help the communication between human and robot to be performed in more smooth and natural way. For example, a greeting can be divided into three primary phases: distance salutation, approach and close salutation. If it happens between two people, the two may nod or wave hand in distance before approaching to each other, and say hello after closing enough. We may also have the experience of being frightened when someone strange approaching and greeting without distance salutation. Exactly some social norms seem hidden in the order of behaviour sequences. As social norms can be defined by the interactions between human beings, we assert that they can be also defined by the interactions between human and robot. If the robot can follow the social norms, it seems helpful to make the verbal communication (e.g. conversation) and nonverbal communication (e.g. gesture) between human and robot more easy and natural. We assert that it is necessary for the robot to achieve involvement in the conversation between human and robot.

There are two major challenges in field of human-robot interaction (Monica, 2001). The first is to build robots that have the ability to learn through social interaction with humans or with other robots

in the environment. The second challenge is to design robots that exhibit social behavior, which allows them to engage in various types of interactions. In this paper we focus on the latter and present an approach trying to provide a promising solution to the issue of how to apply behaviour patterns concluded from human's daily life communication to improve the involvement capacity of robot, so that the robot can communicate and interact with human in more effective and natural way.

The main contributions of this paper include the proposition of a novel two-layered architecture of human-robot interaction system. The proposed architecture can not only make the human-robot interaction system social by taking use of social interaction pattern through Bayesian Network, but also make the system robust by detecting rhythmic repetitive gesture through building dynamical system method. In this paper, the waiter robot implemented in part is used as an example of the robot which is expected for robust and social interaction. The waiter robot can respond for requests from customers, offer drink and remove empty cups etc. in the situation such as a party hall. Such a robot has to be able to recognize various agreements of human society. For example, if a customer raises his hand or waving his arm, it may means that he has some requests to call the waiter, and the robot should be able to recognize the movement and give some appropriate response. A robust waiter robot should also be able to distinguish the requesting movements from those when customer simply raises his hand without anything requests. This paper shows how to recognize human's social interaction using Bayesian Net, and how to robustly understand and recognize human's repetitive gesture using an entrainment principle.

The remainder of the paper is organized as follows. Section 2 explains why social interaction pattern is needed to create involvement and engagement in human-robot interaction. Section 3 introduces the architecture of the two-layered approach. Section 4 describes how the system works. Section 5 draws the conclusions and outlines directions of future work.

2 Social Interaction Pattern and Robustness to Create Involvement and Engagement in Human-robot Interaction

2.1 Social norm and gesture

In order to make the robot naturally involve in or engage in the communication between human and robot, it seems necessary for the robot to be able to follow social norms of human society. In order to enable natural human-robot communication, we suggest conceive the robot as social agent which has ability of mutual understanding at some level. So it is also necessary for the robot to understand and follow the social norms of human society.

Social norm means some social behaviour patterns which can be expressed by some behavior sequences. As long as we determine the specific situation, we can conclude some general behaviour patterns which we called "schema".

A schema is a probabilistic representation of social interaction patterns which is concluded according to social norms in human society. In different situations, the behaviours in schema often occur in different order and some also repeat for many times.

The communication behaviours are often classified into verbal mode and nonverbal mode. The former is mainly based on verbal modality using natural or man-made language. And the latter focuses on nonverbal modality using bodily language such as facial expression, gesture and posture. Roughly speaking, the verbal modality emphasizes explicit side of user's intention and the nonverbal modality emphasizes more on implicit side of intention.

We argue that the nonverbal modality plays more important role when a human wants to convey his tacit intention to others. The gesture is one kind of mostly used nonverbal movements. We pay more attention on repetitive gesture, because it is very natural that people often prefers to repeat his movement when the communication partner can not understand him. We believe the rhythm information hidden in the repetitive gesture may contain some important information. It may also reflect the tacit intention of human at some extent. Among the tacit gesture, the repetitive gesture can be easily and certainly caught, we make it a key to realize the robust robot.

2.2 Making Human-Robot Interaction Robust

A key to achieve a robust architecture for robot is using a representation of rhythmic and repetitive gesture. When a user wants to convey his intention to others by nonverbal way, it is natural to use repetitive gesture. In our opinion, the repetitive gesture can be categorized as follows:

(1)Attracting attention: e.g. waving.

(2)Direction indication: e.g. leftward, rightward, forward, backward, upward, downward.

(3)Movement indication: e.g. come here, get out.

(4)Speed indication: e.g. slow-down, hurry up etc.

Many people may have the experience of helping the driver to back the car into a narrow parking area by direction indication gesture (backward, forward, etc.). In this situation, it is obviously more

easy and natural to convey one's intention to the partner by gesture than by word.

Additionally, taking advantage of rhythmic information will be helpful to decrease noisy data and improve the robustness of the interaction. Observations in various research areas suggest that human communicational behaviour is significantly rhythmic in nature, for instance, in the way how words are grouped in time (speech rhythm) or how they are accompanied by body movements (gestures). Animals and humans exhibit many kinds of behavior where the frequencies of gestures are related by small integer ratios (like 1:1, 2:1 or 3:1). Especially if the gestures continue for a while, gestures often tend to cycle in the ratio 1:1 or 1:n or m:n (where m and n are very small integers). (Robert Port et al. ,1998). For example, most joggers notice that during steady-state jogging, one's breathing tends to lock into a fixed relationship with the step cycle with two or three steps to each breath cycle, or perhaps three steps to twice breaths. In (Wachsmuth, 2002) "rhythm" is defined as relative timing between adjacent and nonadjacent elements in a behaviour sequence, i.e., the locus of each element along the time line is determined relative to the locus of all other elements in the sequence. The "gestures" can be understood as body movements which convey information that is meaningful in some way to a recipient. There is evidence that communication among humans is strikingly rhythmic in nature. When this is true, then this observation should also be relevant in human-robot communication.

The mutual drawing-in phenomenon (sometimes it is also called entrainment) is the mediation between the system of human being and the system of robot. The drawing-in phenomenon makes the loop of human-robot interaction. Suppose the state of human to be x, the state of robot to be y, the rhythm of human's repetitive gesture at the beginning of communication to be f(x), the rhythm of robot to be g(y). So the systems of human and robot will be as follows:

$$\text{Human}: dx = f(x) + h(x,y) \quad (1)$$

$$\text{Robot}: dy = g(y) + r(x,y) \quad (2)$$

Here the part of h(x y) and r(x y) will bear the task of interaction between human and robot. Human's rhythm of operation and robot's rhythm of operation are drawn to some stable states by h(x y) and r(x y).

Considering the most simple condition, suppose $h(x,y) = r(x,y) = 0$, and the f(x) and g(y) takes a simple formation as follows:

$$f(x) = A * \mathrm{Cos}(\alpha * x + \theta) \quad (3)$$

$$g(y) = B * \mathrm{Cos}(\beta * y + \gamma) \quad (4)$$

For the convenience of describing the rhythm, we define one evaluation index named "attractiveness degree" (abbr. as AD) as $AD = \frac{f(x)}{g(y)}$.

If α equals to β, and θ equals to γ, the AD will be A/B, so as long as A times B or B times A, AD will be some integer, so human and robot can synchronize with each other.

If α equals to β, and θ does not equal to γ but both keep constant, as long as A times B or B times A, according to intuitive experience, the two systems will synchronize with a time delay.

If α does not equal to β, even if θ equal to γ and A equals to B, the two systems will not synchronize.

Therefore, the α and β seem playing important role here. As we know, most functions can be expressed approximately by some combination of groups of some cosine or sine functions. If we can work out the average frequencies or frequency related parameters of the repetitive gesture, relation between rhythm of human and robot may be calculated successfully.

In real world applications, the f(x), g(x), h(x,y) and r(x,y) may take different formations and should be defined according to specific situations respectively.

Although the meaning of repetitive and rhythmic gesture maybe exist different explanation by different people at different situations, we believe that there must be some common meaning (e.g. urgency) can be expressed by some features (such as amplitude, frequency, etc.) and can be understood by different people in some specific situation. Certainly there might also be some kinds of culture-dependent repetitive gestures, for instance, the repetitive "come here" gesture. We argue that repetitive gesture can improve the robustness of human-robot interaction. What is more, there are some other reasons why we choose repetitive gesture. Firstly, it is easy to extract closed orbit from human's gesture, so that period and frequency can be extracted more easily. Secondly, it is easy for computer to process the periodic data. Because the computer needs not distinguish where the starting point is and where the ending point is. What is more, it is also helpful to decrease influence of noisy data when processing the periodic data by calculating average values of different period data.

3 The Architecture of the System

3.1 Two-layered Approach

In this section, we will introduce the details about the architecture of our approach. The architecture of

the system is shown in Fig 1.

As shown in figure, the robot system accepts nonverbal information from human by input devices and human can see robot's reaction by observing its movements. There are two layers in our proposed robot system: the upper layer and the low layer. The lower layer is in charge of event detection and the upper layer coordinates the behavior of robot by handling the evidences provided by the lower layer.

Next we will explain the details of how the information flows in the system. At first, the data of human's gesture should be input into the system by some kinds of input devices (such as motion capture). Then the perception module will perform the pre-processing task and transfer control to autonomous module when meeting some urgent situations (e.g. sonar data of being collided with obstacle) or some predefined actions (e.g. stop gesture) are input. The information about gaze and distance will be processed by this module too. After pre-processing, the recognition module will save the interaction patterns into database and dispatch result to processing modules (including probability module, synchronization/modulation module and autonomous module). If human's behavior can be matched with some known interaction patterns, i.e. schemata, the probability module which adopts schema based interaction approach (see next section) will calculate the probability variables according to Bayesian Net, and decide what kind of behaviour the robot should choose. Then the result will be transferred to motion selection module to generate detail motor control commands. The module will also save often-used motions into motion behavior database to facilitate the motion selection in the future. If the gesture is unknown and the gesture sequences are recognized as repetitive gesture, the "attractiveness degree" will be calculated and synchronization/modulation module which adopts entrainment based interaction approach (see next section) will be used to create dynamical systems for robot and human, so that the robot can work out the reaction behaviour to be performed. If the gesture can not been recognized, the autonomous module will take over the control and choose the next movement for the robot.

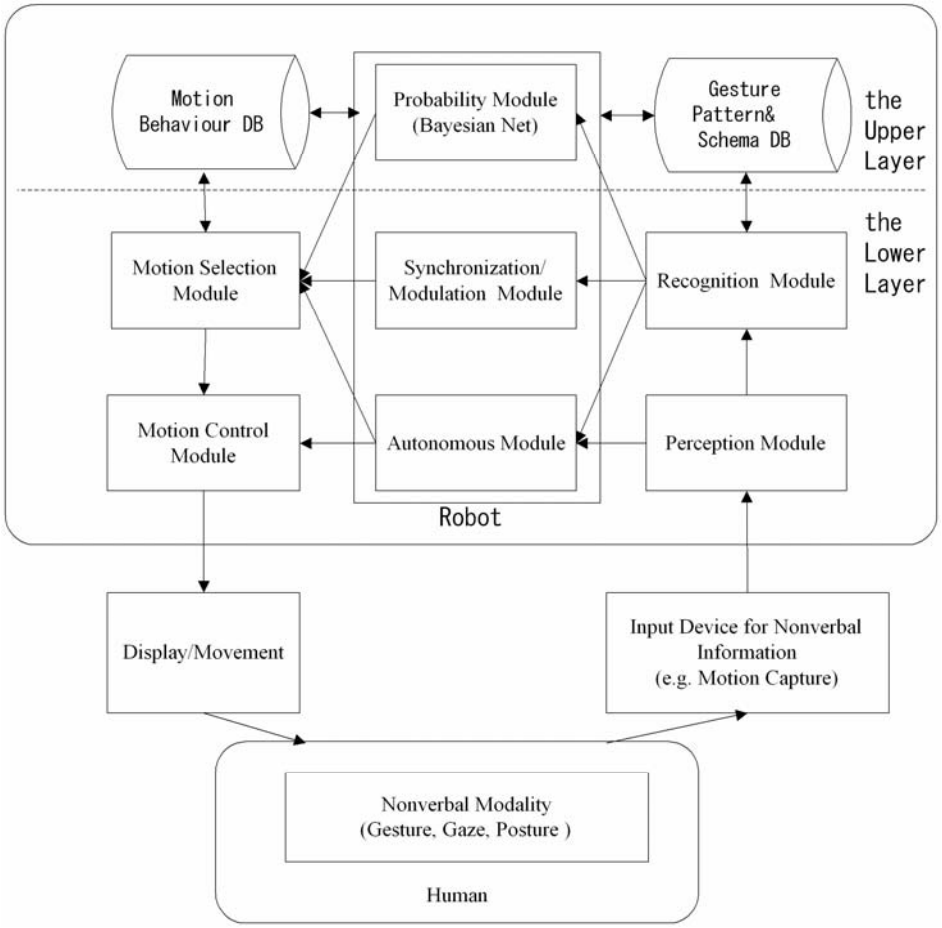


Figure 1: Architecture of Two-layered Human-Robot Interaction System

The human can watch the reaction of the robot and repeats or changes his gesture, so that the robot can improve the interaction with him and his intention can be conveyed to robot in better way. Through this way, the communication between human and robot can proceed in more smooth and natural way.

One of predominant features of the synchronization/modulation module locates at usage of repetitive gesture. The repetitive gesture may convey human user's intention in most cases during communication and interaction between human-human or human-robot. And the probability model (schema based interaction using Bayesian Network) is a reasonable method to take usage of the social norm information hidden in gesture patterns. The dynamical system method (entrainment based interaction) can represent the repetitive gesture in a more basic and natural way.

In the system, the lower layer is controlled by the upper layer. Meanwhile the lower layer provides evidences to the upper layer.

3.2 The Upper Layer: Schema Based Interaction

The conventional research about communication between human and robot usually presumed that both sides already understand the purpose of communication. It can be said that "the atmosphere of communication" has been already established. However, we must premise on the state before the beginning of communication, in that the purpose of communication in the real world is various. In the state when communication has not yet started, human usually can not convey his intention by verbal communication method, such as conversation. Therefore, it is necessary for the robot to infer the tacit intention from the interaction loop by watching human's nonverbal information and giving some reactions. Basing on this loop, the communication atmosphere can be established and the next steps, such as conversation can proceed on smoothly and naturally.

Considering the example of waiter robot, information necessary for the waiter robot to establish communication process includes: inter-personal distance, direction of gaze and acknowledge movement (ACK), such as raising the hand, nodding, etc. In the case of communication between human beings, there are many social habits to follow. We can conclude some patterns from this kind of social habits, and call them interaction schema. We believe that the interaction schema may help the interaction between human and robot.

Interaction schema describes matching of typical operation of human and robot (HATAKEYAMA 2004). The robot infers the human user's intention and environment situation according to the interaction schema, so that the robot can decide its next movement. Moreover, the flow of communication can be designed by putting some schemata in order as a sequence. By being designed appropriately, state of schema can change in accordance with situations or the purposes, so that interaction loop can be constituted and communication atmosphere can be established.

Considering the instance of waiter robot which can provide service to human in a party, we can get some ideas about what is schema based interaction approach. Fig. 2 shows one example of interaction schema. Before the beginning of communication, it is necessary for the waiter robot to collect the information includes: inter-personal distance, gaze direction and acknowledge (ACK) action. In addition, the history record of human-robot interaction and current action of robot are also needed for the robot to decide the next action. By using schema based interaction, the communication atmosphere can be established, so the human and robot can begin to communicate with each other. For example, if the robot senses a human at far distance who is waving and keeps waving when the robot approaching to the human or waves its hand toward the human, it is obviously that the human may want to communicate with the robot. When the both approach to each other and near enough, they may gaze at each other or start to say hello to each other. By the similar method, the robot can recognize the human's intention by detecting other similar repetitive gesture or action of human. The repetitive gesture can be processed by one novel kind of interaction method named entrainment based interaction which we will introduce in next section.

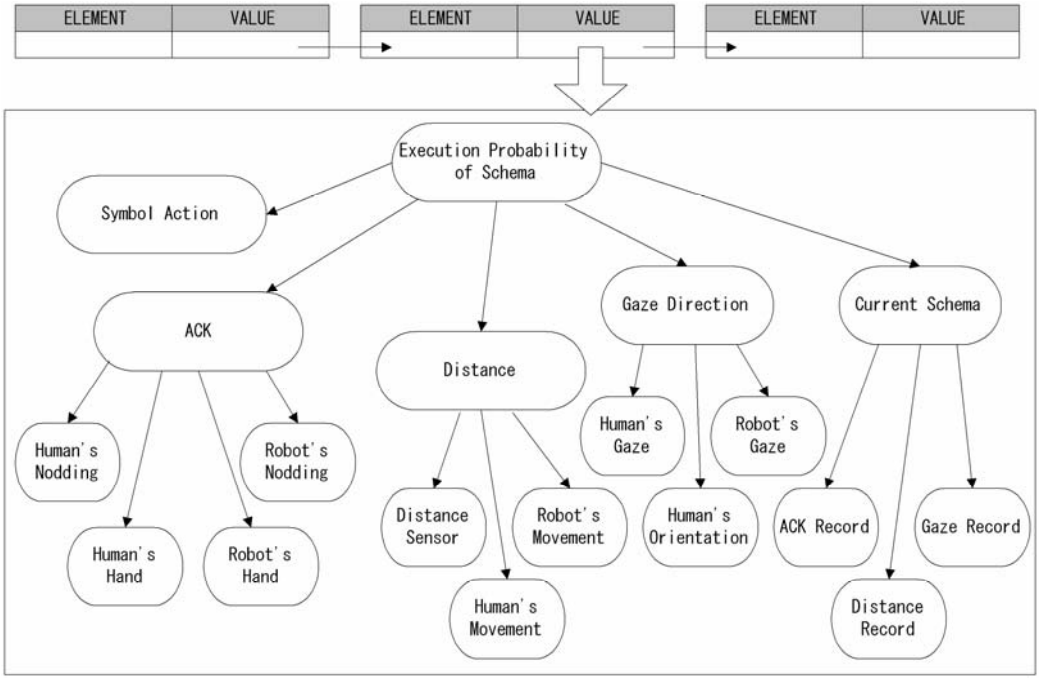


Figure 2: Interaction Schema

3.3 The Lower Layer: Entrainment Based Interaction

In order to explain what is the "entrainment based interaction", we will introduce the concept entrainment in advance. The "entrainment" in this paper is defined as a kind of phenomenon that the rhythm of human user's gestures and those of the robot become synchronized during the human-robot interaction. User's tacit intention can be conveyed to robot by repetitive gestures.

The entrainment based interaction can be classified into two kinds: human-oriented entrainment and environment-oriented entrainment. The former focused on the synchronization phenomena usually happened during the conversation between human beings. For example, the listener often unconsciously synchronizes the rhythm of his voice, posture and gesture (bodily movement) with those of the speaker, so that the conversation between the two people can proceed smoothly and naturally. The project "E-COSMIC" led by Watanabe developed InterRobot (OKADA 2001) using facial expression, posture and gesture to support the remote communication between human beings. In this case, how to match the voice and the bodily movement became important. The later, environment-oriented entrainment, focused on how to make the human or robot adapt to the changing environment easily and naturally. The typical research topic is about bipedal movement for humanoid robot (OKADA 2001). In this field the dynamical system is widely used to model the environment. Our research tried to model the human-robot interaction by using dynamical system, which was seldom used to model the human-oriented entrainment before. Since the complicated structure of human beings, it seems quite difficult to model the rhythm of human's movements. In our system, we used method of building dynamical system for human and robot respectively, and by using synchronization and modulation in entrainment based interaction approach, the two systems can interact with each other easily. Therefore the human can convey his intention to robot in more natural way.

The outline of entrainment based interaction is shown in Fig. 3(TAJIMA 2004). At first the human user instructs the robot by repetitive gestures with his hand. The robot extracts orbit from gestures and builds a dynamical system converging to the orbit in low dimensional space by "synchronization".

Then robot can operate in high dimensional space according to dynamical system. After synchronization, the user can change the rhythm of his gesture while observing reactions of the robot. The robot will modify the dynamical system to converge to new orbit of human's gestures. Modification of the dynamical system can be conducted by transforming the linear function used in old dynamical system. We call it "modulation". Then the robot will operate according to the new dynamical system. By this way, the human user can convey his tacit intentions to the robot by entrainment based interaction.

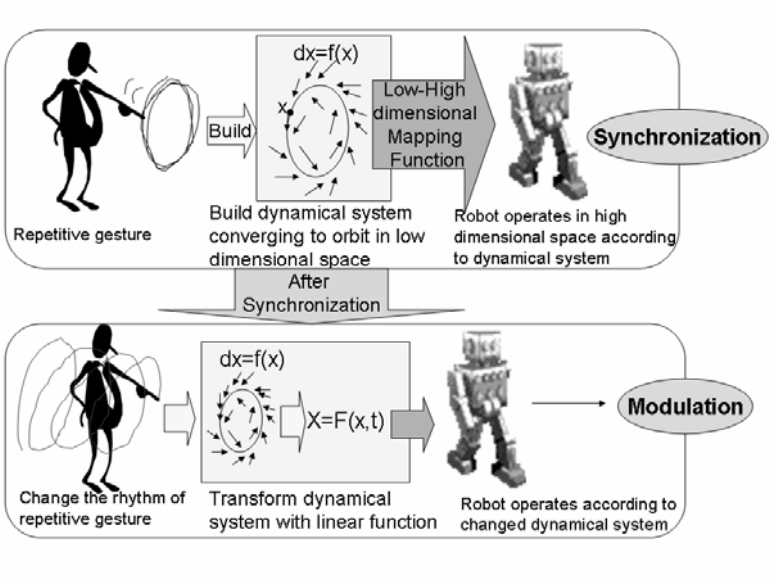


Figure 3: Outline of Entrainment Based Interaction

4 How the System Works?

4.1 Application Instance: Waiter Robot System

In this section, we will explain briefly about how to implement a human-robot interaction system basing on our proposed architecture. In order to explain in an easy-to-understand way, we use the instance of waiter robot and set the party field as our application background where waiter robot provides service for the human customers. We take an instance of schema sequence as our example, which is constituted of some interaction schemata. In this case, we set three tasks: asking for seconds, asking for removing empty cups, approach to robot to get drinks. And the procedure is as follows:

-Communication partners look at each other
-Raise and waving hand at distance
-Approach to partner
-Begin to communicate with partner

Exactly in more complicated situations, there are a lot of behaviours that are available to be chosen. As long as the probability of each state is calculated, it is not difficult for the robot to decide the next action by choosing the state with high probability value.

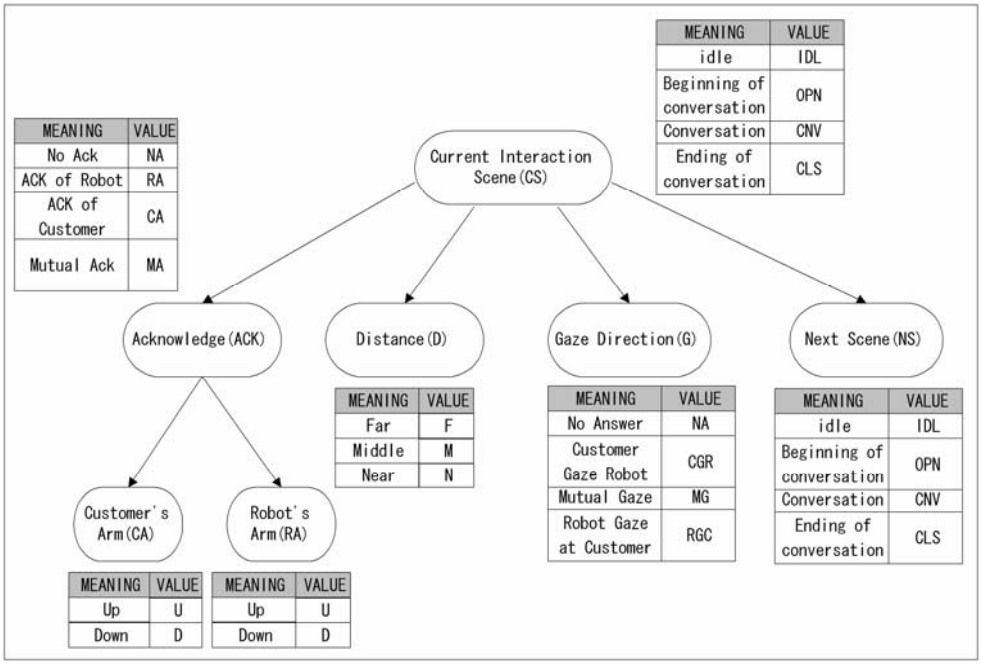


Figure 4: Schema Based Interaction using Bayesian Network

4.2 Upper Layer: Establishment of Communication

Interaction schema can be implemented by Bayesian Network. To decide next scene to transit to, the system will calculate the probability by Bayesian network from human's action, robot's action, symbol action and action records. According to Bayes' Theorem $P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)}$, the posterior probability P(H|E) means the probability of hypothesis H after considering the effect of evidence E, likelihood P(E|H) gives the probability of evidence E assuming the hypothesis H is true, the term P(H) gives prior probability of hypothesis H, and last term P(E) is independent of H and can be regarded as a normalizing factor. One example of Bayesian network is shown in Fig. 4. The probability variable is represented by elliptic nodes, and the causality is represented by edge connecting nodes. The table behind the node list probability variable's possible values of the node.

The causality between stochastic variables is represented by conditional probability. For example, from the value of variable "ACK", the variables "Customer's Arm(CA)" and "Robot's Arm(RA)" can be determined by two matrices as follows.

$$ACK \begin{array}{c} \\ NA \\ RA \\ CA \\ MA \end{array} \begin{array}{c} CA \\ D \quad U \\ \begin{bmatrix} 0.85 & 0.15 \\ 0.85 & 0.15 \\ 0.15 & 0.85 \\ 0.15 & 0.85 \end{bmatrix} \end{array} \qquad ACK \begin{array}{c} \\ NA \\ RA \\ CA \\ MA \end{array} \begin{array}{c} RA \\ D \quad U \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \\ 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix} \end{array}$$

The left matrix expresses the probability P(CA|ACK) and the right matrix expresses P(RA|ACK). The P(CA|ACK) expresses the probabilities of CA, i.e. human customer lifts his arm(CA=U) or puts down his hand(CA=D) in various possible conditions of ACK including NA(no acknowledge), RA(acknowledge of Robot), CA(acknowledge of customer) and MA(mutual acknowledge). Similarly, other causalities can be expressed by matrices as follows. For example, P(CA=D|ACK=NA)=0.85,P(CA=U|ACK=NA)=0.15, P(CA=D|ACK=RA)=0.85, etc. Here the P(CA=D|ACK=NA)=0.85 means that when there are not any acknowledge, the customer will put down his hand at the probability of 0.85.

$$CS \begin{array}{c} \\ IDL \\ OPN \\ CNV \\ CLS \end{array} \begin{array}{c} ACK \\ NA \quad RA \quad CA \quad MA \\ \begin{bmatrix} 0.65 & 0.15 & 0.15 & 0.05 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0.05 & 0.15 & 0.1 & 0.7 \\ 0.3 & 0.2 & 0.1 & 0.4 \end{bmatrix} \end{array}$$

$$CS \begin{array}{c} \\ IDL \\ OPN \\ CNV \\ CLS \end{array} \begin{array}{c} D \\ F \quad M \quad N \\ \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.5 & 0.4 & 0.1 \\ 0.05 & 0.15 & 0.8 \\ 0.05 & 0.15 & 0.8 \end{bmatrix} \end{array}$$

$$CS \begin{array}{c} \\ IDL \\ OPN \\ CNV \\ CLS \end{array} \begin{array}{c} G \\ NA \quad CGR \quad RGC \quad MG \\ \begin{bmatrix} 0.45 & 0.3 & 0.2 & 0.05 \\ 0.05 & 0.4 & 0.4 & 0.15 \\ 0.1 & 0.2 & 0.2 & 0.5 \\ 0.1 & 0.3 & 0.3 & 0.3 \end{bmatrix} \end{array}$$

$$CS \begin{array}{c} \\ IDL \\ OPN \\ CNV \\ CLS \end{array} \begin{array}{c} NS \\ IDL \quad OPN \quad CNV \quad CLS \\ \begin{bmatrix} 0.5 & 0.4 & 0.05 & 0.05 \\ 0.3 & 0.15 & 0.5 & 0.05 \\ 0.1 & 0.05 & 0.4 & 0.45 \\ 0.7 & 0.2 & 0.05 & 0.05 \end{bmatrix} \end{array}$$

Given the values of a priori probability P(CS),

$$\begin{cases} P(CS = IDL) = 0.5 \\ P(CS = OPN) = 0.1 \\ P(CS = CNV) = 0.3 \\ P(CS = CLS) = 0.1 \end{cases}$$

We can work out the value of P(ACK=NA):

P(ACK=NA)= P(ACK=NA|CS=IDL)*P(CS=IDL)
+ P(ACK=NA|CS=OPN)*P(CS=OPN)
+ P(ACK=NA|CS=CNV)*P(CS=CNV)
+ P(ACK=NA|CS=CLS)*P(CS=CLS)
=0.65*0.5+0.1*0.1+0.05*0.3+0.3*0.1
=0.38

Similarly, we can get other results as follows.

P(ACK=RA)=0.18,P(ACK=CA)=0.155,
P(ACK=MA)=0.285,
P(CA=D)=0.542,P(CA=U)=0.458

By using Bayesian Network, the robot system can decide what to do in the next step according to probability value through probabilistic reasoning from incomplete noise-containing sensor data. Through the methods mentioned above, the robot can involve in the communication with human following some social-norm like patterns. And the communication atmosphere can be established to help the interaction between human and robot proceed smoothly.

4.3 Lower Layer: Detection of Repetitive Gesture

The key point of the lower layer in our proposed architecture is to detect repetitive gesture by building a dynamical system through synchronization

and/or modulation. The dynamical system can be built through following steps.

Firstly, the human user's repetitive gestures can be input by motion capture with position sensors fixed on his hand(s). Then system will normalize the data after eliminating the direct-current elements.

Secondly, position data series can be represented by a series of s-dimension vector x. With auto-correlation function, period T can be extracted from data series. So that one period of data can be extracted.

Next, we use low-pass filter to extract low frequency elements, we called them attractors here. Specifically, we use DFT(Discrete Fourier Transfer) to filter out high frequency noise data and use IDFT(Inverse DFT) to get the equation x=A(t) which can represent the orbit with period T.

At last, a dynamical system can be constructed from the orbit using polynomial expression (OKADA M. et al., 2002).

Next, we will introduce the detail of how to build dynamical system from the orbit in three conditions.

In the first condition, when the point x(t) is on the orbit, say point x(t), the moving vector dx is defined as

$$\mathrm{dx} = \mathrm{x}(t+1) - \mathrm{x}(t) \tag{5}$$

so that the points near the orbit can converge to the orbit.

In the second condition, when the point x(t) is near the orbit, the nearest point on the orbit is defined as $\mathrm{x_m}(t)$, the moving vector dx can be expressed as following equations.

$$\mathrm{dx} = \mathrm{x}_m(t+1) - \mathrm{x}_m(t) + al^2\{\mathrm{x}_m(t) - \mathrm{x}(t)\} \tag{6}$$

$$l = \sqrt{\{\mathrm{x_m}(t) - \mathrm{x}(t)\}^2} \tag{7}$$

The coefficient a in (6) is a positive constant, and dx can be expressed as a N-dimensional polynomial expression as (8). The term l in (6) and (7) means the modulus of vector $\{\mathrm{x}_m(t) - \mathrm{x}(t)\}$. $a_{p_1 p_2 \cdots p_s}$ can be inferred from (8) by least square method.

$$\mathrm{d}x = \sum_{n=0}^{\mathrm{N}} \sum_{\substack{p_1 + p_2 + \ldots + p_s = n \\ p_k \text{ is positive integer}}} a_{p_1 p_2 \cdots p_s} x_1^{p_1} x_2^{p_2} \ldots x_s^{p_s} \tag{8}$$

In the third condition, when the point $\mathrm{x}(t)$ is far from the orbit, we define a dynamical system using other equations, for example, $\mathrm{dx} = -kx + c(k > 0)$, so that it can attract to the center point c of the circle orbit. Through the approaches noted above, a dynamical system $\mathrm{dx} = \mathrm{g}(x)$ can be achieved.

After finished building a dynamical system by synchronization, it is easy to build a dynamical system by modulation. We only need to modify the linear function f(x) used in dynamical system obtained from synchronization. A new function expression can be created by adding new elements such as time t to the dynamical system. Function f(x) can be obtained by comparing the equation of the orbit with the series of position sensor data. That is to say, function f(x), such as $\mathrm{f}(x) = ax + b$, (a and b are both constant) can be defined in advance and the coefficients of function f(x) can be obtained by the least square method. Different applications may have different forms of function f(x).

4.4 Integration of Two Layers

The system can work well only if we can integrate the two layers successfully. As we noted before, the lower layer can detect repetitive gesture and other nonverbal information, and send results to the upper layer, so the upper layer can make decision basing on Bayesian Network through selecting appropriate schema.

In the situation of party with a crowd of people, if one customer wants to draw attention of other people, he may keep waving his hands for a while. Because the moving object seems much easy to be seen. Suppose this kind of behaviour is defined as social norm, and saved as social interaction pattern in form of schema in our system. The low layer of the system can achieve the data from input devices, and the waving gesture can be detected as repetitive gesture by synchronization/modulation module, so that the probability module in upper layer can calculate the probability value for the behaviour of customer. At last the robot can give a reaction according to the result of probability. Following this way, other repetitive gestures can also be processed. For example, when the human customer wants to ask the robot to approach him quickly, he may repeat to beckon his hands for many times at high frequency. By detecting event including repetitive gesture, gazing and nodding, the lower layer can provide something about the orbit, the rhythm of the gesture and other movements of human to help the upper layer to choose more suitable reaction.

5 Conclusion and Future Work

In this paper, we proposed a novel two-layered human-robot interaction approach. Using the application instance of waiter robot, we provide the proof of the concept of the proposed system. We believe that the system can make the human-robot interaction social and robust. Through Bayesian network method, we tried to make the robot follow social interaction patterns by using schema which is generated according to social norms in human society. Meanwhile we improve the robustness of the robot system by detecting rhythmic repetitive gesture

through entrainment method which builds dynamical system for the repetitive gesture.

Except for the example of greeting and waiter robot, the proposed approach can also be applied in some other practical situations. For example, when a man wants to help the other (driver) to park the car into a narrow parking area, the man often prefers communicating with the other by repetitive gesture rather than using language. When a man wants to instructs what the other should do or what direction should move in a noisy environment where voice does not work for communication, the gesture seems one of the best choices to convey intentions.

However we only provided the proof of concept of our approach, the effectiveness of the approach had not been proved enough. Therefore we plan to implement whole or part of the system to achieve some practical results. We assume that the capacity of mutual adaptation and mutual adjustment may act important role in human-robot interaction. We also plan to give further research on this topic.

References

CHRISTOPH Bartneck, Jodi Forlizzi , A Design-Centred Framework for Social Human-Robot Interaction. *Proceedings of RoMan 2004, Kurashik,* 2004.

HATAKEYAMA M., Human-Robot Interaction based on Interaction Schema (Master Thesis), *University of Tokyo, Japan* ,2004.

MONICA N. Nicolescu and Maja J Mataric**,** Learning and Interacting in Human-Robot Domains, *Technical Report Institute for Robotics ad Intelligent Systems Technical Report IRIS01-395,University of Southen California,* 2001.

OKADA M., YASUMURA M., SASAKI M. (eds.), Embodiment and computer, *KYORITSU Publication, Japan,* 246-256, 2001.

OKADA M., TATANI K., NAKAMURA Y.**,** Polynomial design of the nonlinear dynamics for the brain-like information processing of whole body motion, *Proc. of IEEE International Conference on Robotics and Automation(ICRA2002), Washington D.C., U.S.A.* 2:1410-1415,2002.

ROBERT Port, Tajima,K.&Cummins,F. Speech and Rhythmic Behavior, *In The Nonlinear Analysis of Developmental Processes(G. J. P. Savelsburgh, H. van der Maas, & P. C. L. van Geert, editors), Royal Dutch Academy of Arts and Sciences,*1998.

TAJIMA Takashi., XU Yong and Nishida Toyoaki., Entrianment Based Human-Agent Interaction, *Proc. of IEEE Conf. on Robotics, Automation and Mechatronics(RAM2004), Singapore,*1042-1047, 2004.

UN, United Nations and the international Federation of Robotcis. *World Robotcis 2002, nternational Journal of Computer Vision,UN,New York,* 2002.

WACHSMUTH I., Communicative Rhythm in Gesture and Speech, *In P. Mc Kevitt, S. O'Nullain & C. Mulvihill (eds.): Language, Vision and Music,, Amsterdam: Benjamins,* 117-132, 2002.

Establishing Natural Communication Environment between a Human and a Listener Robot

Yoshiyasu OGASAWARA[*]
Masashi OKAMOTO[*]
[*]Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
yoshiyas@kc.t.u-tokyo.ac.jp, okamoto@kc.t.u-tokyo.ac.jp
Yukiko I. NAKANO[†]
[†]Research Institute of Science and Technology for Society, Japan Science and Technology Agency
Atago Green Hills MORI Tower 18F, 2-5-1 Atago, Minato-ku, Tokyo, 105-6218, Japan
nakano@kc.t.u-tokyo.ac.jp
Toyoaki NISHIDA[‡]
[‡]Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract

The progress of technology makes familiar artifacts more complicated than before. Therefore, establishing natural communication with artifacts becomes necessary in order to use such complicated artifacts effectively. We believe that it is effective to apply our natural communication manner between a listener and a speaker to human-robot communication. The purpose of this paper is to propose the method of establishing communication environment between a human and a listener robot. In our method, their common intention is formed by joint attention and redundancy of behaviour.

1 Introduction

In recent years, the functions of familiar artifacts, such as an electric appliance and a personal computer, gets complicated as they are being advanced rapidly. To fully use the functions of such artifacts without giving a burden to the user is difficult with the present method, which the user has to learn in advance. The user should be able to tell the artifacts what she wants to do, that is, her 'intention'.

But, the user's intention does not always appear from the start. In most cases it is gradually appearing and becoming clear during the process of the communication. Therefore, it is important for the user to establish natural communication with the complicated artifact.

However, there are many examples where the natural communication is not established between humans and artifacts. One of the typical examples is making a video letter with a video cam. In this case the speakers in video letters often present unnatural way of speaking and behaviour. But, they will not do such a way of speaking when they talk to their close persons. We assume that this difference is caused by whether natural communication between a speaker and a listener is established or not.

The human listener responds to the speaker's behaviour with various ways. The listener implicitly conveys his listening attitude to the speaker through his responses or gestures. These behaviours and responses establish the communication and let the speaker feel relaxed. On the contrary, in making a video letter, the video camera is placed at the listener's position instead of a human listener. Then the communication is not established because the camera does not response to the speaker at all. Therefore, the speaker feels stressed and cannot behave as usual.

The purpose of this research is to build a listener robot whose natural behaviours as a listener allows its user to speak and behave in unrestrained ways. And then we propose the method of establishing the environment of natural communication where a human explains to the robot with gestures such as pointing. As a result it will be shown how the users get to exercise the functions of complicated artifacts effortlessly.

2 Related works

Many researches have been made on building a natural communication robot. Some of them discuss the matter in view of the listener's behaviour.

Watanabe and Ogawa (2001) developed InterRobot for the purpose of smoothing the voice conversation between the remote places. When the human speaks to InterRobot, it generates the gestures such as nodding from the user's speech, and it reacts as if it were listening to the user. On the other hand, speech information is sent to the remote place, and InterRobot duplicates the speaker's gestures conjectured from the information.

Kismet, which Breazeal and Scassellati (1999) developed, is one of the robots aiming at the human-like behaviour. Kismet can recognize a thing in its sight to which it should pay attention by vision processing technology. It can turn its neck and eyes toward a human face or an object that moves quickly.

Although these works try to make robots behave like a human, what he is talking about is not taken into consideration. Therefore, it is difficult for the human to explain about some topic to the robot in a natural way.

Ozeki et al. (2001) developed the video contents creation supporting system. In this system, the user speaks toward a set of cameras with specific gestures and words for operating cameras.

In this method, it is difficult to do the usual way of speaking because there is no communication with the artifact. If the system is able to communicate with users, it is expected that users can speak naturally. Such ability will make it possible for a human to use the system with ease, even if the system is updated and its function becomes complicated.

3 Natural communication with artifacts

In our research, we aim to establish natural communication between a human and a listener robot. In this section, we introduce the idea of 'User involvement' (Okamoto et al., 2004) as the basic framework in designing the computer-mediated communication environment. In enhancing the user involvement in human-to-robot communication, joint attention is significant for the communicative reality to be established there. Moreover, social skills for communication are also introduced to smooth the communication between a human and a listener robot.

3.1 User involvement

There are many discussions about realizing natural human-computer interaction. In our research we focus on the idea of 'User involvement' that Okamoto et al. (2004) put forward. User involvement means the cognitive way humans willingly engage in the interaction with computers, or the way in which humans are, on the contrary, forced to be involved in a virtual world which computers display or in a human-to-robot communication. The requirements are considered as follows:

- **Cognitive/Communicative reality should be achieved:** The user should feel the virtual object/world, or the human-to-computer interaction as "real".
- **Two (or more) cognitive spaces should be linked:** The user should move in and out smoothly at least two cognitive spaces such as his/her viewpoint (here) and what he/she sees (there).

Our goal is to achieve the communicative reality in the main so as to enhance the user involvement by implementing natural responses and reactions as a good listener into the listener robot. For it is difficult to fully establish the cognitive reality using a robot, in that current humanoid robots do not have so sufficient appearances or facial expressions as to make humans feel as if they the robots were living. Moreover, as Reeves and Nass (1996) points out, humans are likely to behave toward artifacts as if they were humans. Therefore, if a robot reacts to humans in unnatural ways, then they will assume it is churlish and non-cooperative and will not keep communicating with it.

From the point of the view of the user involvement, it can be said that each of the participants in communication originally lies in a different cognitive space before the communication begins. But, once the communication starts, there has to be something to connect those cognitive spaces, which functions as a reference point for one participant to access another. Then the communicative reality for the participant is achieved.

In human communication, what connects the participants' cognitive spaces is that which are cognitively shared in the participants, such as exchanging verbal information, establishing eye contact, matching action and reactions, joint attention, and so on. We believe that those factors that enable human communication will be applied to human-robot communication as well.

We focus on the 'joint attention' in particular in order to establish natural human-robot communication environment. Specifically, the speaker-listener communication using a listener robot is described.

3.2 Model of speaker-listener communication environment

Figure 1 illustrates the model of the speaker-listener communication environment. A speaker and a listener keep exhibiting their behaviours, such as gestures or utterances, to each other in the communication. During the process, one participant's intention is approaching that of the other, and then the common intention is formed. The common intention here means rather the 'intentionality', which Dennett (1987) suggests, than the 'intention' as used in a usual context. For the common intention concerns what is mutually aimed both by the speaker and the listener.

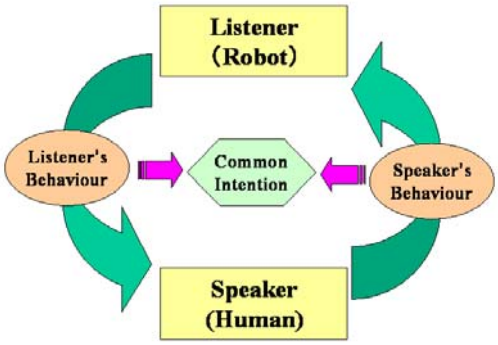


Figure 1: Model of speaker-listener communication environment

Take the explanation task for communication example. The state of the communication environment changes depending on what is being explained, or what is being attended to. When applying the model to the human-robot communication with a listener robot, the robot needs to change their behaviours according to the state of the communication environment. Therefore, in order to decide the robot's behaviour at each moment of the communication process, it is necessary to fully analyze what comprises the given domain of explanation.

3.3 Joint attention

Joint attention is the interaction where a speaker and a listener cognitively share an object that they attend to. For example, the listener looks at the thing the speaker points at. In human communication, these interactions are being done unconsciously and make the participants feel the communication natural. In fact the function of joint attention was implemented into such a humanoid robot as Kismet (Breazeal et al., 1999) and Infanoid (Kojima, 2000), and was proved to be effective for establishing natural human-robot communication. Therefore, joint attention is also one of the important requirements to establish communicative reality between the speaker and the listener.

In view of the user involvement, the joint attention leads humans to be mutually involved in the same cognitive space in the communication, which is the overlapped part of both cognitive spaces of speaker and listener. As a result, the joint attention enables the listener to accesses the speaker's attention object through the joint attention as a reference point, and vice versa. Therefore, joint attention achieves the smooth transition among the cognitive spaces of the participants, which helps to establish the communicative reality for them.

3.4 Social Skills

We propose the idea to use the social skills, as a policy for establishing more advanced communicative reality. Social skill is the theory to establish human relations effectively in view of the communication techniques. Aikawa (2000) describes the social skills for listening to the partner's speech. Table 1 summarizes them.

Table 1: Skills for listening to speech

Capable Pose	No interruption, No rush
Open Question	Prompting, Explaining more in details
Reflection	Verbal response, Repeat, Paraphrase, Summary
Using Non-verbal Channel	Posture, Gaze, Nodding, Distance, Hand's movement
Decoding Speaker's Non-verbal Channel	Voice (pause, speed, pitch), Emotion, Gaze, Hand's movement

In these skills, verbal response and nodding are important for building the listener robot. Many people have experienced that the listener's nodding improves the rhythm of the speaker's speech. It is difficult for robots to understand the contents of the speech, but it is possible for them to behave as if they were listening to the speaker through using these skills.

4 Analysis of speaker-listener communication

In this section we analyze the characteristics of human communication between a speaker and a listener in order to define the appropriate behaviours of our listener robot. First we describe the human behaviours for establishing joint attention according to psychological studies. Secondly we analyze the videos which captured explanatory scenes of humans and show the observations of the videos.

4.1 Behaviours for establishing joint attention

As shown in the previous section, joint attention is one of the significant factors that constitute speaker-listener communication.

Tomasello (1999) suggests that joint attention is divided into the following 3 types according to the development levels of infants:

- **Check Attention:** Attention to the partner, or the object he shows.
- **Follow Attention:** Attention to the object that the partner points at or his eye gaze turns to.
- **Direct Attention:** Making the partner pay attention to the object with voice, eye gaze, and so on.

This classification indicates that it is necessary for establishing joint attention to properly react to the behaviours such as showing by hand, pointing by finger and turning eye gaze.

Moreover, Clark (2003) explains about the functions of such behaviours as *pointing* and *placing* in human communication. Among them the behaviours for 'directing-to' an object are classified into the following categories:

- *Pointing* (finger)
- *Sweeping* (arm)
- *Tapping* (finger, foot)
- *Nodding* (head)
- *Turning* (torso, face)
- *Eye Gazing*
- *Speaking* (frequently accompanied by head and face)

In particular *eye gazing* is considered to be the most attention-getting behaviour. Additionally, where an object is placed affects the listener's attention. For instance, placing an object onto a desk or in front of the hearer gets his attention much.

Therefore, each of these behaviours should be counted as attention behaviour of a speaker.

4.2 Observations of explanatory movie

We observed an explanatory video movie so as to confirm that those attention behaviours are actually used for explanatory communication. The data we used is 40 minutes of an educational material video for DIY (Do It Yourself) in which a professional instructor explains how to use machine tools against the TV camera.

As a result of the observation, it was proved that most of the attention bahaviours were actually used in explaining scenes. In particular, pointing, showing and eye gaze were most frequently used, when the camera also focused on the object.

Moreover, some of the attention behaviours were frequently used simultaneously (e.g. Showing + Pointing + Gazing + Speaking "*This is...*"). During the large part of the instruction the instructor was attending either to an object to be explained or to the camera in front of her. We also found that the direction of her gaze continuously changed between the object and the camera at short intervals while she was speaking.

4.3 Analysis of listener's response behaviours against speaker

Since the explanatory task observed in the previous section is toward a TV camera alone, it is different from the actual explaining communication between a human speaker and a listener. We thus made two movies of explanatory scenes between humans and analyzed the listener's behaviours against the speaker's ones using a video annotation tool, Anvil[1] (Figure 2).

Figure 2: The video data of explanatory scene

The task to be explained by the speaker was how to assemble a piece of furniture (a metal rack). The speaker was one of the authors and the listeners were two students.

Table 2: Analysis of listener's behaviours

	Listener's response to speaker's gaze		
	Joint attention	Gaze toward speaker	Nodding toward speaker
Listener 1	76.4	85.3	64.7
Listener 2	84.7	71.4	47.6

(%)

Table 2 shows the result of the analysis. When the speaker attends to an object, the listener attends to it at more than 75%. Therefore, joint attention between the speaker and the listener is achieved at high frequency. Moreover, when the speaker turns his gaze on the listener, the listener turns his gaze

[1] http://www.dfki.de/~kipp/anvil/

back to the speaker at more than 70%. In many cases the listener gives a nod in concurrence with his gaze, but its frequency differs greatly in individuals.

Among those exchanges occurred in the communication the most frequent behaviour transition during a short period is as follows:

1. The speaker turns his gaze on the listener.
2. The listener turns his gaze back to the speaker.
3. The listener gives a nod (or does nothing).
4. The speaker looks at the object to be explained.
5. The listener looks at it.

We assume that this transition occurs because the speaker wants to confirm the listener's attention.

In addition, the following characteristics were commonly observed regarding the two listeners:

- Among the speaker's behaviours, showing the object by hand and turning gaze especially attract the listener's attention.
- The listener usually attends to the object after the speaker's multiple behaviours (e.g. showing + gaze).
- When the speaker moves or changes his posture, the listener attends to the speaker himself instead of the object.

4.5 Summary

The observations suggest that the speaker-listener communication in explanatory task has the following characteristics:

(1) On attention behaviours:
- According to the speaker's attention with pointing, gaze, or posture, and then **the listener attends to the same object**.
- Among the attention behaviours **showing by hand**, **pointing** and **gaze** are the most affective ones, and are **often used simultaneously**.

(2) On communication modes:
- There are different communication modes in explanation: (a) the speaker **attends to an object** and explains about it to the listener, (b) the speaker **attends to the listener** and talks to him, (c) the speaker glances at the listener to **confirm the listener's response**.
- In other words, **the communication mode changes according to each relation among the speaker, the listener and the object**.

5 Requirements for a listener robot

In this section we describe the requirements for building a listener robot based on the observations and the analysis in previous sections.

5.1 Establishing joint attention

As shown in Section 4, it is frequently observed that both the speaker and the listener attend to the same object in explanation task. Therefore, implementing the ability of joint attention into a listener robot is required.

The correspondence of the listener's proper reaction with the precedent actions of the speaker is necessary for the communication between a speaker and a listener to be established. It is through this repetitive process that natural speaker-listener communication is achieved. Among the correspondent action and reaction couplings, the most fundamental and effective one is seemingly joint attention.

In order to achieve joint attention, the listener has to recognize the speaker's attention and its target from the observation of the speaker's behaviours, and should react to the attention appropriately and instantly.

Moreover, it is also essential for establishing joint attention whether the attention behaviours are put out by the speaker intentionally or not. For reacting against non-intentional behaviours would become unnatural in communication. Therefore, we proposed the method of detecting the intensity of the speaker's intention based on the redundancy of the behaviour.

5.2 Modality of attention behaviour

Table 3 summarizes attention behaviours in view of each modality based on the observations in Section 4.

Table 3: Modality of attention behaviour

Modality	Behaviours
Hand movements	Pointing, Grasping, Showing
Eye gaze	Direction (head, eyes)
Posture	Approaching, Direction (body) Standing-up, Sitting-down
Speech	Deictic words, Verbal response, Mentioning

A listener and a speaker use these types of behaviours to express their own attention or to interpret the other's. In Table 3, hand movements and

eye gaze are assumed to strongly represent attention. If a listener robot is able to recognize such attention behaviours of the speaker, then the establishment of joint attention will be easier.

The other types of behaviours are also used for representing attention. However, they are often used with a certain degree of redundancy. The redundancy of attention behaviour is described in the next subsection.

5.3 Redundancy of attention behaviour

We observed that attention behaviours are frequently represented with more than one modality or in repeating fashion. Such redundant manners strongly suggest that those behaviours are intentional. In other words, the redundancy of behaviour manner helps a listener to recognize the speaker's intention.

We propose that the redundancy of attention behaviours should be applied to the design of natural communication environment using a listener robot. The speaker's redundant behaviours strongly suggest the intentionality of his behaviours. Recognizing the redundancy enables the listener robot to easily understand the speaker's intention. The effectiveness of applying the informative redundancies to an agent's sensors and actuators is also suggested by Pfeifer and Scheier (1999).

It is assumed that there are the following two types of redundancy in the speaker-listener communication:

- **Redundancy of modality:** Using multiple modalities of behaviours simultaneously
- **Redundancy of time:** Using the repetitive or persistent behaviours

Since the redundant behaviours are often intentional, communicative robots should react rapidly and appropriately against those intentional behaviours of humans. Conversely, the robots also can convey strong intention to humans using redundant actions or behaviours. This effect of redundancy is discussed in Tajima (2004) through the human-robot communication experiment.

A human can adjust himself to a robot with ease if the robot can interpret redundancy. It is often observed that a human behaves in more redundant ways when he cannot communicate with others smoothly. Thus, even the robots with insufficient recognition abilities can communicate with a human as long as he is willing to use the redundant ways of communication.

5.4 Modes of Communication

Intentional attention behaviours are not always used in the process of explanatory communication. It is required for a listener robot to behave properly even if there are no such intentional behaviours of a speaker.

The analysis in Section 4 suggests that the proper behaviours of a listener robot should be determined depending on the relations between the speaker, the listener and the target. We call that a communication mode. According to each of the communication modes the listener robot should be able to change its behaviours properly. The modes of communication are classified into the following four types in view of the speaker's attention (also see Figure 3):

- Talking-*to* mode:
- Talking-*about* mode
- Confirming mode
- Busy mode

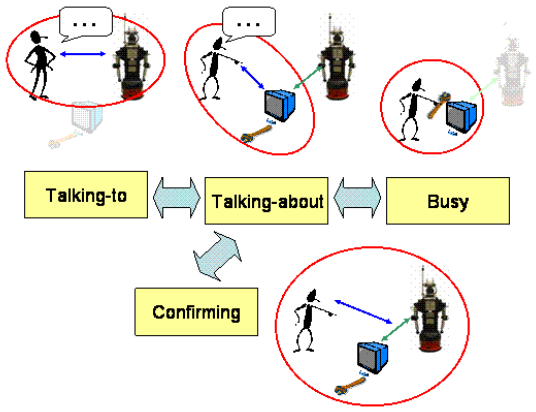


Figure 3: The modes of communication

In general, the speaker engaged in the explanation task attends to either the listener or the target to be explained. In the '**talking-*to*'** mode, the speaker is mainly watching the listener and is involved in the cognitive space based on the relation between the speaker and the listener. As the speaker in the 'talking-*to*' mode expects the listener to be involved in the same conversation, the listener should pay attention to the speaker himself.

On the other hand, when the speaker is mainly watching the target to be explained, he is in the '**talking-*about***' mode. In this mode, the speaker expects the listener to cognitively share the target. Therefore, the listener should attend to the target in turn.

Additionally, in the cases where the speaker switches his gaze between toward the listener and toward the target, he is interested in the relation of the listener and the target, that is, whether or not the listener is paying attention to the target. We thus call the mode as the '**confirming**' mode in that he tries

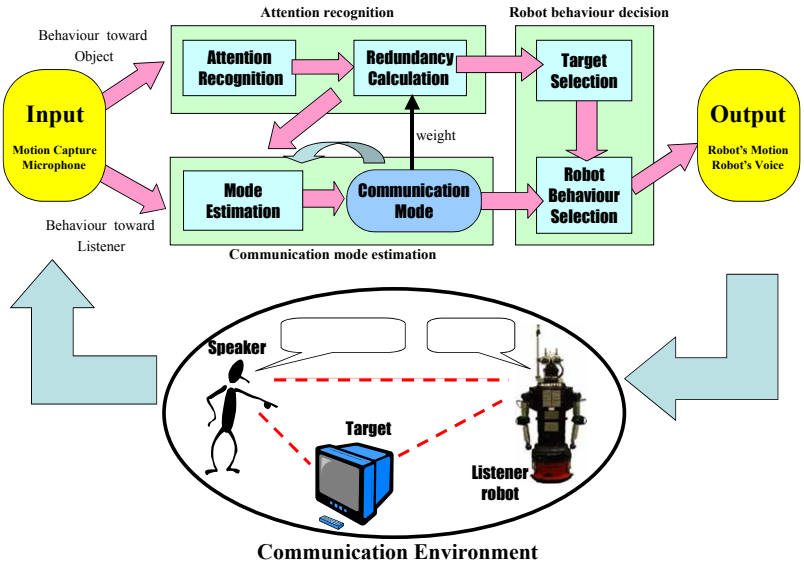


Figure 4: Architecture of listener robot

to confirm the listener's attention. In the confirming mode, the listener should turn his gaze back to the speaker and respond with nods or verbal responses.

The '**busy**' mode is the mode when the speaker is devoting himself to his work without talking to his listener. In this mode the speaker is attending more to the target than in the talking-*about* mode, and tends to ignore the listener. This situation is not favorable for explanation task, but frequently occurs when the speaker is not skillful. A listener robot can keep attending to the target during the busy mode, which is one of the advantages for using robots.

These modes of the speaker's attention differ in which cognitive space the speaker is involved. The speaker creatively uses the multiple modes to establish communicative reality. If the listener robot does not appropriately react according to the specific communication mode, the speaker will not be able to smoothly switch these modes, and then he will not feel the communication with the robot as real.

6 Constructing a Listener Robot

In this section, we describe the implementing method of the attributes of establishing natural speaker-listener communication environment, suggested in Section 5, into the listener robot.

6.1 Target

In this research, we construct a listener robot, which is designed to participate in such human-robot communication environment as the situation where a human explains the procedure of assembling a piece of furniture or an appliance toward the listener robot.

This task of explaining the procedures of assembling furniture implies the basic behaviours that are usually used by the speaker and the listener. Therefore, constructing a listener robot involved in the task will be also helpful to solve the problems of many other explanation tasks.

6.2 Hardware

In constructing a listener robot, we use Robovie[2], a humanoid robot. Robovie is nearly as tall as a human and can move its hands, head, and eyes. Additionally, it can make a move with its wheels.

The motions of the speaker's body or the tools he uses are recognized via the motion capture. The markers of the motion capture are attached to the speaker's head, arms, body, and the objects to use or to point at in his explanation. In addition, one microphone is used to measure the speech sound volume of the speaker.

6.3 Architecture of listener robot

The architecture of the listener robot is shown in Figure 4.

The robot receives motion capture data of the speaker's behaviours and the speech sound as input. As the result of processing the data, the robot outputs bodily expressions as feedback to the communication environment.

[2] http://www.mic.atr.co.jp/~michita/everyday-e/

- **Input:** Motion capture data, and speech sound.
- **Output:** Robot behaviours (head, eyes, arms, head nod, voice, and body direction)

The following subsections describe the methods used in each module in the robot system architecture.

6.3.1 Attention recognition

In order to recognize the attention behaviours of a speaker, the listener robot processes each data from respective markers attached to all the possible targets. The details of the process are described below.

First, the confidence for attention toward each target is calculated from the positions of the motion capture markers to decide the following behaviours:

- Eye gaze (head direction)
- Pointing
- Grasping
- Repetitive hand gestures (e.g. tapping)
- Physical relationship with objects (distance, body direction)

Respective bahaviours are recognized based on simple calculation of such as distance and angles between markers. For example, gaze direction is estimated from the directions of two markers attached on the speaker's head. Grasping is recognized by calculating the distance between hand and object. The precise direction of eyes is difficult to recognize from motion capture information. Therefore, we are not concerned it in this listener robot.

Repetitive gesture of hands is calculated by employing methods proposed by Tajima (2003) and Anuchitkittikul (2004). Their method detects human's repetitive gestures by using an autocorrelation function and the Fast Fourier Transform (FFT).

The calculation of the confidence for respective gestures is based on the relational network of sensor data and behaviours as illustrated in Figure 5. Outputs from this process are calculated by using the method proposed by Hatakeyama (2004). This theory applies Bayesian Network to the processing of sensor information and the decision of robot's behaviour. In this method, inputs and outputs are represented as random variables. The causal relationships between inputs and outputs are expressed as a conditional probability table.

This method has an advantage over other methods that describe the relationship between input and output as rules, because it is more robust against the noisy input and the changes of communication environment.

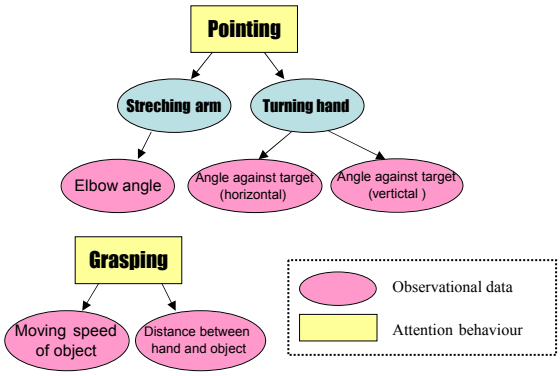


Figure 5: Network for attention recognition (partial)

Each node of the network was decided by us based on the observations in Section 4.

Furthermore, the confidence values of the respective behaviours are regarded as scores, and then the scores are weighed and added up into the redundancy of attention behaviours toward the target (Redundancy of modality).

Weighing scores depend on the current communication mode and the duration of behaviours (Redundancy of time). For example, in the Talking-*to* mode, weighed values are less than in any other mode because gazing attention to the target does not frequently occur in the Talking-*to* mode.

6.3.2 Communication mode estimation

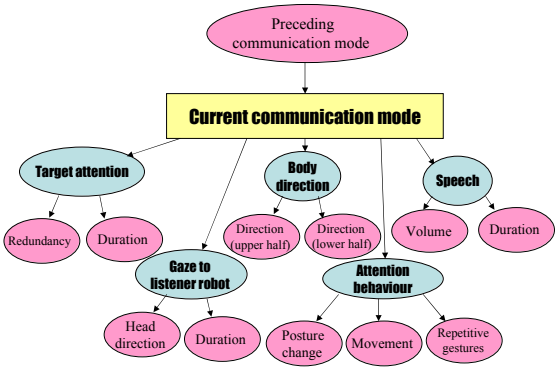


Figure 6: Estimation of communication mode

The communication modes suggested in Section 5 are classified according to what relation is established between the speaker, the listener and the target. Our listener robot decides each communication mode based on the speaker's behaviours.

Which communication mode the speaker and the listener are involved in is estimated from the following inputs using the similar algorithm to that of attention recognition (also see Figure 6):

- Target attention (with the largest redundancies)
- Gaze to the listener robot
- Body direction toward the listener robot
- Attention behaviour (e.g. posture change)
- Speech (volume, duration)
- Preceding communication mode

The general conditions for recognizing each communication mode are as follows:

- **Talking-*to***: with frequent speech, speaker's body and face toward the listener
- **Talking-*about*:** with attention behaviours and gaze to the target
- **Confirming:** with momentary gaze toward the listener during target attention
- **Busy:** with no speech, with continuous attention to the target

6.3.3 Robot behaviour decision

The listener robot's behaviours are decided by if-then rules based on the redundancy of attention behaviour and the communication mode. The object with the largest redundancy is selected as the target for the listener robot's attention.

The manners of the listener robot's behaviours are summarized as follows:

- **Talking-*to*:** turns the head to the speaker and randomly nods when the speech is aborted.
  - **Low redundancy:** *no attention*
  - **Medium redundancy:** *gaze attention in a short period*
  - **High redundancy:** *slight head move and gaze attention in a short period*
- **Talking-*about*:** uses various methods.
  - **Very low redundancy:** *no attention*
  - **Low redundancy:** *gaze attention*
  - **Medium redundancy:** *head attention after gaze attention*
  - **High redundancy:** *prompt head attention*
  - **Very high redundancy:** *head move and verbal responses*
- **Confirming:** takes a glance at the speaker, nods with verbal responses, and then returns to the original state. Nodding and verbal responses are randomly produced.
- **Busy:** in the same fashion as talking-*about* mode, but with no utterances,

6.4 Examples of explanation for the listener robot

(a) Talking-*to* mode

(b) Talking-*about* mode (Grasping + Pointing)

(c) Confirming mode

Figure 7：Example scenes of explanation for the listener robot

Figure 7 shows examples of explanation for the listener robot. These pictures illustrate three speaker's modes when the speaker explains how to assemble the metal rack using tools.

Picture (a) shows that the speaker is talking to the robot and the robot turns its head to the speaker. In picture (b) the speaker attends to a tool with multiple behaviours. Then the robot turns its head to the tool, when the joint attention between the speaker and the robot is established. In picture (c) the speaker looks at the robot so as to confirm its reactions during his attention to the tool. At the time the robot glances at the speaker with its eyes and nods to the speaker.

As a result, the listener robot efficiently recognized speaker's intentions using the redundancy of speaker's natural behaviours. Moreover, considering communication modes helped a lot to realize the proper reactions of the robot for explanatory task.

7 Conclusion

In this paper, we proposed the method to establish the natural communication environment with the artifact such as a robot. We developed the listener robot applying this method in the explanation of assembling furniture. The listener robot established natural joint attention with a human speaker using the redundancies of attention behaviour.

In the future, we will examine psychological burden of the user in the explanation task with the listener robot. Additionally, we will analyze more detailed behaviours of human speakers and listeners and define more proper communication modes and robot's behaviours in each mode.

Finally, we will develop the system supporting creation of the video contents using this listener robot. If this robot's behaviour makes a speaker relaxed, the speaker is expected to be able to create good video contents.

References

Mitsuru Aikawa. *Techniques of Human Relation – Psychology of Social Skills (in Japanese)*. Number 20 in Selection of Social Psychology. Saiensu-Sha, 2000.

Burin Anuchitkittikul, Masashi Okamoto, Hidekazu Kubota, and Toyoaki Nishida. Gestural interface for the creation of personalized video-based content. In *Proceedings of the 2nd International Conference on Information Technology for Application (ICITA 2004)*, China, 2004.

Cynthia Breazeal and Brian Scassellati. A context dependent attention system for a social robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)*, 1146–1151, Stockholm, Sweden, 1999.

Herbert H. Clark. Pointing and placing. In *Pointing: Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Assoc Inc., 2003.

Daniel C. Dennett. *The Intentional Stance*. The MIT Press, 1987.

Makoto Hatakeyama. Human-Robot Interaction based on Interaction Schema (*in Japanese*). Master's thesis, Graduate School of Information Science and Technology, The University of Tokyo, 2004.

Hideki Kozima. Infanoid: An experimental tool for developmental psycho-robotics. In *Proceedings of International Workshop on Developmental Study*, Tokyo, 2000.

Yoshiyasu Ogasawara, Takashi Tajima, Makoto Hatakeyama, and Toyoaki Nishida. Human-robot communication of tacit information based on entrainment (in Japanese). In *Proceedings of The 18th Annual Conference of the Japanese Society for Artificial Intelligence*, 2004.

Masashi Okamoto, Yukiko I. Nakano, and Toyoaki Nishida. Toward enhancing user involvement via empathy channel in human-computer interface design. In *Proceedings of IMTCI*, 2004.

M. Ozeki, Y. Nakamura, and Y. Ohta. Camerawork for intelligent video production –capturing desktop manipulations. In *Proceedings of International Conference on Multimedia and Expo*, 41–44, 2001

Rolf Pfeifer and Christian Scheier. *Understanding Intelligence*, Bradford Books, 1999.

Byron Reeves and Clifford Nass. *The Media Equation: How People treat computers, television, and new media like real people and places*. CSLI Publications, 1996.

Takashi Tajima and Toyoaki Nishida. Manual-less interaction based on synchronization and modulation (*in Japanese*). In *Proceedings of IEICE Human Information Processing*. The Institute of Electronics, Information and Communication Engineers, December 2003.

Michael Tomasello. *The Cultural Origins of Human Cognition*. Harvard University Press, 1999.

Tomio Watanabe and Hiroki Ogawa. InterRobot for human interaction and communication support. In *Proceedings of world Multi-conference on Systems, Cybernetics and Informatics (SCI2001)*, 466–471, 2001.

Reading of intentions that appear as diverse nonverbal information in face-to-face communication

Yoshimasa Ohmoto
University of Tokyo
Japan
Ohmoto9@dolphin.c.u-tokyo.ac.jp

Kazuhiro Ueda
University of Tokyo
Japan
ueda@gregorio.c.u-tokyo.ac.jp

Takanori Komatsu
Future University-Hakodate
Japan
komatsu@fun.ac.jp

Abstract

In face-to-face communication, humans read their partners' intentions by using diverse nonverbal information. In this study, we focused on a "lie" as being one of the partners' intentions that does not correspond directly to specific nonverbal information. Specifically, we investigated whether a "lie" was discernible in actual communication by using nonverbal information such as gazing, prosody and facial expressions. Participants were asked to play a game of revised "Indian poker" in order to observe lies, which appeared in a situation similar to actual communication actual communication. The result of the application of discriminant analysis by using four of the 13 nonverbal information provided a discrimination rate of about 75% - 85% to determine whether participants told a lie or not. Each participant utilised different nonverbal information with changes in situations regarding communication. On the other hand, in the case of the same participant, a sign of the coefficient of the discriminant function of the same kind of nonverbal information was stable even though the situation changed. Therefore, by analysing the nonverbal information utilised in situations using the communication media, we may also communicate naturally and smoothly, like in a face-to-face communication by using communication media.

1 Introduction

Presently various kinds of communication media are quickly spread in our daily lives. However, it can be said that face-to-face communication is the most natural and smooth form of communication for most people. In face-to-face communication, we not only use verbal information but also diverse nonverbal information for smoother communication (Von Raffler-Engel, 1980; Daibou, 2001). For example, facial expressions and gestures are used to express the intention of irony or a joke. Many communication media have fewer nonverbal communication channels than face-to-face communication. Therefore, we investigate the utilisation of nonverbal information in actual communication and then apply it for the creation and use by the media.

While diverse nonverbal information complements the meaning of verbal information, they are mainly used to determine the meanings of polysemous words or to convey the social meaning about relations and attitudes (Von Raffler-Engel, 1980; Burgoon, 1994). Hayashi (1998) reported that the utterance of the Japanese interjection "eh" with different prosodic information could be interpreted differently to mean different things.

Recently, some researchers have conducted research on a human's intention based on his/her nonverbal information. For example, Vertegaal et al (2001) reported that gaze directional cues of participants could be used as a means of establishing who was talking with whom and Goto et al (2002) created a system that provides its user with following a word inferred from the previous word uttered when the user hesitated to speak. In most studies however, a limited number of nonverbal information was used for reading, this too. Simple intention corresponded directly to specific nonverbal information. In actual communication, we use much more nonverbal information such as, gazing, prosody, facial expressions, gestures and so on. We further need this nonverbal information to read complex intentions, such as, irony, jokes, state of mind of strain or dissatisfaction and so on that do not correspond directly to specific nonverbal information. Therefore, investigating the synthetic use of diverse nonverbal information is required for reading such complex intentions.

In actual communication, communication strategies change according to the change in human relationships and situations. In this study, communication strategy means giving and hiding people's intention for achieving their purpose to communication partners. The manner in which to tell a lie is a kind of communication strategy. Therefore, we expect that the manner to tell a lie will change in the process of communication. A lie is a complex intention and can be define objectively. Then we focused on a "lie" appearing in actual communication and we investigate nonverbal information when people tell a lie. Up to now, many researchers have studied about discrimination of a lie. However, there are very few studies that observe nonverbal information when people tell a lie during actual communication.

Therefore, it is considered necessary to pay attention to the diverse nonverbal information in order to catch the expressions that change with situations in actual communication, such as, the motives and manners to tell a lie, times and partners when a lie is told, individual characters of humans who tell a lie and so on.

This study has two purposes. (1) To confirm that the changes of diverse nonverbal information were observable in a situation similar to actual communication, in which diverse nonverbal information was used and a participant was able to spontaneously select between telling a lie and not and that these changes were useful in discriminating a participant's intention. (2) To investigate by using nonverbal information, whether lies as complex intention, which do not correspond directly to specific nonverbal information, were able to be discriminated or not and the feature of nonverbal information when lies were told by analysing which variables contribute towards the discrimination of lies. For these investigations, we conducted an experiment using a game in which participants spontaneously selected between telling a lie and not.

Below, "intention" means "the inner representation which cannot be directly observed ", "information" implies "the external representation, which can be observed", a "variable" means "a part of information used for analysis" and a "lie" means, "linguistic statement deceived intentionally" (THE YUHIKAKU DICTIONARY OF PSYCHOLOGY, 1999).

The rest of the paper is organised as follows. Section 2 explains the outline of the experiment conducted in this study. Section 3 describes the method of analysis and results. Section 4 contains a discussion and conclusions.

2 Experiment

We conducted an experiment to record the diverse nonverbal information that the participant's made while spontaneously telling a lie during communication.

Table 1. List of points.

Defeated.	-3 points.
Leave the game.	-2 points.
Leave despite holding a top card.	-4 points.
Winner.	Gets all of the other players' lost points.

2.1 Revised "Indian Poker"

Indian poker is a card game. Each player is dealt one card whose face side they cannot view. Each player must place the card on the forehead so that all the other players can see that card. This means that you can see every card except your own. The players then decide whether or not to stay in the game from the number of cards shown and communication with the other players. Finally all the players' cards are turned face up on the table and the player with the highest card wins. Exceptionally, one (A) counts as 14. The losers points are lower when a player drops out of a game than the losers points when he/she is defeated (Table 1).

Under these normal rules, winning or defeat is easily estimated from the cards shown by others', communications for encouraging others to leave the game or not difficult. Therefore, we add the rule "2, 3, 4 cards win when all the other players in the game have picture cards (11, 12, 13)". By adding this rule, winning or defeat cannot be easily estimated. The game of normal Indian poker plus this rule is called "Revised Indian poker".

By adding this rule, whether the other players stay in the game or not are concerns of winning or defeat. The players then communicate with each other in order to obtain information on their own card and the intention of the other players is to decide on whether to stay in the game or not. In this communication, a player is clue to tell a lie in order to make the other players leave the game.

The reason for using the revised Indian poker is that participants are able to choose their own communication strategy on whether to tell a lie or not. This choice is his/her own and without instructions from the experimenter. So lies, which appear according to the change in communication strategy, can be observed in this environment, which is similar to actual communication.

2.2 Setting

Participants were asked to play a game of revised Indian poker. The behaviour and utterances of the

Figure 1. Position of each participant and the cameras.

participants playing this game were recorded on video.

Three participants (players) participated in the experiment. They sat surrounding a table and video cameras for the recording were placed between them (Fig. 1). The video camera recorded the participant who was sitting on the opposite side of the table. In this setting, the participants played revised Indian porker. The participants' utterances and actions were not controlled; they were allowed free communication. The experimenter asked them to play the game of revised Indian poker only after the participants were briefly provided instructions on the rules and strategies of the game.

The three participants were two graduate students and the person conducting the experiment[1]. The two graduate students were acquainted with each other.

2.3 Procedure

The experiment was conducted in the following manner.

1) Three participants sat surrounding a table (Fig. 1).
2) The experimenter briefly provided instructions on the rules and strategies of the game.
3) The experimenter shuffled the cards and dealt a card to each participant.
4) The participants were asked to place the card on the forehead without looking at it.
5) Each participant communicated in order to make the other participants either to leave or stay in the game.
6) Each participant showed his/her card to the others, after deciding on whether to stay or leave the game.
7) After the winner was decided, the losers paid the points to the winner.

[1] The reason why the experimenter participated in the game was that it was difficult to communicate smoothly with the beginners of the game and to detect a lie. This experimenter acted like a usual player.

The procedures of 3–7 were defined as a trial. This trial was repeated about 20 times.

In addition, we conducted the same experiment on the same participants after three months of the first experiment. The numbers of trials were 20 in the first experiment and 17 in the second experiment.

3 Results

3.1 Method

By using the discriminant analysis, an investigation was made into how many lies could be discriminated and the effect of the contribution of variables.

Utterance, a unit of analysis, was taken out from the video, which was recorded during the experiment. This “utterance unit” was defined by the time span from the start time of active utterance by a participant to 0.5 seconds after the utterance ended. This definition was adopted because, at the end of utterances, eyes moved or facial expressions changed in many cases.

The variables of "gaze", "prosody" and "facial expressions" in every utterance unit were elicited.

To sum up, in this experiment; we assume that independent variables were 13 nonverbal information variables in Table 2, while a dependent variable was whether a participant told a lie or not.

The objective of gazing was judged by the direction in which the participant turned his face and his eyes, by the relation between the position of the camera and other participants.

Each variable in a row of “gaze” in Table 2 is explained here. At the top of a row, the definition of "the Rate of Gazing at the Partner of Conversation (RGPC)" is the percentage of time during which a participant gazes at his/her partner of conversation as a portion of the total time of an utterance unit. At the middle of a row, the definition of "the Rate of Gazing at the Useful Object for Communication (RGUOC)" is the percentage of time during which the participant is gazing at the object that has useful information for communication as a portion of the total time of an utterance unit. In this case, these objects are the faces of the other two participants and their cards. At the bottom of a row, the definition of "the Transitional Rate of Gazing at the Object (TRGO)" is the value of the numbers that a participant is gazing 0.4 seconds or more divided by the time of an utterance unit.

It may be difficult for people to pick up subtle changes in facial expressions. We have noticed that people force a smile while telling a lie in many cases for the purpose of deception. It was reported that there is a time difference between the start of the expression of the eyes and that of the mouth in a

Table 2. 13 Nonverbal Information.

Nonverbal Information.	Independent variables.
Gaze. (Three variables.)	The Rate of Gazing at the Partner of Conversation (RGPC).
	The Rate of Gazing at the Useful Object for Communication (RGUOC).
	The Transitional Rate of Gazing at the Object (TRGO).
Prosody. (Nine variables.)	Pitch, (the first half, the second half, change.)
	Power, (the first half, the second half, change.)
	Speed, (the first half, the second half, change.)
Facial expression. (One variable.)	Whether the Mouth Reacted Early rather than the Eyes (WMREE).

forced smile (Nakamura, 2000; Shikura et al, 2001). In this study, "the difference of the reaction time between the eyes and mouth" was regarded as a typical variable of facial expressions. Briefly speaking, "Whether the Mouth Reacted Earlier than the Eyes (WMREE)" was regarded as a variable. The value was set to 1 (truth) when the experimenter could judge that the mouth reacted earlier than eyes. Otherwise, it was set to 0 (false). The value was set to 1 (truth) when only the month reacted.

The experimenter classified nonverbal information of every utterance unit into the variables in Table 2.

3.2 Procedure

We performed discriminant analysis to determine, which variable was useful in discriminating lies among the variables.

There were two groups in the discriminant analysis. One was "an utterance, which is a lie" and the other as "other utterances." Below, "an utterance, which is a lie" is called the "Lie utterance" for short. We define that an "equivocal utterance" is an utterance that is neither a truth nor a lie, (an evaded utterance, a noncommittal answer and so on). There were between 20-30% of "equivocal utterances" in the whole utterance. These “equivocal utterances” were classified as “other utterances.”

By selecting different variables for discrimination showed which variables were useful in discriminating lies. First, an experimenter identified the pairs of variables with 0.8 or more correlation coefficients and removed the variable with a lower F-value. As a result, the experimenter removed one variable at a minimum and five variables at the maximum. Next, the variables were selected by backward elimination. The selected variables by this operation were regarded as the main variables that largely contributed to discriminating lies.

3.3 Results

As one of the participants, (Participant B) did not tell a lie in the first experiment. The other participant is called Participant A.

In all cases, four variables were selected. However, the discrimination rate reached 75-85%. According to the research of Mirror and Stiff (1993), a percentage of correct answers when people judge are at most 70%. Therefore, this result shows that we can discriminate people’s lies by using diverse nonverbal information at the same or with more accuracy as people can. Moreover, discrimination may be made with a comparatively small number of variables, even if much nonverbal information is used.

The actions of Participant A in telling a lie were analysed. The experimenter identified specific actions when Participant A told a lie in the first half of the first experiment as follows; "gaze was not fixed" and "utterances while thinking about something". However, most of these actions were not identified in the second half of the first experiment and during the second experiment. Below, “the first half of the first experiment” is called “1-1”, “the second half of the first experiment” is called “1-2”. In the behaviour identified, differences existed between “1-1” and “1-2” and similarities existed between “1-2” and the second experiment. Therefore, we hypothesised that differences would exist between “1-1” and “1-2” and similarities would exist between “1-2” and the second experiment in nonverbal information too.

In order to verify this hypothesis, we performed discriminant analysis on the data sets of “1-1” and “1-2” (Table 3 (a)). However, no significant change was observed in the discrimination rate. At a glance, similarities were not found between “1-2” and the second experiment, in the variables that largely contributed to the discrimination.

Then, each of the discriminant functions of “1-1” and “1-2” of Participant A classified the data set of the second experiment as unknown data set. The result is shown in Table 4.

When the discriminant function of “1-2” classified the data set of the second experiment, the discrimination rate did not decrease in both the "lie utterances" and "other utterances". However, when the discriminant function of “1-1” classified the data set of the second experiment, the discrimination rate decreased greatly in the "lie utterances" while the discrimination rate did not decrease in the "other utterances". This result shows that “1-2” and the second experiment resemble each other in the man-

Table 3. Results of the discriminant analysis.

(a) Participant A.

Experiments.	The number of utterances.	Discrimination rate.	The number of variables.	The selected variables. (Coefficient of discriminant function, F-value).
First.	Total.			
	Total utterances: 86 Lie utterances: 24 Other utterances: 62	Lie utterances: 74% Other utterances: 82%	4	RGUOC. (-7.0,16) WMREE. (2.2,9.2) Power (change). (0.92,2.1) Pitch (the first half). (0.51,2.1)
	The first half (1-1).			
	Total utterances: 36 Lie utterances: 15 Other utterances: 21	Lie utterances: 79% Other utterances: 83%	4	RGPC. (-5.5,13) WMREE. (0.93,1.0) Pitch (change). (0.47,0.28) Pitch (the first half). (0.27,0.27)
	The second half (1-2).			
	Total utterances: 50 Lie utterances: 9 Other utterances: 41	Lie utterances: 89% Other utterances: 90%	4	WMREE. (5.2,12) Power (change). (2.6,7.2) Pitch (the first half). (1.7,5.6) RGUOC. (-10,4.2)
Second.	Total utterances: 126 Lie utterances: 14 Other utterances: 112	Lie utterances: 79% Other utterances: 83%	4	RGUOC. (-4.1,15) WMREE. (3.8,7.3) Pitch (change). (1.7,8.3) Speed (the second half). (1.1,3.7)

(b) Participant B.

Experiments.	The number of utterances.	Discrimination rate.	The number of variables.	The selected variables. (Coefficient of discriminant function, F-value).
First.	Total utterances: 48	—	—	—
Second.	Total utterances: 95 Lie utterances: 10 Other utterances: 85	Lie utterances: 80% Other utterances: 73%	4	WMREE. (2.0,5.9) RGPC. (-5.1,5.5) Loudness (change). (-4.5,4.6) RGUOC. (4.3,4.5)

ner to tell a lie, but also that differences exist between these and "1-1".

As mentioned above, it is confirmed that the nonverbal information that people showed in telling a lie changes during communication.

One of the reasons why the participants' expression of nonverbal information changed was that the participants were getting habituated in telling a lie while playing revised Indian poker. This habituation would change the communication strategy, for example the manner to tell a lie change from reactive to active. In turn, the change in the communication strategy would change their expression of nonverbal information. Therefore, their discriminant function changed.

In spite of the three months interval, the discrimination rate did not decrease. Therefore, it is suggested that the discriminant function does not change when the situations and the participants are the same.

In all the sets of main variables that largely contributed to discrimination in Table 3, mostly the variables of gaze, prosody and facial expressions were included. This implies that observation of diverse nonverbal information is required. The selected variables differ or their contribution differs in every experiment, participant and situation in both the first and second halves of an experiment. On the other hand, in the case of the same participant, a sign of the discriminant function coefficient of the variables with the highest contribution was stable even when the situation changed from "1-1" to "1-2" and the second experiment. Therefore, it is suggested that there is some consistency, for every participant at least.

Table 4. The results of the discrimination rate.

The result by the discriminant function in the first half.	Lie utterance.: 36%
	Other utterances.: 95%
The result by the discriminant function in the second half.	Lie utterances.: 79%
	Other utterances.: 78%

4 Discussion and Conclusions

In many of the previous studies, a limited number of nonverbal information was used for reading people's intention, and simple intention corresponded directly to specific nonverbal information. The lies as intention did not correspond to specific nonverbal information. In this study, we focused on 13 of the nonverbal information, about gaze (three variables), prosody (nine variables) and facial expression (one variable). These were classified the behaviour of the participants while they were playing a game of revised Indian poker and the data that was elicited was analysed by using a discriminant analysis. The result of the application of discriminant analysis by using four of the 13 variables provided a discrimination rate of about 75% - 85% to determine whether participants told a lie or not. This result suggests that we can discriminate people's lies by the minimum usage of diverse nonverbal information at the same or with more accuracy as people can. Since all the sets of main variables that largely contributed to discrimination, the variables of gaze, prosody and facial expressions were included; the necessity of observing diverse nonverbal information is suggested.

There are very few studies that discriminate lies using an experimental setting in scenes similar to actual communication. In this study, we conducted an experiment using a game in which the communication of players resembled actual communication. As a result, it is confirmed that the variables that largely contributed to discriminating lies changed with situations and in communication. On the other hand, in the case of the same participant, it is suggested that some consistency exists. From these, in order to discriminate lies in actual communication, it is necessary to pay attention to both the change of nonverbal information according to the change in communication strategies and a certain amount of consistency in every individual.

In this study, we discriminated lies as a complex intention by using a small number of nonverbal information when the situations and participants were the same. Therefore, by analysing the nonverbal information utilised in situations using the communication media, we may also communicate naturally and smoothly, like in a face-to-face communication by using communication media.

References

J.K.Burgoon, Nonverbal signals. In M.L.Knapp & G.R.Miller (Eds). *Handbook of interpersonal communication. 2ndEd*. Newbury Park,CA.:Sage. 1994.

I.Daibou. The social meaning of interpersonal communication. *Japanese Journal of Interpersonal and Social Psychology*, Vol.1:1-16, 2001.

M.Goto, K.Itou, S.Hayamizu. Speech Completion: On-demand Completion Assistance Using Filled Pauses for Speech Input Interfaces. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-2002)*, 1489-1492, 2002.

Y.Hayashi. F0 Contour and Recognition of Vocal Expression of Feelings: Using the Interjectory Word "eh". *Technical report of IEICE. Speech*, Vol. 98 No. 177:65-72, 1998.

G.R.Miller, and J.B.Stiff. Deceptive communication. Newbury Park, CA:Sage. 1993

Y.Nakajima, K.Ando, M.Koyasu, Y.Sakano, K.Shigemasu, M.Tatibana, Y.Hakoda. THE YUHIKAKU DICTIONARY OF PSYCHOLOGY. YUHIKAKU, 1999.

T.Nakamura. Analysis of Time Differences in Expressions of Spontaneous and Forced Laughter. *Techinical report of IEICE, IE*, Vol. 100 No. 34:1-8, 2000.

R.Vertegaal, R.Slagter, Gerrit van der Veer, A.Nijholt. Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 301-308, 2001.

W.Von Raffler-Engel. Aspects of Nonverbal Communication. Lisse: Swets & Zeitlinger, 1980.

T.Yotsukura, H.Uchida, H.Yamada, S.Akamatsu, N.Tetsutani, S.Morishima. A Micro-Temporal Analysis of Facial Movements and Synthesis Spontaneously Elicited and Posed Expressions of Emotion Using High-Speed Camera. *Technical report of IEICE*. Vol. 101 No. 300:15-22, 2001.

An Embodied Conversational Agent for Interactive Videogame Environments

Ian Kenny
Middlesex University
The Burroughs
London NW4 4BT
i.kenny@mdx.ac.uk

Christian Huyck
Middlesex University
The Burroughs
London NW4 4BT
c.huyck@mdx.ac.uk

Abstract

In interactive environments inhabited by an agent and a user, it is difficult for the agent to understand what the user is talking about and why. The agent needs to be aware of situational and environmental information to properly interpret the user's utterances. Our current embodied conversational agent is described along with an example of its current performance. Issues of reference resolution that future work will address are explored.

1 Introduction

Most current embodied conversational agent (ECA) research focuses on aspects of multi-modal communication made possible by the embodiment of the agent, for example gaze, gestures and the integration of verbal and body language. In addition, most existing systems separate the agent and user into separate spaces, usually due to a separation in roles, for example, vendor and buyer. Most existing ECAs also occupy a fixed location and are surrounded by objects that neither change nor move.

In a shared dynamic visual environment, such as an interactive videogame, in which a mobile agent and a user perform collaborative tasks, an ECA requires a range of additional capabilities in order to determine what the user said and why. For example, a key task for an ECA in such an environment is the identification of entities referred to by Referring Expressions (REs), e.g. definite noun phrases and pronouns. This task presents many challenges. For example, an RE may refer to an object that has not been mentioned in the dialogue, e.g. an item that is in the shared visual field. This would present a problem for almost all existing reference resolution techniques which deal only with entities in the *discourse* context. The object referred to may also be no longer co-present with the agent and user but was *previously* in the visual field. Again, the object may or may not have been mentioned in the dialogue. An object may also be referred to that was mentioned some time ago in the dialogue. Most approaches to reference resolution would not consider such an item to be particularly 'salient', but it may have some ongoing task-related relevance. In task-oriented spoken dialogue, REs may also refer to abstract entities such as actions and events. Such references are common in spoken dialogue. For example, after issuing a command to an agent the user may subsequently ask "did you do *it*?" where 'it' refers to the action requested in the command. Since most existing reference resolution strategies cannot deal with such abstract entities, they would be unable to identify the referent of the pronoun 'it' in this case.

Reference resolution techniques generally follow a two-stage approach: (1) building the discourse context of entities available for reference and (2) identifying a single referent for each RE. As stated above, most existing techniques add only linguistically evoked entities to the context and then only those evoked by noun phrases. For our agent, this will be insufficient since entities may be evoked by other sources. A major source of possible referents is the environment itself. Before a single utterance has been received from the user, information from the environment must be added to the context. The agent and the user will be able to see objects and events in the locality and will also build a visual history of previously seen entities. Decisions must be taken about which objects and events from the environment, and which properties of those entities, must be added to the context. The agent must consider the user's 'field of view'. What can the user see? Objects and actions may also be referred to that relate to the task at hand hence must also be added to the context. This requires some form of task context in addition to the discourse and visual contexts.

Furthermore, items in the agent's inventory may be referred to therefore a personal context will also be required.

An example utterance will illustrate some of the issues to be addressed. If the user commands the agent to "get the gun" (a more detailed description of our agent's domain is given later), the system must determine which particular gun is being referred to. The gun might be the one that was just being discussed hence will be represented in the discourse context. However, there may be other candidates that need to be considered. Is the user 'looking at' a gun? Is there a gun in the room that the user is *not* looking at but might assume that the agent is aware of because it is in the shared space? Is there no gun in the present location but there *was* one that the user and agent walked past in some other part of the space? Is there an enemy agent standing in front of the agent with a gun in its hand? Of course, there could be multiple candidates therefore some judgement will be needed as to the relative cognitive status of the candidates and their salience.

This paper describes the domain for our agent and the current implementation (section 3), and sets out our future plans (section 4). Firstly, the paper considers some related research in ECAs, natural language understanding systems, videogame agents and reference resolution.

2 Related Work

There is a broad range of work related to our research covering ECAs, NLU systems, videogames and reference resolution. There is a large body of embodied conversational agent research, although this generally focuses on a different type of agent from that described here. Generally such systems do not place the ECA and user in a shared dynamic environment and agents are not usually mobile. Much of this work focuses on interpreting multi-modal input and generating multi-modal output. The agents are embodied to create a more 'human-like' face-to-face communication experience. Agents include REA (Bickmore & Cassell, 2000), Gandalf (Thorisson, 2002), MACK (Stocky & Cassell, 2002) and Steve (Rickel & Johnson, 1999).

MACK, an ECA kiosk, does, in a sense, share a space with the user (i.e. the 'real world'). He can gesture to entities in the shared environment although those entities do not change their position or other properties. Furthermore, MACK is not mobile. Steve shares a virtual reality space with the user and can discuss and gesture towards items in the environment. Steve does have a limited mobility. His natural language understanding ability is limited, however.

The most advanced ECAs, in terms of overall capabilities, appear to be the virtual humans in the 'Mission Rehearsal Exercise' (Hill et al., 2001; Traum & Rickel, 2002). These agents are mobile, can give orders to 'subordinate' agents, can negotiate, are aware of events in their environment, etc. However, the agent and user do not share the same environment. The system also require a lot of equipment.

Although not conversational agents, another set of relevant systems include those that seek to interpret user utterances (usually commands) in some form of visual or text-based environment. These systems do not have 'agents' as such and the environments used are largely static and deterministic. Many, however, pay close attention to the problem of reference resolution. The best-known early system is SHRDLU (Winograd, 1972) that enabled a user to manipulate a simple 'blocks world' using commands. Kelleher describes a system utilising a visual environment containing coloured objects such as trees and houses that the user can manipulate using commands (for example, "make the red tree taller") (Kelleher, 2003). The novelty of the system is its visual saliency algorithm which enables the resolution of references to currently and formerly visible items. In contrast to SHRDLU, Kelleher's system considers only the objects that the user can see (i.e. those on the screen) or has seen recently. The system considers both the linguistic and visual contexts as possible sponsors of referents. Gabsdil, Koller and Striegnitz created a 'text adventure engine' that enables the creation of text adventure games (Gabsdil, Koller, & Striegnitz, 2002). These games generally consist of a story-based scenario, a set of locations, objects and events. The user types in commands to manipulate the objects described and the system responds by describing the updated state of the world. Since the system completely controls what the user knows about the game world, it is also able to restrict possible referents to those that the user can 'see'.

Reference resolution research has mostly focussed on identifying noun phrase antecedents of pronouns in texts, e.g. Hobbs (1986). However, research has shown that in dialogue it is possible for up to 50% of pronouns to have non-noun phrase antecedents (Byron & Allen, 1998; Eckert & Strube, 2000). Approaches to the resolution of references to abstract entities have been proposed by Eckert and Strube (Eckert & Strube, 2000), and Byron (Byron, 2002). These approaches do not, however, take into account how visual information may impact on the selection of referents. Reference resolution approaches that do take into account visual information include Kelleher's system (Kelleher, 2003), the Mission Rehearsal Exercise (Hill et al., 2001; Traum & Rickel, 2002), "Kairai" (Shinyama,

Tokunaga, & Tanaka, 2000) and CommandTalk (Stent et al., 1999).

Of these only the Mission Rehearsal Exercise involves reference resolution in a dynamic 3D environment. Byron however, has described some of the key elements of a system that can resolve references to abstract and non-abstract entities in a 3D environment (Byron, 2003). One key task to be addressed is deciding which entities to add to the expanded 'discourse context' required for visual environments (which Byron terms the 'Model'). For example, which objects, which properties of objects and which events should be added to the Model hence made available for subsequent reference. Also, once the Model has been compiled, how do we determine the relative salience of entities in the Model? For example, which properties cause an object to have greater salience? Is a larger object more salient, one that is moving, one that is moving quickly, one that changes colour, one that is relevant to an ongoing task?

Conversational agent research related to videogames tends to focus on the selection of appropriate utterances based on simulated emotion and personality (Morris, 2002; Drennan, 2003). Implemented systems are rare or non-existent. However, Zubek and Khoo (Zubek & Khoo, 2001) have implemented a 'chatterbot' in a game based on the 'Half-Life' videogame (Sierra, 1998). This agent has very limited conversational ability being based on an adapted version of ELIZA (Weizenbaum, 1966). The agent's utterances are based on typical exchanges in an online videogame environment (rapid changes of focus, bad language, spelling errors, etc.). Although the agent is mobile and shares a dynamic space with the user, it has very limited language 'understanding' ability and does not seek to resolve references to entities in its environment.

3 The Half Life Agent

The purpose of our agent is to perform collaborative tasks with a user. The agent is a 'bot' in the Half Life videogame environment (Sierra, 1998).

3.1 Environment and Domain

This environment is interactive, dynamic and 'real-time'. The user shares the space with the agent (and possibly other agents) and sees the environment from a 'first person' perspective. In our current implementation the user communicates with the agent via the keyboard.

The environment contains objects such as guns and ammunition, and possible actions such as shooting a gun, killing another agent, exploring rooms. The agent collaborates with the user to achieve common goals, for example the death of opponents. Actions such as killing an opponent require that the agent owns a gun and some ammunition, and has located the opponent. Owning a gun requires that the agent has located a gun and picked it up. These objects and actions could be used to represent other domains but we use the one described because it comes naturally to Half Life.

3.2 Current Implementation

To date the focus of implementation has been on the behavioural element of the agent. This follows an autonomous agent model (Allen, 1995). The key element of the autonomous agent is a spreading activation network (Maes, 1989). The network selects actions based on the spreading of activation between *competence modules*. A competence module encapsulates a primitive action, a list of preconditions, an add list, a delete list, and, possibly, a *language module*. For example, the GET_GUN competence module has a precondition of AT-LOCATION-GUN and adds OWNS-GUN. The network selects for action the competence module with the highest activation. The state and goals inject activation into the network that then spreads activation between the competence modules based on successor, predecessor and conflictor links. In the current implementation, conversational actions are chosen incidentally when 'physical' actions are selected (via the language module).

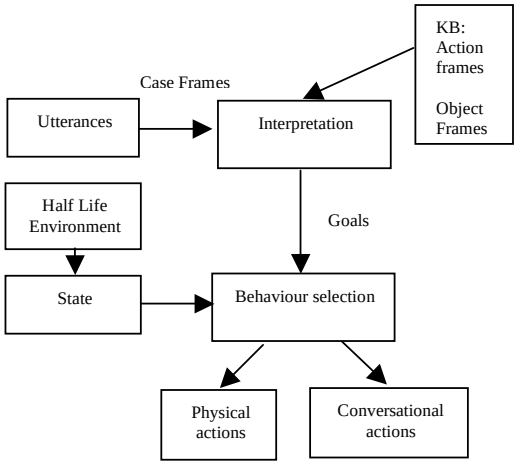


Figure 1: Conceptual overview

The current system, figure 1, has a simple conversational capability. One of our aims is to select conversational actions in the same manner as other actions via spreading activation to fulfil conversational goals. Currently, the agent produces template-based utterances to respond to the user. These responses comprise acknowledgements that it has understood a command, information about sub-goals that must be fulfilled if it cannot fulfil the given goal

immediately, and a confirmation when a goal has been achieved.

The network is set to respond to goals, i.e. goals inject more activation than the state. Goals are determined by interpreting the utterances of the user. Whilst the agent's interpretation ability is at an early stage, it can respond to a set of commands. A command is mapped onto a goal by matching the case frame produced by the parser (Huyck, 2000) against action and object frames in the KB, and then identifying which competence module represents the resulting action. The state in the selected competence module's add list is added to the system's goal list. The spreading of activation will form a path that achieves the goal through the network. An advantage of the activation network approach is that such deliberative behaviour can be achieved without a traditional planning ability.

The interpretation process currently identifies what is being talked about (within limits) but not why. However, the activation network also offers potential for intention recognition. The network is implemented as a graph that relates competence modules to each other via successor and predecessor links etc., hence it can also be used to identify the relationships between actions. Thus, if the agent is asked to "get a gun", links in the network will show that the state of owning a gun is a precondition for killing the enemy, perhaps aiding intention recognition.

3.3 Example scenario

Here we consider a simplified scenario in which the user commands the agent to "kill the enemy". The utterance is passed to the parser that returns a case frame representing the semantics. The case frame looks something like this:

KILL (OBJECT ENEMY)

The system checks that KILL is a valid action in this domain. If the player had said "shoot", or a range of other words, the system would resolve these to the action KILL. In this domain, KILL is the fundamental sense of those verbs. An action frame for KILL is retrieved that identifies that it requires an object. This is identified from the parsed input as ENEMY. ENEMY is checked to see if it is a valid object in this domain, which it is. The object frame for ENEMY is then retrieved.

The system has now identified the action that the utterance is requesting and identified the class of object that the action is to be applied to. It needs to identify the particular enemy object that is being referenced. At this stage there is only one enemy in the environment so the reference is easily resolved. Effective reference resolution is the key aim of our future work.

Now we know which action to perform on which object. The agent will respond to the player with an acknowledgement. If any of the above stages had failed the agent would have announced that it did not understand the utterance. We now need to locate the competence module in the network that represents the action and place the state in its add list into the goal list. The system locates the KILL_ENEMY module. If the state brought about by this action is ENEMY-DEAD, this is now the system's goal. The system will identify whether the preconditions of the KILL_ENEMY module are true. For simplicity's sake, say there is one precondition: OWNS-GUN. If this is not true (i.e. the agent does not have a gun) the network will automatically seek to make this true by spreading activation backwards from the goal to the KILL_ENEMY module and from that to its predecessors. If the competence module GET_GUN adds the state OWNS-GUN then this module will receive activation from the KILL_ENEMY module. The GET_GUN module will be executable because there are no preconditions to this action. The agent will announce that it needs to get a gun first and will go about doing this. Once this goal has been achieved (i.e. OWNS-GUN is true), the KILL_ENEMY module will now be executable. It will have been accumulating activation from the goal and will now become active. The agent will now seek the enemy and try to kill it.

4 Future Work

The key focus of our work is reference resolution in a shared dynamic environment. Our future work consists mostly of adding this capability to our existing agent. The overall approach will be to gradually extend the referring expressions that the system can deal with (i.e. indefinite noun phrases, definite noun phrases and pronominal references to non-abstract and abstract entities). At the same time the types of speech act that can be handled will be extended from commands to assertions and queries. The key elements of this task are the mechanisms for compiling the various contexts required for evoked entities and the mechanisms for selecting a single referent for each referring expression. Our approaches will be based on existing techniques where applicable with extensions to handle the particular problems presented by a mobile agent in a shared dynamic environment.

We also intend to place the selection of conversational behaviours onto the same footing as 'physical' actions. We are considering whether this requires a separate activation network or would benefit from integration with the existing network.

In addition, we wish to collaborate with other researchers in order to share ideas and, possibly, implementations of elements of such a system. This may be through the use of entirely different types of domain, or different videogame domains.

5 Conclusions

Placing an embodied conversational agent and a user in the same environment, where they are both mobile and can take actions, raises concerns not generally addressed in the research. When agents are able to factor situational and environmental information about objects and events into the interpretation process, more robust interpretation will be possible.

Our current system can be used to explore these questions. The next stage in our work is a more complete interpretation process for the agent. This will lead to a more useful and believable agent.

References

J. Allen. Natural Language Understanding (2nd Edition ed.): The Benjamin/Cummings Publishing Company Inc. 1995.

T. Bickmore, & J. Cassell. "How About This Weather?" Social Dialogue with Embodied Conversational Agents. *Paper presented at the AAAI Symposium on Socially Intelligent Agents*. 2000.

D. K. Byron. Resolving Pronominal Reference to Abstract Entities. *Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia. 2002.

D. K. Byron. Understanding Referring Expressions in Situated Language: Some Challenges for Real-World Agents. *Paper presented at the First International Workshop on Language Understanding and Agents for the Real World*, Hokkaido University. 2003.

D. K. Byron, & J. Allen. Resolving Demonstrative Pronouns in the TRAINS93 corpus. *Paper presented at the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*. 1998.

P. Drennan. Conversational Agents: Creating Natural Dialogue between Players and Non-Player Characters. In S. Rabin (Ed.), *AI Game Programming Wisdom 2*. Hingham, MA: Charles River Media. 2003.

M. Eckert, & M. Strube. Dialogue Acts, Synchronising Units and Anaphora Resolution. *Journal of Semantics, 17*(1): 51-89. 2000.

M. Gabsdil, A. Koller, & K. Striegnitz. Natural Language and Inference in a Computer Game. *Paper presented at the 19th COLING*, Taipei. 2002.

R. W. Hill, J. Gratch, S. Marcella, J. Rickel, W. Swartout, & D. Traum. Virtual Humans in the Mission Rehearsal Exercise. *Kynstliche Intelligenz, 17*(4). 2001.

J. Hobbs. Resolving Pronoun Reference. In *Readings in Natural Language Processing*: Morgan Kaufman. 1986.

C. R. Huyck. A Practical System for Human-Like Parsing. *Paper presented at the European Community AI Conference*. 2000.

J. Kelleher. A Perceptually Based Computational Framework for the Interpretation of Spatial Language in 3D Simulated Environments. Dublin City University, Dublin. 2003.

P. Maes. How To Do the Right Thing. *Connection Science Journal, 1*(3): 291-323. 1989.

T. W. Morris. Conversational Agents for Game-Like Virtual Environments. *Paper presented at the AAAI 2002 Spring Symposium on Artificial Intelligence and Interactive Entertainment*. 2002.

J. Rickel, & W. L. Johnson. Virtual Humans for Team Training in Virtual Reality. *Paper presented at the Ninth World Conference on AI in Education*. 1999.

Y. Shinyama, T. Tokunaga, & H. Tanaka. "Kairai" - Software Robots Understanding Natural Language. *Paper presented at the Third Workshop On Human-Computer Conversation*, Italy. 2000.

Sierra. Half Life, from http://games.sierra.com/games/half-life/. 1998.

A. Stent, J. Dowding, J. M. Gawron, E. O. Bratt, & R. Moore. The CommandTalk Spoken Dialogue System. *Paper presented at the 37th Annual Meeting of the Association for Computational Linguistics*, UMIACS, MD. 1999. June 1999.

T. Stocky, & J. Cassell. Shared Reality: Spatial Intelligence in Intuitive User Interfaces. *Paper presented at the Intelligent User Interface Conference (IUI '02)*. 2002.

K. R. Thorisson. Machine Perception of Multi-Modal Natural Dialogue. In P. McKevitt (Ed.), *Language, Vision and Music*. Amsterdam: John Benjamins. 2002.

D. Traum, & J. Rickel. Embodied Agents for Multi-party Dialogue in Immersive Worlds. *Paper presented at the AAMAS '02*, Bologna, Italy. 2002. July 15-19.

J. Weizenbaum. ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM, 9*(1): 35-36. 1966.

T. Winograd. Understanding Natural Language. *Cognitive Psychology, 3*(1): 1-191. 1972.

R. Zubek, & A. Khoo. Making the Human Care: On Building Engaging Bots. *Paper presented at the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*. 2001.

Identifying Affectemes: Transcribing Conversational Behaviour.

Lesley Axelrod Kate Hone
Brunel University, Uxbridge Campus,
Department of Information Systems and Computing
Uxbridge UB8 3PH
lesley.axelrod@brunel.ac.uk; kate.hone@brunel.ac.uk

Abstract

As we spend increasing amounts of time interacting with technologies, we require them to enhance our engagement, enjoyment and fun. Systems that display a degree of social intelligence, including emotional interaction, may improve usability, playability, and help to sustain our interaction. We need to assess users' communicative intentions and affective states during interaction. Successful conversation depends not only on words, but on continual transmission of affective information via numerous communication channels. Humans have a colossal tacit knowledge of this interplay, but fully analyzing and measuring such complex interaction is problematic and lengthy. This paper describes the development of a practical methodology to assess user behaviour and communicative intent during human-computer interaction. Interaction Analysis techniques identify discrete affective messages 'affectemes' and their components. A pilot of this methodology is described, examining user interaction with standard and Wizard-of-Oz (WOZ) generated affective systems.

1 Introduction

Emotion recognition in computing (ERIC) is the focus of a project at Brunel University. It explores the use of recognition technologies from a user's perspective and empirically tests assumptions about human behaviour which underlie their use. It aims to establish the extent to which people will naturally express themselves when they know they are interacting with a device that can recognise their affective states and to identify the conditions under which the application of such detection can lead to improvements in subjective and/or objective measures of system usability. Our first experiment includes comparison of users' communicative and affective behaviour in different conditions and has led us to review and develop methods to assess multimodal interaction.

1.1 Definitions

Despite many emotion theories, models and taxonomies, with recent contributions from interdisciplinary work and new arenas such as artificial intelligence, agreement has not yet been reached on definitions of terms such as 'emotion' 'mood' etc. Picard (1997) suggests 'affect' as an all encompassing term that we can use, even without complete agreement as to all the details of its components.

Affect consists of complex neurophysiological systems that help organise and regulate other systems such as cognition, memory and problem-solving. The affective state of an individual is critical for their functioning during work, rest and play, and there is growing recognition that organisations benefit from management of their employees' emotional labour (Bunting, 2004).

Picard (1997) uses the term 'emotional expressions' to differentiate outward expressions of emotion from the internal emotional experiences of individuals and to describe what is revealed to others by affective displays. These 'emotional expressions' are the focus of the proposed methodology.

1.2 Complex human communication

In real life situations human communication is complex, rule based, and multi-modal. Conversation is not only based on spoken or written words and the manner in which they are produced. Word meaning can be altered by context or timing or by the use of multimodal non verbal cues. Non verbal expressions can replace words in conversational interaction. As soon as we are exposed to emotional expressions we acquire assumptions and knowledge about gender, age, background, etc. and about affective states. We use our five senses as well as our background knowledge about the norms for individuals, environments, time and place.

We send and receive an interplay of:

- appearance (occulesics), eg. fashion, architecture;
- movement (kinesics) whole body, gesture, facial;
- voice (vocalics) content, tone, pitch, rate;
- touch and smell - (haptics and olfactics);
- space - (proxemics) territoriality, personal space;
- time - (chronemics) context, contagion.

Synchronous use of these modalities may duplicate, add or amend information. Such 'redundancy' increases the chance of successful communication, add emphasis to messages and richness to the experience. Eg. Baveles (1996) describes gestures used in a narrative to enhance spoken information, to symbolize movements and to add information about direction, position and force of movement.

Following social rules such as those for politeness, reciprocity and turn taking are essential to successfully sustain communication. There are different rules about non verbal behaviour for narrators and listeners, and some variations depending on factors such as personality, age, sex, culture, etc. In human to human interaction we are soon frustrated and feel negative affect if our communication partners do not communicate effectively or fail to demonstrate appropriate recognition and expression of affect.

Reeves and Nass (1998) showed that people tend to treat new media like real people. As we spend an increasing amount of time interacting with technologies in all spheres of life, users may feel increasingly alienated by applications that fail to demonstrate socially intelligent affective responses. We may need to develop enriched communicative environments if we are to extend and sustain the time spent in human- computer interaction (HCI).

1.3 Assessing emotional expressions

There are four commonly used means to assess affective interaction, each with different difficulties:

- bio-physiological measures such as heart rate, give indication of arousal, but are intrusive and difficult to apply in the real world. Data does not map to underlying emotions or emotional expressions.
- measures can be obtained from direct probes, questionnaires or narrative means, but may disrupt the normal flow of affective interaction.
- performance measures such as patterns of mouse behaviour, have been linked to individual characteristics, and click-streams used to build 'digital silhouettes' of users (Scheirer, Fernandez et al 2002, Boston.internet.com 2000).
- observational data can yield rich real world data, but is prone to bias, difficult to automate, complex and lengthy to carry out effectively. Coding systems analyse only specific modalities, (see review by the International Standards for Language Engineering project (ISLE) (2002)). Perhaps best known is Ekman's (2002) Facial Action Coding System (FACS). Time required for training and analysis, makes this impractical for everyday use by HCI practitioners. Until automated systems are efficient and multimodal (and maybe to influence their development), we need a method using human strengths to assess interaction, that is relatively quick, systematic and quantifiable.

1.4 Borrowing methodologies

Successful strategies for sequential analysis of human interaction have been developed by Bakeman and Gottman (1997). Rather than observing models and fitting observation schemes to them, they observe interactions, define mutually exhaustive and exclusive factors, and devise coding protocols to capture those factors and then build models to fit what is observed. Similar strategies could be used to assess emotional expressions during HCI.

The International Phonetic Alphabet (see IPA website, 2004) uses different levels of analysis for different purposes. The IPA was developed over a century ago by observing sounds recognized as distinctive by native speakers of a language, (phonemes) each represented by a unique symbol to form a phonetic alphabet. Phonemic transcription is quick and adequate for many situations. The exact pronunciation of any phoneme will vary depending on individual variants, such as its position in a word and the influence of nearby sounds. Variations of phonemes are called allophones and their differences can be transcribed if necessary. More detailed phonological or phonetic transcription takes longer but gives greater depth.

Our affective coding system uses both sequential analysis methods and levels of coding detail. To help us conceptualise we borrowed from the IPA terminology, calling our coding units 'affectemes' and variations of them 'allafects', broad coding and transcription 'affectemic' and analysis in greater detail, as 'affectic' or 'affectological' transcription.

2 Pilot experiment

An experiment was designed to assess the extent and effects of interaction involving display of emotional expressions at the interface. The design of the experiment was such that in some conditions the system appeared to vary its response on the basis of recognized emotional expressions at the interface, for instance providing clues when the user displayed negative affect. We hypothesized that such interventions would lead to improved task performance and improved satisfaction. The experiment used a Wizard of Oz method (where a hidden human observer controls the computer's 'affective' response) while participants completed a simple on-screen word

puzzle, both described in detail elsewhere (Axelrod & Hone 2004.) The experiment had a 2 × 2 between-subjects factorial design. The factors were:

1. ***Acted affective*** (with two levels; 'standard' vs. 'system appears affective'). This refers to whether the system appeared to act affectively. In the 'standard' condition clues and messages appeared only in response to the user clicking the 'help' button. In the 'system appears affective' condition, if the participant was observed via the one way mirror to use emotional expressions, the game was controlled so that appropriate clues or messages appeared to them.
2. ***Told affective*** (with two levels; 'expect standard system' vs. 'expect affective system'). This refers to whether the participant expected the system to act affectively. In the 'expect standard system' condition participants were told they would be testing the usability of a word game. In the 'expect affective system' condition they were told that they would be testing the usability of a word game on a prototype system, that might recognize or respond to some of their emotions.

There were therefore four experimental conditions in total, representing the factorial combination of the two factors. 60 participants were videoed for 10 minutes each.

Table 1 Participant groups – 2 X 2 factorial design

	System appears to act affectively	System appears to act as standard application
Participants told to expect affective system	Group 1	Group 2
Participants told to expect standard system	Group 3	Group 4

We wanted to identify mutually exclusive and exhaustive emotional expressions that could be counted and rated for valence and intensity. Videos were edited so that both the user and the context of their screen activities could be viewed synchronously. Transana qualitative analysis tool [Woods et al 2004] was very useful to organize video data, to link video to time-logged transcripts and to facilitate coding and reviewing of coded samples.

2.1 Coding Schema

Each 10 minute sample was reviewed in context for an overall feel of the user's individual patterns. To reduce the coding load, time sampling was used with two separate representative 1 minute episodes extracted from each 10 minute interaction (120 samples in all.) A number of transcripts were then attached to each sample.

For the first level of analysis affectemes (distinct episodes of emotional expression) were identified and start and end points were time stamped. Coder's spontaneous comments about reasons for selection and appraisal of these episodes was recorded. Each episode was then rated for perceived valence and arousal, using a rating scale adapted from Bradley & Lang's (1994) self assessment manikin.

Table 2: Transcripts and keywords used

Transcript	Keywords-affectemes	Keyword families examples of allaffects
1 audible activity	speech	
	whisper	
	affect burst	groan, whistle, tut, sigh, inbreath, snort.
	extraneous noises	Foot tap, finger tap, door bang
	data entry	keyboard, mouse click
2 whole body movement	shifts	large shift, small shifts, lean back, postural twirl
	tension	hunched shoulders
	grooming	tuck hair, shunt glasses, bite finger/thumb, scratch
	chin dump	L hand, R hand, both hands, mouth covered, fisted, fingers spread
3 head movements	gesture	nod, shake
	peer	distant, close
	aspect shifts	tilt, turn, L, R, up, down
	chin move	tuck, thrust
4 upper face	brow-raise	bilateral, unilateral L, unilateral R
	frown	slight, deep
	nose	wrinkle, flare
5 lower face	smile	PanAm, zygomatic
	mouthing words	
	fidgets	compress, pursing, rinsing, jaw grind
6 gaze and blinks	eye shifts	flashbulb eyes, narrowed, closed
	screen attention shifts	on screen, off screen, scanning, L, R, up, down
	blinks	
7 keyboard, mouse and on-screen activity	data entry	Keystrokes, mouse clicks
	on-screen activity	Picture clue, short text clue

A coding scheme was developed to reflect the behaviours exhibited involving creation of transcripts for various modalities and allocation of keywords relating to affectemes, as in table 2.

Individuals displayed a wide range of emotional expressions (fig 1) and affecteme patterns were recognised and intuitively named. For example, negative affectemes included the 'fed up chin dump' where chin or cheek was dumped into a cupped hand during contemplation of an onscreen item and the 'angry mutter'. Positive affectemes included the 'satisfied nod' and smiles of success on task completion.

Figure 1. Array of emotional expressions

Consideration was given as to what behaviours triggered perception of episodes of emotional expressions. They were perceived when:

- there was a sudden and large change in a modality, such as a sudden movement or sound;
- there were changes in a large number of modalities, for example sudden vocalisation with posture shifts, gesture and rapid blinking;
- there were no changes for a period of time.

2.2 Reliability and validity

Preliminary results for reliability and validity of this method are encouraging. Six coders rated the same five episodes to test reliability of coding episodes with close agreement as to number and onset and offset boundaries of affectemes. Results for coding scales for valence and arousal, were compared using Cohen's Kappa and agreement levels were found to range from 0.29 to 1.0, with a mode of 1.0 and a mean of 0.78. Percentage agreements ranged from 89% to 100% with a mode of 100% and a mean of 97.24.

Validity was assessed by retrospective walkthroughs by participants reviewing their own videos and coding their own episodes of emotional expression. This showed good agreement with independently coded data, particularly using rating scales for valence and arousal, for example agreeing on positive or negative. There was also some interesting discussion about the participants' appraisal and terminology for emotional states. Words used by participants and independent coders to describe episodes of emotional expressions usually shared valence, although the word used to describe the emotional state differed, for example, a participant described an episode as 'confusion' whilst a coder described the same episode as 'anger'.

3 Results

Preliminary results are encouraging. We have found that both task performance and subjective user satisfaction are significantly increased when the system appears to act affectively. When told the system may recognise affect, participants were significantly more likely to state that they had shown emotions.

Full analysis of behaviours is ongoing. We have established that even with applications they believe to be completely standard, users constantly and consistently display emotional expressions, with individual variations in frequency and type. There are individual variances in behaviours shown, and some common behaviours. It seems that participants' blink rate is significantly higher when they are told the system may recognise affect.

4 Discussion

With an inductive, top-down approach and use of human expertise, we seek to identify and address issues that developers of recognition systems must consider in order to enforce conversational involvement and extend rich interaction. Agents could similarly be assessed to identify the features that cause humans to feel dislike or discomfort.

Recognition systems currently use microanalysis of specific modalities. Micro analysis alone does not enable realistic analysis of conversational intent and such systems lack social intelligence. E.g. Ekman, Friesen et al (2002) instruct that for full analysis using FACS codes must be reviewed in context of other environmental and contextual factors.

Macro analysis can be used as an initial analysis and then be combined with microanalysis as necessary. Macro analysis involving identification of a set of common emotional expressions and their key components could lead to improved systems. For example, happiness might be conveyed by a smile involving mouth and eyes, a breathy tone of voice and frequent postural shifts. We could strip away some modalities and then compare to full modality recognition rates, so for example it might only be necessary to identify mouth movement for 'happiness' identification to take place.

Alternatively for correct identification of some emotional expressions information might be required from more than one modality. For example, without contextual information a smile might be construed as 'happy' when it signals irony or without postural information, fear and anger might be

confused if in anger there is a forward postural shift and in fear a withdrawal.

Systems that rely on one modality, such as speech or text only, may fail to identify mixed messages, a powerful and frequently employed communication strategy, eg. irony is often conveyed by a mismatch between tone of voice and choice of words. Systems need to recognise human inhibition. In context (e.g. where animation is expected) a blank face gives a powerful message. In our pilot, periods of prolonged inaction were rated as expressive episodes.

Systems should reflect habitual communication patterns. Resting expressions vary and humans automatically take this into account when assessing conversation partners, so that someone who has a 'smiley' resting face will have to smile very broadly to show a change, while someone with a more miserable expression at rest, might only need to exhibit a slight smile to achieve the same result.

We strive to convey our emotions in new digital domains, such as text and email, by developing conventions such as 'emoticons'. Viable affective recognition systems are still some way off. If we identify the most common emotional expressions, we might be able to add 'affecticons' to interfaces, allowing users to better communicate by clicking on the affecticon that reflects their current feelings.

Acknowledgements

This work is supported by EPSRC grant R81374/01.

References

Axelrod, L., Hone, K. Smoke and Mirrors: Gathering User Requirements for Emerging Affective Systems. *Proceedings of 26th International Conference on Information Technology Interfaces*, ITI 2004.

Bakeman, R., Gottman, J. M., *Observing Interaction: An introduction to sequential analysis*. Cambridge University Press 1997.

Bradley, M. M., Lang, P. J., Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *J Behav Ther Exp Psychiatry* 25(1), 49-59, 1994.

Bavelas, J. B., *Debunking Body Language: new research on non-verbal communications.* Lecture given at University of Victoria, http://novaonline.nv.cc.va.us/eli/spd110td/interper/message/linksnonverbal.html 1996.

Boston.internet.com. *Predictive Networks Wins* http://boston.internet.com/news/article.php/2371_419481 2000.

Ekman, P., Friesen, W. V., Hager, J. C., *Facial Action Coding System: Investigator's Guide.* Research Nexus. 2002.

Ekman, P., *Emotions Revealed: Recognising faces and feelings to improve communication and emotional life.* Times Books. 2003.

IPA http://www.arts.gla.ac.uk/IPA/index.html

ISLE Survey of Multimodal Coding Schemes and Best Practice *ISLE Natural Interactivity and Multimodality Working Group Deliverable D9.* 2002 http://isle.nis.sdu.dk/

Picard, R.W., *Affective Computing*. Cambridge MA: The MIT Press. 1997.

Reeves, B., Nass, C., *The Media Equation: How people treat computers, television and new media like real people and places.* CSLI Publications. 1998.

J. Scheirer, R. Fernandez, J. Klein and R. W. Picard, "Frustrating the user on purpose: A step toward building an affective computer," *Interaction with Computers* 14 (2) 2002.

Woods, D. K., Fassnacht, C, 2004 Wisconsin Centre for Education Research, TRANSANA website http://www2.wcer.wisc.edu/Transana

Towards Context-Based Visual Feedback Recognition for Embodied Agents

Louis-Philippe Morency* Candace Sidner† Trevor Darrell*

*Computer Sciences and Artificial Intelligence Laboratory
Massachussetts Institute of Technology
Cambridge, MA 02139, USA
{lmorency,trevor}@csail.mit.edu

†Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139, USA
sidner@merl.com

Abstract

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. We investigate how contextual information can improve visual recognition of feedback gestures during interactions with embodied conversational agents. We present a visual recognition model that integrates cues from the spoken dialogue of an embodied agent with direct observation of a user's head pose. In preliminary experiments using a discriminative framework, contextual information improved the performance of head nod detection.

1 Introduction

During face-to-face conversation, people use visual nonverbal feedback to communicate relevant information and to synchronize rhythm between participants. A good example of nonverbal feedback is head nodding and its usage for visual grounding, turn-taking and answering yes/no questions. When recognizing visual feedback, people use more then their visual perception. Knowledge about the current topic and expectations from the previous dialog are also included with the visual perception to recognize nonverbal cues. Our goal is to equip an embodied conversational agent (ECA) with the ability to use contextual information for performing visual feedback recognition much in the same way people do.

In the last decade, many ECAs have been developed for face-to-face interaction. A key component of these systems is the dialog manager, usually consisting of a history (also called a memory) of the past events and current state, and an agenda of the future actions (see Figure 1). The dialog manager uses contextual information to decide which verbal or nonverbal action the agent should perform next. This is called context-based synthesis.

Contextual information has proven useful for aiding speech recognition (Lemon et al., 2002). In these systems, the grammar of the speech recognizer dynamically changes depending on the agent's previous action or sentence. In a similar fashion, we want to develop a context-based visual recognition module that builds upon the contextual information available in the dialog manager to improve performance.

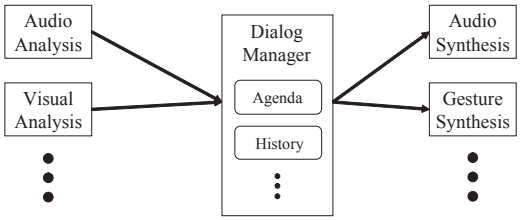


Figure 1: Simplified architecture for embodied conversational agent.

The use of dialog context for visual gesture recognition has, to our knowledge, not been explored before for conversational interaction. Here we present a model for incorporating text-based dialog cues into head-nod recognition. We exploit discriminative classifiers in our work, but other classification schemes could also fit into our general approach.

We have designed a visual analysis module that can recognize head nods based on both visual and dialog context cues. The contextual information is based on the spoken sentences of the ECA, which are readily available from the dialog manager. We use a sum-based technique to detect the head nods based on the spoken sentences and the frequency pattern of the head motion. The experiments were based on 30 video recordings of human participants interacting with an interactive robot.

There has been considerable previous work on gestures with ECA. Bickmore and Cassell (2004) de-

veloped an ECA that exhibited many gestural capabilities to accompany its spoken conversation, and could interpret spoken linguistic utterances from human users. Sidner et al. (2004) have experimented with people interacting with a humanoid robot. They found that more than half their participants naturally nodded at the robot's conversational contributions even though the robot could not interpret head nods. Nakano et al. (2003) analyzed eye gaze and head nods in a computer–human conversation and found that humans monitored the lack of negative feedback. They incorporated their results in an ECA that updated dialogue state. Numerous other ECAs (Traum and Rickel, 2002; Carolis et al., 2001) are exploring aspects of gestural behavior in human-ECA interactions. Physically embodied ECAs—for example, ARMAR II (Dillman et al., 2004, 2002) and Leo (Breazeal et al., 2004)–have also begun to incorporate the ability to recognize human gestures and track human counterparts.

Few of these systems have incorporated tracking of fine motion actions, or visual gesture, and none have included top-down dialog context in the visual recognition process. This paper describes our system for contextual-based visual feedback recognition.

2 Context-Based Visual Analysis

In general, our goal is to efficiently integrate dialog context information from an embodied agent with a visual analysis module. We define a visual analysis module as a software component that can analyze images (or video sequences) and recognize visual feedback of a human participant during interaction with an embodied agent.

Figure 1 is a general view of the architecture for an embodied conversational agent. In this architecture, the dialog manager contains two main subcomponents, an agenda and a history.

The agenda keeps a list of all the possible actions the agent and the user (i.e. human participant) can do next. This list is updated by the dialog manager based on its discourse model (prior knowledge) and on the history. Some interesting contextual cues can be estimated from the agenda:

- What will be the next spoken sentence of our embodied agent?
- Are we expecting some specific answers from the user?
- Is the user expected to look at some common space?

The history keeps a log of all the previous events that happened during the conversation. This information can be used to learn some interesting contextual cues:

- How did the user answer previous questions (speech or gesture)?
- Does the user seem to understand the last explanation?

Based on the history, we can build a prior model about the type of visual feedback shown by the user. Based on the agenda, we can predict the type of visual feedback that will be shown by the user.

Following the definitions of Cassell and Thorisson (1999) for nonverbal feedback synthesis, we outline three categories for visual feedback analysis: (1) content-related feedback, (2) envelope feedback, and (3) emotional feedback. Contextual information can be used to improve recognition in each category.

Content-related feedback is concerned with the content of the conversation. For example, a person uses head nods or pointing gestures to supplement or replace a spoken sentence. For this type of feedback, contextual information inferred from speech can greatly improve the performance of the visual recognition system. For instance, to know that the embodied agent just asked a yes/no question should indicate to the visual analysis module a high probability of a head nod or a head shake.

Grounding visual cues that occur during conversation fall into the category of envelope feedback. Such visual cues include eye gaze contact, head nods for visual grounding, and manual beat gestures. Envelope feedback cues accompany the dialog of a conversation much in the same way audio cues like pitch, volume and tone envelope spoken words. Contextual information can improve the recognition of envelope visual feedback cues. For example, knowledge about when the embodied agent pauses can help to recognize visual feedback related to face-to-face grounding.

Emotional feedback visual cues indicate the emotional state of a person. Facial expression is an emotional feedback cue used to show one of the 6 basic emotions (Ekman, 1992) such as happiness or anger. For this kind of feedback, contextual information can be used to anticipate a person's facial expression. For example, a person smiles after receiving a compliment.

We are developing a framework to integrate the contextual information of a dialog manager with the visual cues recognized by a computer-vision module. To efficiently integrate contextual information,

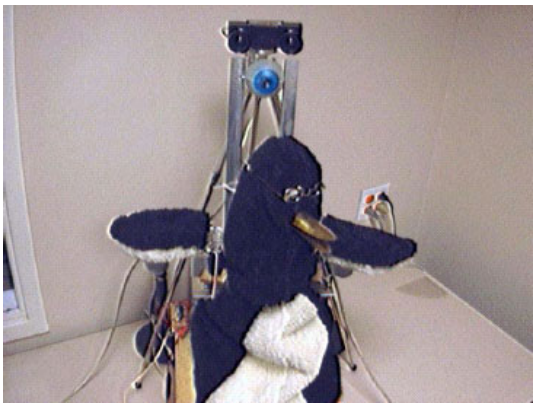


Figure 2: Mel, interactive robot used during our experiment.

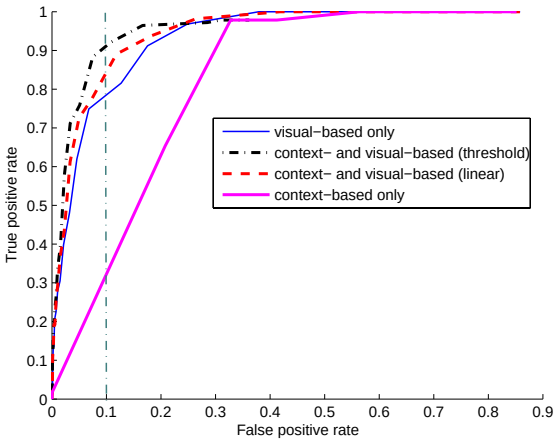


Figure 3: Recognition curves for different head nod detection algorithms.

we need to have a flexible visual recognition algorithm that can deal with multiple sources of information. In this paper, we use a simple cascade of discriminative classifiers to differentiate gestures and to learn contextual events. We show preliminary results indicating that even relatively impoverished dialog cues can have a significant impact on recognition performance. By observing the intra-sentence word-position and whether the sentence was a question, we can significantly improve recognition over visual observation alone.

3 Experiment

Our experiments demonstrate the use of contextual information inferred from an agent's spoken dialogue to improve head-nod recognition. We tested our prototype on 30 video recordings of human participants interacting with an interactive robot, Mel, developed at Mistubishi Electronic Research Lab (MERL) (see Figure 2). Mel interacted with the subject by demonstrating an invention created at MERL. Each interaction lasted between 2 and 5 minutes. Mel's conversational model, based on COLLAGEN (Rich et al., 2001), determines the next item on the agenda using a predefined set of engagement rules, originally based on human–human interaction (Sidner et al., 2003). The conversational model also uses a Sensor Fusion Model (Sidner et al., 2004) to asses engagement information about the user. This module keeps track of verbal—speech recognition—and nonverbal—head-pose estimation and head gesture recognition (Morency and Darrell, 2004)—cues.

For each subject, we had a video sequence of the complete interaction as well as the head pose and velocity for each frame. We labeled each video sequence to determine exactly when the participant noded his head, when the robot spoke, and which type of sentence was spoken by the robot (statement or question).

We trained a first discriminative classifier, a support vector machine (SVM), with data that was a good sampling of natural head gestures (Lee et al., 2004). This SVM was trained using a frequency representation of the head motion over a time window of 1 second. A second SVM was trained for each frame using input features from the spoken dialogue. In this experiment we used two contextual features from the ECA's agenda: the word position inside the spoken sentence and whether the sentence was a question or a statement. The word position is coded between 0 and 1, where 1 represents the middle of a sentence and 0 represents an extremity of the sentence or a pause (between sentences).

Our hypothesis was that including contextual information inside the head-nod detector would increase the number of recognized head nods or, equivalently, would reduce the number of false detections. We tested three different configurations: (1) using only the visual-based approach, (2) using only the contextual information as input, and (3) using the visual approach with the contextual information. Figure 3 shows results for each recognition algorithm when varying the detection threshold. We show the recognition results for two different techniques of combining the discriminative classifiers. For the first technique (black curve in Figure 3), we fixed the context-based threshold to its optimal value and varied the velocity-based threshold. For the second technique (red curve in Figure 3), we linearly combined the output of the two discriminative classifiers.

For a fixed false positive rate of 0.1, 92% of the head nods were detected by the combined approach while only 78% were detected by the visual-based approach and 32% were detected by the context-based method. These results show that we can achieve better performance when integrating contextual information in the visual feedback recognition algorithm.

4 Conclusion and Future Work

Our results show that contextual information can improve visual feedback recognition for interactions with embodied conversational agents. We presented a visual recognition model that integrates knowledge from the spoken dialogue of an embodied agent. By using simple contextual features like word positioning and question/statement differentiation, we were able to improve the performance of our head nod detector from 78% to 92% recognition rate. As future work, we would like to experiment with a richer set of contextual cues and apply our model to different type of visual feedback.

Acknowledgements

Thanks to C. Mario Christoudias.

References

Tim Bickmore and Justine Cassell. *J. van Kuppevelt, L. Dybkjaer, and N. Bernsen (eds.), Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*, chapter Social Dialogue with Embodied Conversational Agents. Kluwer Academic, 2004.

Breazeal, Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. In *The Third International Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004*, pages 1028–1035. ACM Press, July 2004.

De Carolis, Pelachaud, Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *Proceedings of IJCAI*, Seattle, September 2001.

Justine Cassell and Kristinn R. Thorisson. The poser of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 1999.

Dillman, Becher, and P. Steinhaus. ARMAR II – a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics*, 1(1):143–155, 2004.

Dillman, Ehrenmann, Steinhaus, Rogalla, and R. Zoellner. Human friendly programming of humanoid robots–the German Collaborative Research Center. In *The Third IARP Intenational Workshop on Humanoid and Human-Friendly Robotics*, Tsukuba Research Centre, Japan, December 2002.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200, 1992.

Lee, Lesh, Sidner, Morency, Kapoor, and Trevor Darrell. Nodding in conversations with a robot. In *Extended Abstract of CHI'04*, April 2004.

Lemon, Gruenstein, and Stanley Peters. Collaborative activities and multi-tasking in dialogue systems. *Traitement Automatique des Langues (TAL), special issue on dialogue*, 43(2):131–154, 2002.

Louis-Philippe Morency and Trevor Darrell. From conversational tooltips to grounded discourse: Head pose tracking in interactive dialog systems. In *Proceedings of the International Conference on Multi-modal Interfaces*, College State, PA, October 2004.

Nakano, Reinstein, Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003.

Rich, Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human–computer interaction. *AI Magazine, Special Issue on Intelligent User Interfaces*, 22(4):15–25, 2001.

Sidner, Kidd, Lee, and Neal Lesh. Where to look: A study of human–robot engagement. In *Proceedings of Intelligent User Interfaces*, Portugal, 2004.

Sidner, Lee, and Neal Lesh. Engagement when looking: Behaviors for robots when collaborating with people. In *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, pages 123–130, University of Saarland, 2003. I. Kruiff-Korbayova and C. Kosny (eds.).

D. Traum and J. Rickel. Embodied agents for multiparty dialogue in immersive virtual world. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773, July 2002.

Interactive e-Hon: Translating Web Contents into a Storybook World

Kaoru Sumi[*] Katsumi Tanaka[†]
[*] National Institute of Information and Communications Technology
Hikaridai, Seika-cho, Kyoto, Japan
kaoru@nict.go.jp
[†]Kyoto University
Yoshida-Honmachi, Kyoto, Japan

Abstract

This paper describes a medium, called Interactive e-Hon, for helping children to understand contents from the Web. It works by transforming electronic contents into an easily understandable "storybook world." In this world, easy-to-understand contents are generated automatically by creating 3D animations that include contents and metaphors, and by using a child-parent model with dialogue expression and a question-answering style comprehensible to children.

1 Introduction

We are awash in information flowing from the World Wide Web, newspapers, and other sources, yet the information is often hard to understand; laypeople, the elderly, and children find much of what is available incomprehensible. Thus far, most children have missed opportunities to use such information, because it has been prepared by adults for adults. The volume of information specifically intended for children is extremely limited, and it is still primarily adults who experience the globalizing effects of the Web and other networks. The barriers for children include difficult expressions, prerequisite background knowledge, and so on.

Our goal is to remove these barriers and build bridges to facilitate children's understanding and curiosity. When parents explain difficult ideas to their children, they choose words and concepts that their children know. Computers could potentially be applied as communication aids in this process to support children's understanding and help parents to explain . Therefore, in our research, we are presently considering the applicability of such systems for facilitating understanding in children.

This paper describes a medium, called Interactive e-Hon, for helping children to understand difficult contents. It works by transforming electronic contents into an easily understandable "storybook world." Interactive e-Hon uses animations to help children understand contents. Visual data attract a child's interest, and the use of concrete examples like metaphors facilitates understanding, because each person learns according to his or her own unique mental model [1][2], formed according to one's background. For example, if a user poses a question about something, a system that answers with a concrete example in accordance with the user's specialization would be very helpful. For users who are children, an appropriate domain might be a storybook world. Our long-term goal is to help broaden children's intellectual curiosity [3] by expanding their knowledge.

Attempts to transform natural language (NL) into animation began in the 1970s with SHRDLU [4], which represented a building-block world and showed animations of adding or removing blocks. In the 1980s and 1990s, HOMER [5], Put-that-there [6], AnimNL [7], and other applications appeared. In these applications, users operated human agents or other animated entities derived from NL understanding. Recent research has examined the natural behaviour of life-like agents in interactions between users and agents. This area includes research on the gestures of an agent [8], interactive drama [9], and the emotions of an agent [10]. The main theme in this line of inquiry is the question of how to make these agents close to a human level in terms of dialogicality, believability, and reliability.

In contrast, our research aims to make contents easier for users to understand, regardless of agent humanity. Little or no attention has been paid to such media translation from contents with the goal of improving users' understanding.

2 Interactive e-Hon.

Figure 1 shows the system framework for Interactive e-Hon. Interactive e-Hon transforms the NL of electronic contents into a storybook world that can answer questions and explain of the answers in a dialogue-based style, with animations and metaphors for concepts. Thus, in this storybook world, easy-to-understand contents are created by paraphrasing the original contents with a colloquial style, by creating animations that include contents and metaphors, and by using a child-parent model with dialogue expression and a question-answering style comprehensible to children.

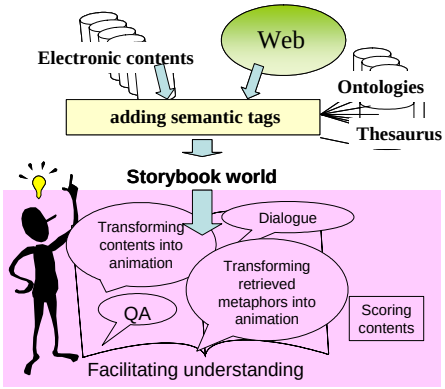


Fig. 1. Interactive e-Hon: This system transforms the natural language of electronic contents into a storybook world by using animation and dialogue expression.

Interactive e-Hon is a kind of word translation medium that provides expression through the use of 3D animation and dialogue explanation in order to help children to understand Web contents or other electronic resources, e.g., news, novels, essays, and so on. In the system's NL processing, each subject or object is transformed into a character, and each predicate is transformed into the behavior of a character. For a given content, an animation plays in synchronization with a dialogue explanation, which is spoken by a voice synthesizer.

This processing is based on text information containing semantic tags that follow the Global Document Annotation (GDA) [1] tagging standard, along with other, additional semantic tags. Tags with several semantic meanings for every morpheme, such as "length," "weight," "organization," and so forth, are used. To provide normal answers, the system searches for tags according to the meaning of a question. To provide both generalized and concretized answers, after searching the tags and obtaining one normal answer, the system then generalizes or concretizes the answer by using ontologies. Recently, the Semantic Web [11] and its associated activities have adopted tagged documentation. Tagging is also expected be applied in the next generation of Web documentation.

In the following sub-sections, we describe the key aspects of Interactive e-Hon: the information presentation model, the transformation of electronic contents into dialogue expressions, the transformation of electronic contents into animations, and the expression of conceptual metaphors by animations.

2.1 Content presentation model

Our system presents agents that mediate a user's understanding through intelligent information presentation. In the proposed model, a parent agent (mother or father) and a child agent have a conversation while watching a "movie" about the contents, and the user (or users in the case of a child and parent together) watches the agents. In this model, the child agent represents the child user, and the parent agent represents his or her parent (mother or father). For this purpose, the agents take the form of moving shadows of the parent and child. There are agents for both the user or users (avatars) and others (guides and actors), and the avatars are agentive, dialogical, and familiar [12]. Thus, we designed the system for child users to feel affinities with ages, helping them to deepen their understanding of contents.

According to the classification scheme of Thomas Rist [13], a conversational setting for users and agents involves more cooperative interaction. This classification includes various style of conversation, e.g., non-interactive presentation, hyper-presentation/dialogue, presentation teams, and multi-party, multi-threaded conversation. With its agents for the users and for others, and with its process of media transformation from contents (e.g., question-answering, dialogue, and animation), Interactive e-Hon corresponds to a multi-party, multi-threaded conversation. In addition, it has a close relationship with the contents being explained.

2.2 Transformation from contents into dialogue expressions

To transform contents into dialogues and animations, the system first generates a list of subjects, objects, predicates, and modifiers from the text information of a content. It also attempts to shorten and divide long and complicated sentences.

Then, by collecting these words and connecting them in a friendly, colloquial style, conversational sentences are generated. In addition, the system reduces the level of repetition for the conversational

[1] http://i-content.org/GDA
Internet authors can annotate their electronic documents with a common, standard tag set, allowing machines to automatically recognize the semantic and pragmatic structures of the documents.

reduces the level of repetition for the conversational partner by changing phrases according to a thesaurus. It prepares explanations through abstraction and concretization based on ontologies, meaning that it adds explanations of background knowledge. For example, in the case of abstraction, "Antananarivo in Madagascar" can be changed into "the city of Antananarivo in the nation of Madagascar," which uses the ontologies, "Antananarivo is a city," and "Madagascar is a nation." Similarly, in the case of concretization, "woodwind" can be changed into "woodwind; for example, a clarinet, saxophone, or flute." These transformations make it easier for children to understand concepts.

In the case of abstraction, the semantic tag "person" adds the expression, "person whose name is"; "location" adds "the area of" or "the nation of"; and "organization" adds "the organization of". In the case of concretization, if a target concept includes lower-level concepts, the system employs explanations of these concepts.

2.3 Transformation of contents into animations

Interactive e-Hon transforms contents into animations by using the word list described in the previous sub-section. In an animation, a subject is treated as a character, and a predicate is treated as the action. An object is also treated as a character, and an associated predicate is treated as a passive action. One animation and one dialogue are generated for each list, and these are then played at the same time.

Many characters and actions have been recorded in our database. A character or action involves a one-to-many relationship. Various character names are linked to each character. Various action names are linked to each action, because often several different names indicate the same action. Actions can be shared among characters in order to prepare a commoditized framework of characters.

If there is a word correspondence between the name of a character and a subject or object in the list, the character is selected. If there is no word correspondence, in the case of the semantic tag "person," the system selects a general person character according to an ontology of characters. When there is no semantic tag of "person," the system selects a general object, also according to an ontology of characters.

2.4 Searching and transformation of metaphors into animations

If a user does not know the meaning of a term like "president," it would be helpful to present a dialogue explaining that "a president is similar to a king in the sense of being the person who governs a nation," together with an animation of a king in a small window, as illustrated in Fig. 2. People achieve understanding of unfamiliar concepts by transforming the concepts according to their own mental models [1][2]. The above example follows this process.

The dialogue explanation depends on the results of searching world-view databases. These databases describe the real world, storybooks (with which children are readily familiar), insects, flowers, stars, and so on. The world used depends on a user's curiosity, as determined from the user's input in the main menu. For example, "a company president controls a company" appears in the common world-view database, while "a king reigns over a country" appears in the world-view database for storybooks, which is the target database for the present research. The explanation of "a company president" is searched for in the storybook world-view database by utilizing synonyms from a thesaurus. Then, the system searches for "king" and obtains the explanation, "A company president, who governs a company, is similar to a king, who governs a nation." If the user asks the meaning of "company president," the system shows an animation of a king in a small window while a parent agent, voiced by the voice synthesizer, explains the meaning by expressing the results of the search process.

In terms of search priorities, the system uses the following order: (1) complete correspondence of an object and a predicate; (2) correspondence of an object and a predicate, including synonyms; (3) correspondence of a predicate; and (4) correspondence of a predicate, including synonyms.

Commonsense computing [14] is an area of related research on describing world-views by using NL processing. In that research, world-views are transformed into networks with well-defined data, like semantic networks. A special feature of our research is that we directly apply NL with semantic tags by using ontologies and a thesaurus.

3 Application to Web contents

Web contents can easily be created by anybody and made available to the public. These contents differ from publications, which are written by professional writers and edited by professional editors, in that they are not always correct or easy to understand. Because these contents may include errors and unknown words (like neologisms, slang words, and locutions), they tend to be ill-defined. In this section, we thus discuss practical problems and solutions in transforming Web contents into a storybook world.

For example, we might try to transform the actual content, "the origin of the *teddy bear*'s name," from a Web source into an animation and a dialogue (Fig. 2).

Fig. 2. A sample view from Interactive e-Hon. (adapted from the original Japanese version)

In this case, e-Hon is explaining the concept of a "president" by showing an animation of a king. The mother and child agents talk about the contents. The original text information can be seen in the text area above the animation. The user can ask questions to the text area directly.

The following is a dialogue explanation for this example:

"This is the teddy bear's story. Do you know what a teddy bear is?"
"No, what is it?"
"It's a stuffed toy bear. This is the story of why it is called a teddy bear."
"Tell me the story."
"Well, it comes from a president of the United States, a country."
"What is a president?" (Here, an animation using the retrieved metaphor is played.)
"A president is similar to a king as a person who governs a country."

3.1 Transformation of Web contents into dialogues

As described above, the system first generates a list of subjects, objects, predicates, and modifiers from a content's text information; it then divides the sentences in the text. For example, it might generate the following lists from the long sentences shown below:

(Original Sentence 1)
"It is said that a confectioner, who read the newspaper, made a stuffed bear, found the nickname "Teddy," and named it a "Teddy bear."
(List 1) MS: modifier of subject; S: subject; MO: modifier of object; O: object; MP: modifier of predicate; P: predicate.

- S: confectioner, MS: who read the newspaper, P: make, O: stuffed bear;
- S: confectioner, P: find, O: nickname "Teddy," MO: his;
- S: confectioner, P: name, MP: "Teddy bear";
- S: it, P: said.

(Original Sentence 2)
"But, the president refused to shoot the little bear and helped it."
(List 2)
- S: president, P: shoot, O: little bear;
- S: president, P: refuse, O: to shoot the little bear;
- S: president, P: help, O: little bear.

The system then generates dialogue lines one by one, putting them in the order (in Japanese) of a modifier of the subject, the subject, a modifier of an

object, the object, a modifier of the predicate, and the predicate, according to the line units in the list. To provide the characteristics of storytelling, the system uses past tense and speaks differently depending on whether the parent agent is a mother or a father.

Sometimes the original content uses reverse conjunction, as with "but" or "however" in the following example: "but…. what do you think happens after that?"; "I can't guess. Tell me the story." In such cases, the parent and child agents speak by using questions and answers to spice up the dialogue. Also, at the ending of every scene, the system repeats the same meaning with different words by using synonyms.

3.2 Transformation of Web contents into animations

In generating an animation, the system combines separate animations of a subject as a character, an object as a passive character, and a predicate as an action, according to the line units in the list.

For example, in the case of Original Sentence 2 above, first,

- president (character) shoot (action)
- little bear (character; passive) is shot (action; passive)

are selected. After that,

- president (character) refuse (action)

is selected. Finally,

- president (character) help (action)
- little bear (character; passive) is helped (action; passive)

are selected.

This articulation of animation is used only for verbs with clear actions. For example, the be-verb and certain common expressions, such as "come from" and "said to be" in English, cannot be expressed. Because there are so many expressions like these, the system does not register verbs for such expressions as potential candidates for animations.

3.3 Handling errors and unknown words

One problem that Interactive e-Hon must handle is dealing with errors and unknown words from Web contents, such as neologisms, slang words, locutions, and new manners of speaking. The text area in the system shows original sentences. Erroneous words and unknown words are thus shown there, but they are exempt from concept explanation by metaphor expression.

In generating dialogue expressions using such words, the resulting dialogues and animations may be strange because of misunderstood modification. In the case of a subject or predicate error, an animation cannot be generated. In the Interactive e-Hon system, if an animation is not generated, the previous animation continues to loop, so errors may prevent the animation from changing to match the expressions in a dialogue. If both the animation and the dialogue work strangely, the text area helps the user to guess the original meaning and the reason for the problem. In addition, new or unknown words can be registered in the NL dictionary, the animation library, and the ontologies.

In fact, our example of "the origin of the teddy bear's name" from the Web may exhibit some errors in Japanese, such as the equivalent of "Teodore Roosevelt" or "Othedore Roosevelt". In such cases, since the original text is shown in the text area, and most of the variant words corresponding to "Roosevelt" are related to "the president," this was not a big problem.

4 Experiment using subjects

We conducted an experiment using real subjects to examine whether Interactive e-Hon's method of expression through dialogue and animation was helpful for users. We again used the example of "the origin of the teddy bear's name." Three types of contents were presented to users and evaluated by them: the original content read by a voice synthesizer (content 1), a dialogue generated by Interactive e-Hon and read by the voice synthesizer (content 2), and a dialogue and animation generated by Interactive e-Hon and read by the voice synthesizer (content 3). We still had open problems, namely, the questions of (1) what sort of media could humans understand easily at the very beginning, and (2) what kinds of cases led them to change their evaluations. The subjects were Miss T and Miss S, both in their 20s; child K, five years old; and child Y, three years old.

Both women understood content 2 as a dialogue but found content 1 easier to understand because of its compaction. They also found content 3 easier to understand than content 2 because of its animation. Miss T, however, liked content 1 the best, while Miss S favored content 3. As T commented, "Content 1 is the easiest to understand, though content 3 is the most impressive." In contrast, S commented, "Content 3 is impressive even if I don't hear it in earnest. Content 1 is familiar to me like TV or radio." She also noted, "The animations are impressive. I think the dialogues are friendly and may be easy for children to understand."

Child K (five years old) said that he did not understand content 1. He felt at first that he understood content 2 a little bit, but he could not express it in his own words. He found content 3, however, entirely different from the others, because he felt that

he understood it, including the difficult word *kobamu* in Japanese, which means "refuse." Child Y (three years old) showed no recognition of contents 1 and 2, but he seemed to understand content 3 very well, as he was able to give his thoughts on the content by asking (about President Roosevelt), "Is he kind?"

In this experiment, we observed that there was a difference in the results between adults and children, despite the limited number and age range of the subjects. At first, we thought that all users would find it easiest to understand content 3 and would like it and be attracted by it. In fact, the results were different. We clearly observed that adults, who understood the original contents, and children, who did not, had different reactions.

We assume that contents that are within a user's background knowledge are easier to understand through regular reading, as in the case of the adults in this experiment. In contrast, for contents outside a user's background knowledge, animation is expected to be very helpful for understanding, as in the case of the children. Further experiments may show that for a given user, difficult contents outside the user's background knowledge can be understood through animation, regardless of the user's age.

5 Evaluation

Interactive e-Hon's method of expression through dialogue and animation is based on NL processing of Web contents. For dialogue expression, the system generates a plausible, colloquial style that is easy to understand, by shortening a long sentence and extracting a subject, objects, a predicate, and modifiers from it. For animation expression, the system generates a helpful animation by connecting individual animations selected for the subject, objects, and predicate. The result is expression through dialogue with animation that can support a child user's understanding, as demonstrated by the above experiment using real subjects.

In the process of registering character data and corresponding words, or an action and its corresponding words, which are one-to-many correspondences, certain groups of words that are like new synonyms are generated via the 3D contents. These groups of synonyms are different from NL synonyms, and new relationships between words can be observed. This can be considered for a potential application as a more practical thesaurus based on 3D contents, as opposed to an NL thesaurus.

Reference terms (e.g., "it," "that," "this," etc.) and verbal omission of a subject, which are open problems in NL processing (NLP), still remain as problems in our system. As a tentative solution, we manually embedded word references in the GDA tags. A fully automatic process knowing which words to reference will depend upon further progress in NLP.

As for the process of transforming dialogues, Interactive e-Hon generates all explanations of locations, people, and other concepts by using ontologies, but granular unification of the ontologies and user adaptations should be considered from the perspective of determining the best solution for a given user's understanding.

6 Conclusion

We have introduced Interactive-e-Hon, a system designed for facilitating children's understanding of electronic contents by transforming them into a "storybook world." We have conducted media transformation of actual Web contents and demonstrated the effectiveness of this approach via an experiment using real subjects. We have thus shown that Interactive e-Hon can generate satisfactory explanations of concepts by applying both animations and dialogues that can be readily understood by children.

Interactive e-Hon could be widely applied as an aid to support the understanding of difficult contents or concepts by various kinds of people with different levels of background knowledge, such as the elderly, people from different regions or cultures, or laypeople in a difficult field.

As future works, we will consider expanding the databases of animations and words and applying Interactive e-Hon to several other kinds of contents.

References

1. Philip N. Johnson-Laird: Mental Models, Cambridge: Cambridge University Press. Cambridge, Mass.: Harvard University Press (1983).

2. D. A. Norman, The Psychology of Everyday Things, Basic Books (1988).

3. Hatano and Inagaki: Intellectual Curiosity, Cyuko Shinsho (in Japanese) (1973).

4. Terry Winograd, Understanding Natural Language, Academic Press (1972).

5. Vere, S. and Bickmore, T: A basic agent. Computational Intelligence, 6:41-60 (1990).

6. Richard A. Bolt: "Put-that-there": Voice and gesture at the graphics interface, International Conference on Computer Graphics and Interactive Techniques archive, Proceedings of the 7th annual conference on computer graphics and interactive techniques, ACM Press (1980).

7. N. Badler, C. Phillips, and B. Webber, Simulating Humans: Computer Graphics, Animation and Control. Oxford University Press (1993).

8. Justine Cassel et al.: BEAT: the Behavior Expression Animation Toolkit, Life-Like Characters, Helmet Prendinger and Mitsuru Ishizuka Eds., pp. 163-187, Springer (2004).

9. Hozumi Tanaka et al: Animated Agents Capable of Understanding Natural Language and Performing Actions, Life-Like Characters, Helmet Prendinger and Mitsuru Ishizuka Eds., pp. 163-187, Springer, (2004).

10. Stacy Marsella, Jonathan Gratch and Jeff Rickel: Expressive Behaviors for Virtual World, Life-Like Characters, Helmet Prendinger and Mitsuru Ishizuka Eds., pp. 163-187, Springer (2004).

11. D. Fensel, J. Hendler, H. Liebermann, and W. Wahlster (Eds.) Spinning the Semantic Web, MIT Press (2002).

12. Toyoaki Nishida, Tetsuo Kinoshita, Yasuhiko Kitamura and Kenji Mase: Agent Technology , Omu Sya (in Japanese) (2002).

13. Thomas Rist et al.: A Review of the Development of Embodied Presentation Agent and Their Application Fields, Life-Like Characters, Helmet Prendinger and Mitsuru Ishizuka Eds., pp. 377-404, Springer (2004).

14. Hugo Liu and Push Singh: Commonsense reasoning in and over natural language. Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES-2004) (2004).

Sustainable Knowledge Globe: A System for Supporting Content-oriented Conversation

Hidekazu Kubota Yasuyuki Sumi Toyoaki Nishida

Dept. of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan.
kubota@ii.ist.i.kyoto-u.ac.jp sumi@i.kyoto-u.ac.jp nishida@i.kyoto-u.ac.jp

Abstract

The purpose of this paper is to support a sustainable conversation. From a view point of sustainability, it is important to manage huge conversation content such as transcripts, handouts, and slides. Our proposed system, called Sustainable Knowledge Globe (SKG), supports people to manage conversation content by using geographical arrangement, topological connection, contextual relation, and a zooming interface. By using SKG, a user can construct his content in the virtual landscape, and then explore the landscape along a conversational context. This paper describes the approaches to realize the concept above, and discusses the effectiveness of SKG based on experiments.

1 Introduction

A sustainable conversation is indispensable for piling our knowledge upon knowledge. In a conversation, sustainability means production of conversation resources for the next conversation. Consider a regular meeting for example. People generally write a transcript (or notes) of a meeting to build up a store of knowledge. They cannot continue to discuss constructively without transcripts of past meetings. There are many other examples of conversation resources such as a proceedings of an annual conference, notes of educational dialogue, a slide (OHP or PowerPoint), a handout, and so on.

The purpose of this paper is to support a sustainable conversation. From a view point of sustainability, a sustainable conversation is content-oriented. In such a conversation, the speaker explores in a lot of past content according to current context, and then pile new idea upon them. However, he cannot manage the past content when it grows too large. It is difficult for the speaker to follow the context in huge content because it lacks a good overview. To solve the problem above, we proposes a method for exploring wide topics along the variable context in huge contents. Our essential idea is to manage conversation content in a virtual landscape. The spatial world is generally a good container of a lot of objects. People can look over objects if they are arranged spatially. People can grasp a location of an object by using a lot of spatial clues such as left and right, high and low, near and far, and so on. Therefore the virtual landscape is expected to be effective for exploring huge content intuitively.

We have developed a system for supporting content-oriented conversation, called Sustainable Knowledge Globe (SKG). By using SKG, a user can construct his content in a virtual landscape, and can also explore the landscape along a conversational context. The appearance of the landscape is a global surface like a terrestrial globe that is familiar for us. SKG enables a user to organize the conversation content on a global surface by using geographical arrangement, topological connection, and contextual relation. SKG also enable a user to look over huge content by using a zooming interface.

2 Previous works

There are many presentation tools for constructing conversation content. Microsoft PowerPoint manages sequential slides. A web browser manages linked documents. They are good tool for a presentation that is prepared in advance, however a conversation tends to spread widely from the supposed context.

Workscape (Ballay, 1994), Web Forager (Card, Robertson, and York, 1996), and Data Mountain (Robertson, Czerwinski, Larson, Robbins, and Dantzich, 1998) are the 3-dimensional file management systems using spatial clues, however they are not suitable to support a conversation because they

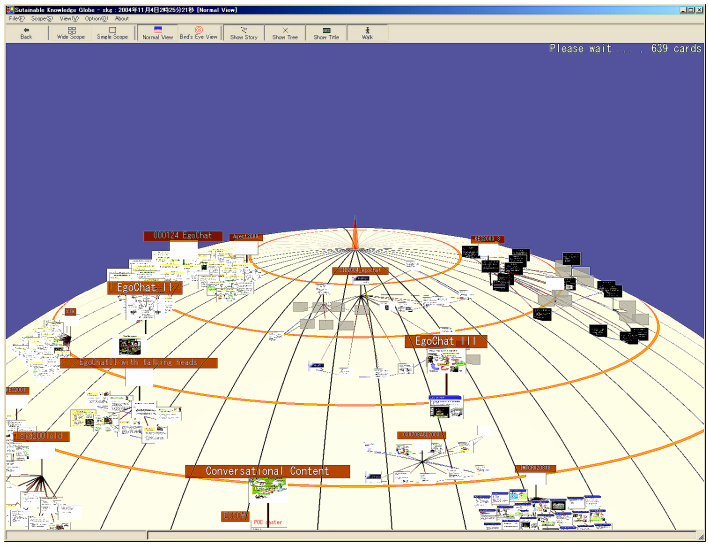


Figure 1: Screen shot of SKG

cannot manage a sequential content. A sequential content is important for a presenter to follow the story. And besides, they can manage at most 100 content. It is too little content for people to have a conversation in multiple contexts.

3 Sustainable Knowledge Globe

We have developed Sustainable Knowledge Globe (SKG) that is a system for supporting content-oriented conversation. Figure 1 is a screenshot of SKG. The shape of SKG's virtual landscape is similar to a terrestrial globe. Each content item is represented as a content card, and laid on the global surface. Here a user can explore and construct his content on any place of the surface by rotating and zooming the globe.

The virtual landscape consists of content cards, their geographical arrangement, their topological connection, and their contextual relation (Figure 2). The geographical arrangement of the cards enables a user to grasp locations of the cards by using spatial clues. The topological connection between the cards shows a narrative structure of a conversation. A tree structure represents a category of the cards. In Figure 2, A1, A2 and A3 is the children of the card A, and A4 represents a subcategory of the card A. A story structure is the other topological structure that shows a story line of the cards. The sequence from A1 to C2 is a story. The tree structure is clear for the user to explain hierarchical content, and the story structure is helpful to follow the story line of the presentation.

The contextual relation gives notice of the context of a conversation to a user. Geographically neighbour cards loosely relate with each other. Such neighbourhood are helpful for a productive digression.

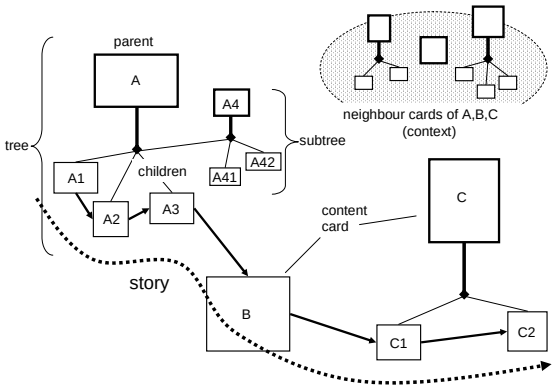


Figure 2: Model of the virtual landscape

Moreover, SKG provides a zooming interface to visualize huge content. It is indispensable to store the conversation content accumulatively.

The followings subsections describe how SKG supports a content-oriented conversation in detail.

3.1 Content card

A piece of content is called a "content card". A content card consists of three parts: an embedded file, a title of a card, and an annotation of a card. Figure 3 is a description of a content card. The <card> element represents a unit of a content card. The <card> element contains three child element: <title>, <url>, and <annotation>. The <title> element contains a title text of a card, the <url> element contains an URL of an embedded file (e.g. a document, an image, a movie clip or a slide and so on), and the <annotation> element contains an annotation text of a card. On the global surface, the content card stands upright with a thumbnail image (Figure 4).

```
<?xml version="1.0" encoding="UTF-8"?>
<card version="1.0">
<title>Overview of EgoChatIII</title>
<url> files\5_2.jpg</url>
<annotation>
We have developed EgoChatIII. EgoChatIII  is a
system for  a virtual conversation between humans
and agents.  This slide shows overview of
EgoChatIII.
</annotation>
```

Figure 3: Description of a content card

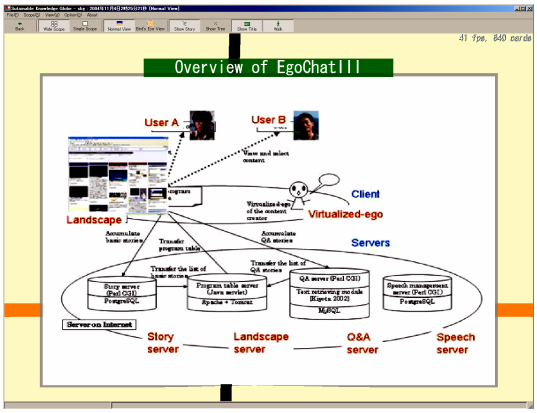


Figure 4: The closest view of a card

The content card is an extended form of a knowledge card (Kubota, 2002). A knowledge card allows only a picture file, besides a content card allows a document file to make it more expressive. The simplest method for creating a new content card is dragging and dropping a file from the desktop. The user can also create cards using a built-in card editor and exchange his cards with other users via the network server system.

3.2 Geographical arrangement

The content space of SKG is a sand-colored sphere with latitude and longitude lines, the landmarks for the north and south poles, and the equator. This is a virtual globe like a terrestrial globe, but not same as the earth. On this virtual globe, the user can move his content anywhere by using the mouse.

The followings are why SKG uses such a virtual globe for content management. The first argument is about the dimension of the space, and the second is about the shape of the land surface.

3.2.1 Dimension of the space

For managing files on generic PC systems, the 2-dimensional desktop metaphor is used. However, human depth perception doesn't effectively work in 2-dimensional representation. Consequently, there are many studies of 3-dimensional virtual space for managing files. They can be classified into two groups by using the land or not. Workscape (Ballay, 1994), Web Forager (Card, Robertson, and York, 1996), and Data Mountain (Robertson, Czerwinski, Larson, Robbins, and Dantzich, 1998) locate files on the virtual land. The land is a good point of reference for a user to grasp the geographical location of files, while it restricts available space onto the land. Cone Trees (Robertson, Mackinlay, and Card, 1991) freely locate files in a virtual space, while there is no foothold. It is suitable to manage abstract connections such as a hierarchical structure of files. From the consideration above, SKG uses 3-dimensional virtual space with the land because we focused on geographical arrangement of content.

3.2.2 Shape of the land surface

There could be many kinds of topologies of the land surface: a finite plane, an infinite plane, 2-dimensional torus, a sphere, and so on. In our prototype system (Hur, 2004) that uses a finite plane, it seems to be difficult for a user to expand the content on the edges of the plane. An infinite plane doesn't have this problem, however it is difficult for people to grasp infinite space. 2-dimensional torus also doesn't have edges, however such topology may be unfamiliar for people.

Therefore, SKG uses a sphere like a terrestrial globe that is familiar for us to explore by using latitude and longitude.

3.3 Topological connection

In addition to the geographical arrangement, the topological connection is also needed to organize the content. SKG enables a user to connect any cards by using a tree structure and a story structure.

3.3.1 Tree structure

A tree structure is a standard way of representing categories. By categorizing content cards, the user can easily retrieve a set of cards.

The relationship between a parent and a child is represented by an arc. A parent card is supported by a rod to attract attention (Figure 5).

SKG allows plural trees on a virtual globe because the arrangement is limited if it allows only one tree. An independent card that doesn't belong to a tree is also allowed by the same reason.

The region (circle) of a tree is displayed when the user modifies a tree structure. The user can easily connect a child card with a parent card by dragging and dropping a child on the parent card or its region. The disconnecting operation is the inverse.

The tree manipulations such as connecting, disconnecting, moving, deleting, and so on operate recursively.

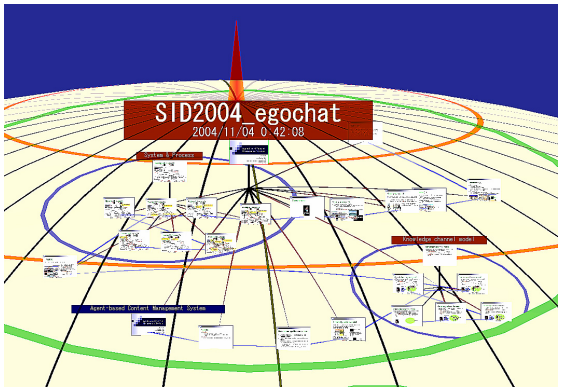


Figure 5: Tree structure

Moreover, the user can scale down a lot size of a tree to store huge content cards on one global sphere. By scaling down, the distance between a parent and a child shortens and the size of a card gets smaller.

3.3.2 Story structure

A sequential content is important for a user to follow a story. In SKG, a story structure is a directional list that shows a story line between content cards. The relationship between a previous card and a next card is represented by a directional arc that can be across trees. The story is so orderly and cross-boundary that the user can easily have a conversation about cross-contextual content.

Figure 6 is an example of the story. It shows a story from the card (1) (named “Process 1-1”) to the card (6). The user can easily go back and forth in the story by clicking arrow buttons.

The user can modify a story structure by using right click menu. He can also import content as a story by sorting in last modified date.

The colours of arcs are different between a tree and a story. Drawing arcs of trees and stories can be switched ON and OFF not to confuse a user.

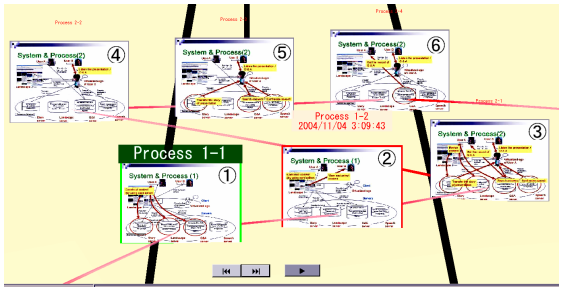


Figure 6: Story structure

3.4 Zooming interface

A zooming graphical interface (Bederson and Hollan, 1994) is effective to manage huge content on a finite screen size. It enables a user to look over the whole content by changing a scale, or to observe specific content by changing focus. Zooming interfaces can be classified into liner and non-linear zooming in general. The linear zooming (Bederson and Hollan, 1994) magnifies and shrinks whole the information like a multi-scale map. The non-linear zooming (Furnas, 1986; Lamping, Rao, and Pirolli, 1995) distorts an arrangement of information to focus specific information. SKG uses the linear zooming interface to manage content because it aims to make a good use of geographical information that should not be distorted. Figure 7 shows a zoom out view where the panorama of the content cards can be seen.

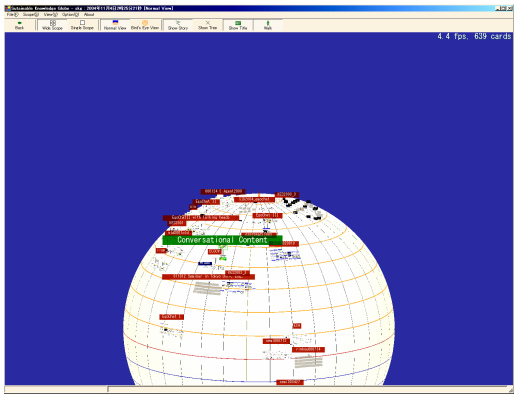


Figure 7: Zoom out view

In SKG, the linear zooming is implemented by using fractal approach (Koike and Yoshihara, 1993) that is an algorithm to zoom a tree structure. It keeps both users' cognitive load and system's response time to be constant by keeping number of visible tree nodes to be nearly constant at any scale. The original algorithm assumes that there is one tree in the target space, whereas there are many trees and independent nodes in the user’s field of vision (namely ‘3D view volume’) in SKG. To solve this problem, we assume one virtual root node that is the parent of all trees and nodes in the field of vision.

The degree of details of the displayed content card is calculated by the fractal approach. The content card has five levels of details: a high detailed thumbnail, a middle detailed thumbnail, a low detailed thumbnail, non thumbnail, and a filled circle. The more cards are displayed, the less their details appear. The ancestor cards are displayed with more detail, while the descendant cards are displayed with less detail. The child cards under the minimum threshold are displayed as only a filled circle that shows the region of a tree.

3.5 Contextual relation

To support a conversation in a rich context, SKG provides a method for managing contexts of the content card. Every content card has contexts that represent the background information of the card. The user can record the contexts and jump to one of them in anytime. A context consists of a latitude value, a longitude value, and a zooming value that decide the user's field of vision. The default context is the tree view where all descendant cards are in the field of vision. If the card has no child, the default context is the closest view where the card fills the screen (Figure 4). When the user clicks a content card, a virtual sphere rotates and zooms automatically to focus the card and show the default context of the card. The default context can be changed by the user.

Figure 8 and Figure 9 are examples of the context. In both figures, the same card (named "Concept") is focused, but contexts are different. Figure 8 is a tree view that emphasizes the children of the card, while Figure 9 is a wider view that emphasizes the story of the card by showing directional arcs back and forth.

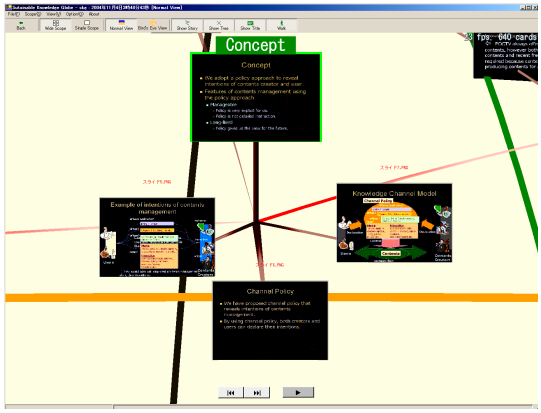


Figure 8: Example of the context (tree view)

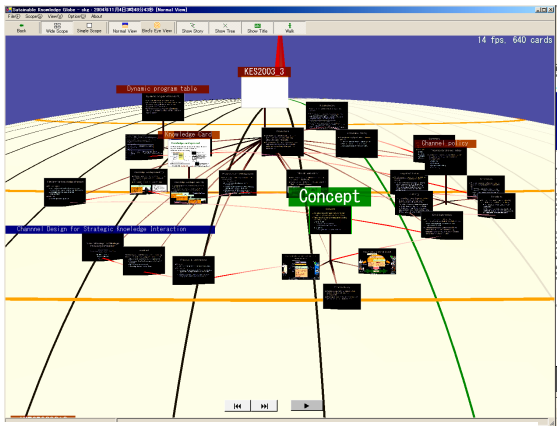


Figure 9: Example of the context (wide view)

3.6 Exploring the landscape

Just after starting SKG, the user defined home position is shown, and then a user can explore the landscape by using a wheel mouse. The sphere has three degrees of freedom (latitudinal direction, longitudinal direction, and Z-axis direction) because a standard mouse device seems to be not suitable to operate a sphere with many degrees of freedom.

A user also changes the view angles of a sphere. A normal view is an oblique view that has a depth, while a bird's eye view is an overhead view that has less occlusion problems.

4 Experiments

We have evaluated effectiveness of SKG through two experiments in practical conversational situations: the one is in a group situation and the other is in a personal situation.

We have recorded the proceedings of a meeting of ESL[1] working group by using SKG. The participants discussed a study of an intelligent room. The virtual landscape of SKG was projected on a large screen during the meetings. One operator imported handouts, wrote the speeches down, and located on the surface of the globe. The speeches in the same topic were grouped into one or two cards, and the cards are almost grouped into a tree by date. The operator also manipulated the globe according to participants' demands that they wanted to focus the specific content.

The meetings are held 10 times from August to November in 2004. The total time length of the meetings is 20 hours. The average number of participants is six. As a result of the experiment, we have acquired 151 content that include 12 trees and 3 stories. We had an interview with the participants about the effectiveness of SKG and got positive comments as bellows:

- I can overview the whole information in the meetings.
- The landscape is useful for me to look again the past proceedings.

We also got a comment that it is difficult for the speaker to operate SKG by a mouse device. The user should not be engaged with mouse operations in conversational situations because it disturbs communication using natural gestures. We are now developing a novel immersive browser that can improve the operativity of SKG by using physical in-

[1] http://esl.sfc.keio.ac.jp/esl_e.html

terfaces like a motion capturing system in a surrounding information space.

We have also experimented on SKG in personal situations. Three subjects have constructed their own landscapes to manage their conversation contents. The contents are mainly research slides and movies, and the rest of them are leisure photos, bookmarks and memos. The average number of content cards is 4,000. They have used their landscape instead of using other presentation systems to talk about their research and some other topics in several workshops and meetings.

The presentations using SKG have suggested us that SKG makes it easier to refer to old other presentations because the recent content and huge past contents are on the same landscape. The subjects often showed a panorama of the topics and sometimes went out of his way to the neighbour cards. Here, SKG seems to support an unexpected talk and a productive digression.

Various narrative structures are formed on their landscapes. Figure 10 shows the examples of how they arranged their cards. Arrangement (a) is a square arrangement that was thought out by Subject I, who said that the square is so regular that he can easily explore his catalogue of CDs and shoes. Arrangement (b) and (c) were thought out by Subject II. Arrangement (b) is a star arrangement where a parent card is centered and its children spread on all sides. Subject II said that this is suitable for representing parent-child relationships of the content. Arrangement (c) is a distorted map where photos are located on a cognitively distorted map of subject II. Photos in England and Italy are located near the Japanese cities because these are all of his journeys. The arrangement of the story is also characteristic. Figure 11 shows the examples of the story arrangement. The story in arrangement (A) by subject II flows horizontally from back left to front right, while the story in arrangement (B) by subject III flows vertically from left front to back right. In the both arrangements, the card queues are turned at the ends of the semantic section 1 and 2. When a user watches the story from the front, (A) emphasizes the flow of the story from left to right, while (B) emphasizes the beginning of the semantic section of the story. Arrangement (C) by subject III and (D) by subject I are smooth arrangements without turn. (C) is a compact spiral that saves space. (D) is a clockwise arrangement that shows the flow of time by using clock metaphor.

It seems that SKG enables subjects to externalize narrative structures in their minds voluntarily. They are subjective rather than objective; they reflect a kind of cognitive map rather than ontological map. Arranging the content landscape on SKG likes an arrangement of the room that brings us customized environment for managing huge objects.

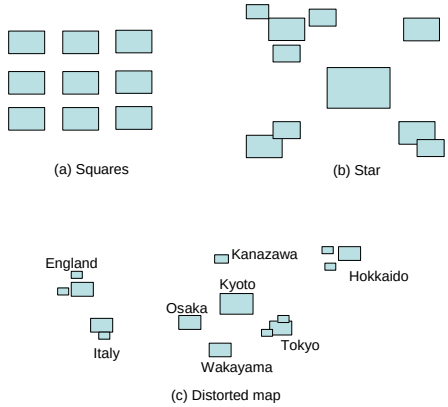


Figure 10: Examples of the card arrangement

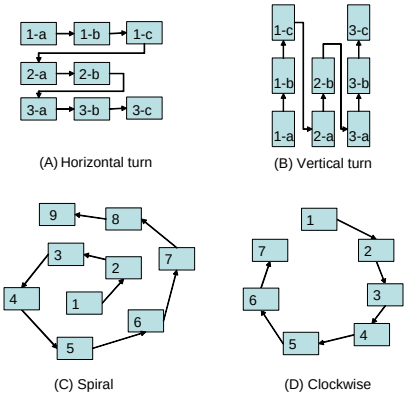


Figure 11: Examples of the story arrangement

5 Discussion

We are now studying automatic content creation from conversations to reduce a load of manually creating content cards (Kubota, Takahashi, Saitoh, Kawaguchi, Nomura, Sumi and Nishida, 2005). Our essential idea is conversation quantization that is a technique of approximating a continuous flow of conversation by a series of objects called conversation quanta each of which represents a point of the discourse. There is a good perspective that conversation quanta can be extracted from the real world conversation by expanding the ubiquitous sensor room system (Sumi, Mase, Mueller, Iwasawa, Ito, Takahashi, Kumagai and Y. Otaka, 2004), where conversational activities can be captured by environment sensors (such as video cameras, trackers and microphones ubiquitously set up around the room), wearable sensors (such as video cameras, trackers, microphones, and physiological sensors) and LED tags.

Our another perspective is integrating SKG with IPOC (Immersive Public Opinion Channel) (Na-

kano, Okamoto and Nishida, 2004; Nakano, Murayama and Nishida, 2004). So far we have developed the global view of content, but the immersive view is insufficient. The immersive view is expected to entrain participants to the subject by embodied fashion and improve operations in landscape by using physical interface. IPOC allows for expanding content landscape in a photo realistic immersive environment. Users can interact with conversational agents in a story-space, which is a panoramic picture background and stories are embedded in the background.

6 Conclusion

In this paper, we proposed Sustainable Knowledge Globe that is a system for supporting content-oriented conversation by using geographical arrangement, topological connection, contextual relation, and a zooming interface. We evaluated effectiveness of SKG through two experiments in practical conversational situations. As a result, we got good suggestions that users can easily access his old content and construct their content on SKG according to a narrative structure in their minds.

References

Joseph M. Ballay. Designing Workscape: an interdisciplinary experience. *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, 10-15, 1994.

B.B. Bederson and J.D. Hollan. Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics. *Proceedings of ACM UIST '94*, 1994.

S.K. Card, G.G. Robertson, and W. York. The WebBook and the Web Forager: an information workspace for the World-Wide Web. *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, 111-117, 1996.

G.W. Furnas. Generalized fisheye views. *Human Factors in Computing Systems CHI '86 Conference Proceedings*, 16-23, 1986.

J. Hur and T. Nishida. Knowledge Editing Support using Mutual Adaptive Knowledge Externalization Method. *In Proceedings of the 18th Annual Conference of the Japanese Society for Artificial Intelligence*, 2E1-05, 2004.

H. Koike and H. Yoshihara. Fractal Approaches for Visualizing Huge Hierarchies. *Proceedings of the 1993 IEEE Symposium on Visual Languages*, 55-60, 1993.

H. Kubota, S. Kurohashi and T. Nishida. Virtualized-egos using Knowledge Cards. *Seventh Pacific Rim International Conference on Artificial Intelligence (PRICAI-02) WS-5 International Workshop on Intelligent Media Technology for Communicative Reality (IMTCR2002)*, 51-54, 2002.

H. Kubota, M. Takahashi, K. Saitoh, Y. Kawaguchi, S. Nomura, Y. Sumi and T. Nishida. Conversation Quantization for Informal Information Circulation in a Community. Social Intelligence Design 2005 (SID2005), March 24-26, Stanford, CA, USA, 2005. (to appear)

J. Lamping, R. Rao, and P. Pirolli. A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, 1995.

Y. Nakano, M. Okamoto and T. Nishida. Enriching agent animation with Gestures and Highlighting Effects, in Proceedings of International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTCI 2004), 2004.

Y. I. Nakano, T. Murayama and T. Nishida, Engagement in Situated Communication By Conversational Agents, in Proceedings of International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTCI 2004), 95-101, 2004.

G. Robertson, J.D. Mackinlay, S.K. Card. Cone Trees: Animated 3D Visualizations of Hierarchical Information. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'91)*, 189-194, 1991.

G. Robertson, M. Czerwinski, K. Larson, D.C. Robbins, D. Thiel, and M.V. Dantzich. Data mountain: using spatial memory for document management. *Proceedings of the 11th annual ACM symposium on User interface software and technology(UIST '98)*, 153-162, 1998.

Y. Sumi, K. Mase, C. Mueller, S. Iwasawa, S. Ito, M. Takahashi, K. Kumagai and Y. Otaka, Collage of Video and Sound for Raising the Awareness of Situated Conversations, in Proceedings of International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTCI 2004), 167-172, 2004.

Supporting the Creation of Immersive CG Contents with Enhanced User Involvement

Masashi OKAMOTO[*]
Kazunori OKAMOTO[*]
[*]Graduate School of Information Science and Technology, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
okamoto@kc.t.u-tokyo.ac.jp, kazu@kc.t.u-tokyo.ac.jp

Yukiko I. NAKANO[†]
[†]Research Institute of Science and Technology for Society, Japan Science and Technology Agency
Atago Green Hills MORI Tower 18F, 2-5-1 Atago, Minato-ku, Tokyo, 105-6218, Japan
nakano@kc.t.u-tokyo.ac.jp

Toyoaki NISHIDA[**]
[**]Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract

As computer technologies have been advanced recently, people tend to feel stressed against poor or unsatisfactory interaction with computers. In order to solve such a situation we propose the human-computer interaction design theory of 'User involvement', which is expected to give guidelines for constructing a natural interaction environment for humans and computers. Specifically, we show the effectiveness of our theory in building a support system for immersive CG contents.

1 Introduction

Nowadays as computer technologies have been advanced rapidly, we often encounter the scene where people feel stressed against poor or unsatisfactory interaction with computers. For instance, in operating a word processor or in communicating with interactive QA system, the users are frequently annoyed with a cumbersome agent or irrelevant responses from the system. On the other hand, interesting TV programs and computer games do not lose their popularity from their watchers and players. What makes the difference?

Our answer is that it is because many of computer-mediated interactive contents leave the cognitive involvement of their users out of consideration that such problems occur. We do not want to insist that the users should participate in designing computer programs. Instead, the designers and the engineers should make out the cognitive activity of the users who use computers feeling *reality* in the virtual world before their eyes or in human-to-computer interaction in which they are engaged. On these occasions the users are deeply into another world or willingly keep communicating with virtual agents or robots. In other words, they are *involved* in the human-to-computer interaction environment.

We thus call such a cognitive activity or state of computer users 'User involvement' (Okamoto et al., 2004). Considering the user involvement helps to make an affective and favourable design of human-to-computer interaction environment.

In this paper, we first propose the idea of user involvement in brief. Specifically, we place much emphasis on the relations of reality and empathy based on cognitive linguistics research. And then a cognitive model for movie contents is shown based on observations on TV programs in Section 2. Lastly, as an example of the systems with enhanced user involvement, our ongoing work on a supporting system for the immersive CG contents creation is described. It is built to support a user who wants to create CG contents with little effort.

2 User Involvement in Human-Computer Interaction

The connotations of 'human-computer interaction' range from an explicit communication with a conversational agent or an intelligent robot to implicit interactions in playing computer games or watching CG contents. When a user consciously communicates with a robot, it can be said that he is *engaged* in that communication in the sense of Sidner et al. (2003). Thus the *engagement* might be applied to explicit human-computer communication.

In this paper we use '*(User) involvement*' instead so that it should include the situation where a user is involuntarily involved in a human-computer interaction or a virtual world. Our research focuses on finding out how the user involvement can be established and enhanced through a good design of human-computer interaction environment.

2.1 Requirements of user involvement

Our definition of the 'User involvement' and the main requirements to establish it are as follows:

User involvement. The cognitive way humans willingly engage in, or are forced to be involved in a virtual world which computers display, in a human-to-robot communication, or in a computer-mediated community.

Requirements.

1. Cognitive/Communicative/Social reality should be achieved.
2. Two (or more) cognitive spaces should be linked, and the user should cognitively move in and out those spaces.

In this approach the 'reality' is classified into the following three dimensions:

- **Cognitive reality.** The way of seeing objects, events and their relations in the real/virtual world as real.
- **Communicative reality.** The sense of reality that is achieved through communication with others.
- **Social reality.** The collective and intersubjective sense of reality based on sharing thoughts or opinions with one another.

In this research we mainly focus on establishing cognitive reality and communicative reality because social reality is considered to concern computer-mediated communities or online communities on the Internet.

2.2 Astigmatic model of user involvement

Since the user involvement is strongly related to our sense of reality and is common to both verbal and nonverbal communications, linguistics researches help to comprehend how it works. In particular recent cognitive linguistics suggest many important characteristics of human cognition in conceptualizing the world.

Langacker (1993) points out the reference-point ability, which enables us to conceptualize an entity at a distance using a mental path from a more accessible entity as a reference point. Applying this concept to the user involvement, it can be said that people conceptualize an unfamiliar entity in another world utilizing a more accessible one as a reference point that stands on both our world and the other.

At the time both of the worlds (i.e. cognitive spaces) need to be linked or overlapped with the reference point. That is not limited to the relation between a real space and a virtual space. At the very start of our life, we are living in two spaces, that is, thinking in the inner space and acting in the outer world using our body as a reference point.

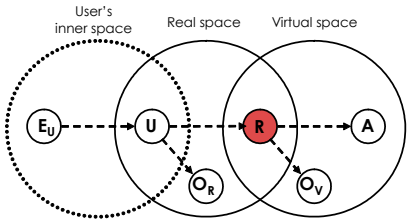


Figure 1: Astigmatic model

Figure 1 shows an example of how the computer user conceptualizes each object in different cognitive spaces using each reference point, especially in a virtual agent system. We call this model as the astigmatic model of user involvement, because multi-spaces are overlapped and linked there.

A computer user (U) accesses an object in a virtual space (O_v) via some reference point (R) that is easily accessible for the user and, at the same time, is a constituent of the virtual space. For instance the correspondent movements of a mouse and its pointer function as a reference point in that it connects our real world and a virtual world in the computer monitor. Then the user feels the interactions with the computer as real. In other words, the *cognitive reality* for a virtual world is established.

Similarly, in our everyday lives, E_U (ego of user) conceptualizes O_R (object in real space) using U (i.e. his self/body) as a reference point. In fact we are not living in a single world but in two worlds at least: our inner world (i.e. our mind) and the outer world

(i.e. the real world). That is the way our cognitive reality for the real world is established.

2.3 Empathy Channel

What becomes a reference point that links our real world and a virtual world on a computer display? We insist that one of the dominant candidates is an empathized ego in the virtual space. As seen in many computer games, a variety of virtual egos of the users exist in monitor, such as a fighter in fighting games and a user's car in driving simulation games. The users can play and step into the game through empathizing with such egos.

On the other hand, we often empathize with protagonists or persons in novels or movies, and virtually experience their lives. Such empathy is achieved by common characteristics between a reader and a character in the story: gender, ethnicity, emotions and experiences in similar situations.

As the empathy connects two cognitive spaces in these occasions and enables the user (or reader) to step into another world, we call it 'Empathy Channel'. The user can acquire and utilize another cognitive viewpoint through the Empathy Channel. Then the user can interact or communicate in the virtual world with a sense of reality. Part of our main goal is to design a good Empathy Channel in human-computer communication environment.

In next section we propose the several methods of enhancing user involvement for movie contents, and analyze TV programs to show how the effective transition of camera shots in them are based on the user involvement to reduce cognitive burden for their audience.

3 Enhancing User Involvement in Movie Contents

In order to enhance the user involvement, human-computer interaction environment should be designed to make the users step into the environment with little effort. Specifically, there should be two kinds of artful devices, that is, establishing a channel to the environment and reducing user's cognitive burden. Thus, it is necessary for making attractive movie-like contents to establish an Empathy Channel in the virtual world and realize smooth shot transition.

In this section we make the following two design suggestions: (1) virtual settings using Empathy Channels for cognitive reality (2) shot transition based on Cognitive Overlapping and Empathy Channel for communicative reality. Furthermore, we analyze popular TV programs and verify the effects of our suggestions.

3.1 Virtual settings with Empathy Channel

As for a conversational agent system, the static aspect is how the virtual world with agents should be arranged. We suggest that the virtual settings using Empathy Channel is one of the effective arrangements.

Nowadays many conversational agent systems are seen as desktop character or web application. Microsoft Agent[1] and Ananova[2] are its typical examples. However, since they work as user's partner, the user will not regard them as his alter ego. As a result, those agents frequently convey the impression that they are just annoyances or unnoticed strangers. There are many reasons for the consequence, but the most important one is that those agents were designed to communicate solely with the users. As observed in human communication, one has to communicate with others via verbal and nonverbal channels. But, natural verbal communication is difficult for conversational agents even if they are 'conversational' because a highly intelligent system is required. On the other hand, it is well-known that one of the nonverbal devices to be established for natural communication is eye contact. However, it is also known that a current agent in a computer display cannot achieve eye contact with its user facing the display. Eventually, the computer users tend to feel those agents as "fake" partners to communicate with.

We thus make a suggestion that a conversational agent should be an empathized ego instead of his partner in two ways. One is to use two or more agents that communicate with each other in the virtual world and to make one agent function as empathized ego of the user. Then the user will empathize with the agent, and will enter the virtual world through it (i.e. Empathy Channel).

The other idea is to use the back image of agent as Empathy Channel (Okamoto et al., 2004). Miyazaki (1993) experimented the empathetic effects of a back image for reading a story. According to the experiment, the picture book featuring the back images of its protagonist make its readers involved in the story more than in the same story with pictures drawn from the observer view. In brief, the back images helped the readers to experience the virtual world as if it were their own.

To sum up, the virtual settings in conversational agent system should be arranged to set Empathy Channels for cognitive reality by applying agent-to-agent communication and back images of the empathized agent.

[1] http://www.microsoft.com/msagent/

[2] http://www.ananova.com/video/

3.2 Shot transition with Cognitive Overlapping

In movie contents there are two aspects to be considered for enhancing the user involvement, that is, a static aspect and a dynamic one. Regarding the dynamic aspect of enhancing the user involvement, when making movie-like CG contents that has storytelling progression, smooth shot transition should be realized so as to reduce cognitive burden for the audience.

As many linguistic researches suggest, our discourse and storytelling are based on the consistent flow of new and old information. For example, the following story clearly indicates such information structure:

> *Once upon a time there was a king that wanted a new castle.* ***The king*** *hired the best castle builder in the land to build the castle. Prior to starting* ***to build the castle****, the king got an architect to construct the plans for the castle and* ***the plans*** *were drawn up quickly*[3]*…*

In cognitive linguistics the information structure is considered to be motivated by our cognitive ability of *figure-ground* perception. In the above example, the underlined parts are *figure* while the bold parts become *ground* (Cf. Talmy, 1978; Langacker, 1987).

Nevertheless, not all discourses and stories follow such explicit information flow, but it is certain that the figure-ground alteration would reduce the cognitive burden in reading or listening to a story. We believe that the same structure can be applied to movie contents[4]. See Figure 2 below.

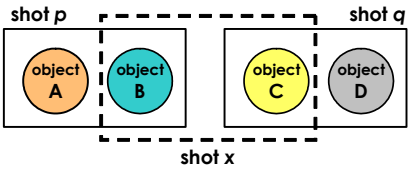


Figure 2: Cognitive Overlapping

In this figure a square corresponds to a camera frame. When shot *p* featuring two objects is just followed by shot *q* featuring other different two objects, the audience is forced to relate the adjacent two shots in his mind, which will be a high cognitive task for him because one movie consists of so many shots that keep changing continuously. In this case, if shot *x*, containing an object already appeared in shot *p*, is inserted between shot *p* and *q*, then the cognitive burden for the audience will be reduced. It is because the object commonly captured in shot *p* and *x* changes its cognitive status from *figure* to *ground* through the shot transition and becomes a reference point for the audience to conceptualize the following shot.

It is assumed that such overlapping is not limited to the visual images of objects. Since movie contents consist of images and sounds, auditory overlapping will also work in enhancing user involvement. For instances, movies or TV dramas sometimes contains the scenes where the shots are changing continuously but narration or actor's voice keeps unchanged during the shot transition.

The overlapping based on figure-ground alteration helps to achieve cognitive reality, so we call it 'Cognitive Overlapping'. The types of shot transition with Cognitive Overlapping are illustrated in Figure 3.

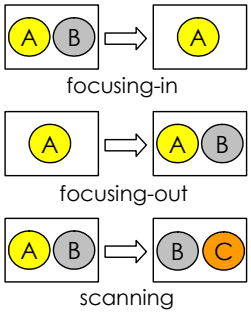


Figure 3: Shot transition with Cognitive Overlapping

Although there are a variety of camerawork techniques such as zoom, tilt, pan, and reverse angle, we classified the shot transition with Cognitive Overlapping into these three types of camerawork. Our classification reflects the semantic relations of shot transition based on captured objects in each shot.

3.3 Shot transition with Empathy Channel

However, we have to admit that non-overlapping transition exists especially from a person shot to an object shot. It seems to contravene our hypothesis, but that is not so.

In such a shot transition type, a person in the previous shot occasionally gives a pointing gesture, gaze, or verbal reference toward the object which lies out of the frame but is recognizable or accessible for him. His attention behaviour leads the audience to make mental contact to the hidden object through the attention as a reference point. As a re-

---

[3] http://www.umsl.edu/~sauter/analysis/fables/fall2002/king_castle.html

[4] Many researchers in film studies have pointed out that language and movie have a lot in common. See Monaco (1977).

sult, the audience gets ready to accept the next shot featuring the object alone (see Figure 4).

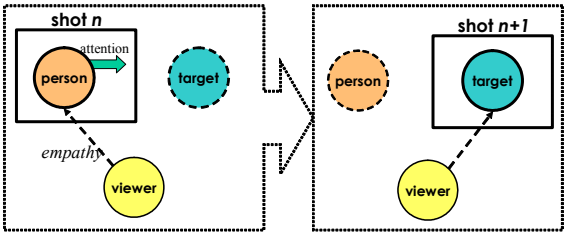


Figure 4: Shot transition with Empathy Channel

In other words, through pointing gesture, gaze or verbal reference by a person in the camera frame, the audience can empathize with him and can effortlessly relate the following shot to the previous one. Such a communicative cue for attention functions as the Empathy Channel for communicative reality.

For establishing a natural human-communication environment, Empathy Channels for both cognitive and communicative reality need to be considered. We thus suggested the virtual settings using Empathy Channels for cognitive reality and the shot transition based on Cognitive Overlapping and Empathy Channel for communicative reality, but our ideas still remain speculative.

3.4 Analysis of the shot transition in TV program

In order to verify the speculation for shot transition, we observed and analyzed certain popular TV programs. This section shows the analyzed data, the analysis, the result and some discussions.

3.4.1 Data

The TV program we analyzed is a 30-minute one providing the information about popular items in a certain shop through the conversation between a TV host as *guide* and a shop owner or clerk as *explainer*. The whole data we used is three programs of 90 minutes.

We divided all the shots in the programs into the following seven types of shots according to what was captured in each shot in view of Cognitive Overlapping (see also Figure 5):

Type 1: The shot featuring the guide
Type 2: The shot featuring the explainer
Type 3: The shot featuring objects to be explained
Type 4: The shot featuring the guide and the explainer
Type 5: The shot featuring the guide and the objects
Type 6: The shot featuring the explainer and the objects
Type 7: The shot featuring the guide, the explainer and the objects

Type 1 Type 3 Type 4

Figure 5: Some examples of shot type

3.4.2 Analysis and results

These programs consist of 485 shots, 78 shots of which is counted in Type 1, 56 shots in Type 2, 117 shots in Type 3, 57 shots in Type 4, 56 shots in Type 5, 54 shots in Type 6, and the other 67 shots in Type 7. As the programs we analyzed were for the information about shops and their items, the most frequent shot type was Type 3, which features an object to be introduced to the viewing audience. Each of the transition rates from one shot to another is shown in Table 1 (Note: the transition between the same shot types is excluded).

Table 1: The shot transition rate (%)

From \ To	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
Type 1		14.1 (12.8)	17.9 (12.8)	**30.8**	**19.2**	9.0 (3.8)	**9.0**
Type 2	17.9 (16.1)		16.1 (8.9)	**30.4**	1.8 (1.8)	**16.1**	**17.9**
Type 3	17.1	7.7		10.3	**24.8**	**22.2**	**17.9**
Type 4	**26.3**	**40.4**	14.0 (8.8)		**1.8**	**7.0**	**10.5**
Type 5	**33.9**	7.1 (7.1)	**42.9**	**0.0**		**0.0**	**16.1**
Type 6	3.7 (1.9)	**16.7**	**55.6**	**1.9**	**1.9**		**20.4**
Type 7	**6.0**	**4.5**	**58.2**	**6.0**	**13.4**	**11.9**	

In this table, a bold figure means the rate of overlapping transition. The rest represents non-overlapping shot transition. Moreover, each figure in a bracket means the non-overlapping transition occurred with pointing gesture or gaze toward the target in the following shot.

3.4.3 Discussion

This result shows that overlapping shot transition is frequently used in TV programs since the transition occupies 77.9% of the whole transition. It also suggests that it is effective for establishing cognitive reality to use Cognitive Overlapping. Specifically,

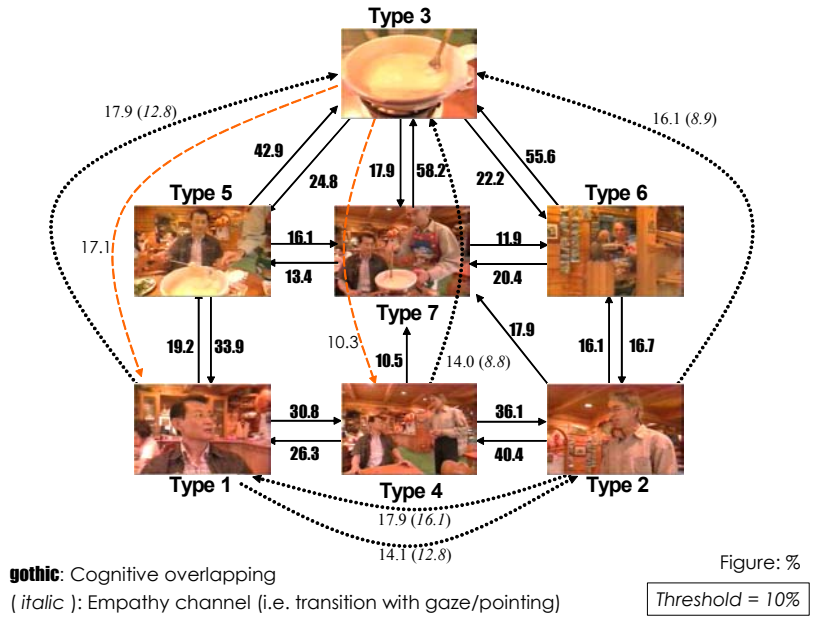


Figure 6: The shot transition model

as illustrated in Figure 3, the *focusing-out* shot transition is used to introduce new information to the audience, while the *focusing-in* to focus the object or the person to pay attention to for the audience.

Although the overlapping transition occupies most part of the whole transition, the rest should be explained in view of the user involvement. We found that 61.6% of the non-overlapping transition is the transition from person shots to object/person shots, that is, the shot transition in which Empathy Channel can be established.

Actually, 72.7% of those transitions include attention behaviours, such as pointing gesture or gaze by the person toward the object(s) in the next shot. Therefore, it can be said that communicative reality via Empathy Channel also enhanced the user involvement. Furthermore, the total rate of the shot transition using either Cognitive Overlapping or Empathy Channel covers as much as 87.8% of the whole.

We summarize the result as the shot transition model in Figure 6. The threshold is fixed at 10%, those transition rates under which are not shown in this model. As seen in Figure 6, there remain a few transitions that would be possible and expected from our theory, but are practically in very low rate or never happen. We assume it might depend on the content or the information flow of the data we used, but more detailed examination is needed.

In next section, we describe our ongoing work for supporting immersive CG contents creation based on the observations here and the virtual settings with Empathy Channel described in Section 3.1.

4 Supporting immersive CG contents creation

In creating CG contents, only skillful artists and creators have so far been able to produce affective and immersive contents with high user involvement. Based on the user involvement theory and the analysis of TV program described in the previous sections, this section proposes a system that enables the user to create immersive CG contents with little effort.

4.1 Previous methods

Virtual Director (Manos et al., 2000) and TVML (Hayashi et al., 1997) are some examples that focus on creating CG contents based on scripting languages. In this approach, all particular events, characters' gestures and also camerawork need to be described in the scripting language beforehand. Thus, creating CG contents using this approach requires sophisticated skill and is quite difficult for amateurs.

In addition, the systems described in Ariyasu et al. (1999) and Douke et al. (2000) try to help the user create CG contents by automatically generating agent arrangement inside a CG set and camerawork. The cost of creating CG contents in such systems can be reduced by using templates, which are derived from the observation of those TV programs that provide information to audience. According to these works, the templates play as tacit rules to enhance the comprehension that the audiences have for each TV program.

Instead of defining templates, in this study, we use the shot transition network proposed in Section 3 as rules for selecting camerawork. This is because the shot transition network can generate more various patterns of camerawork derived from a real TV program than template-based approach. Moreover, our method has an advantage over the previous ones in generating character agents playing on the background photos taken by non-professional users.

4.2 Virtual environment setting

The system we propose is designed for novice users to easily produce movie-like CG contents with enhanced user involvement. Thus, the virtual world setting is simplified and optimized for a specific task, that is, to introduce and explain about interesting objects or monuments to audience by conversation between agents.

The virtual setting of the system is supposed to construct an agent-to-agent communication environment based on Empathy Channel for cognitive reality as illustrated in Figure 7.

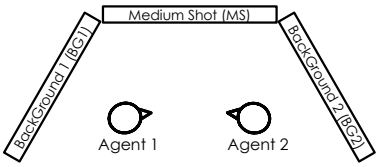


Figure 7: Virtual environment setting

The basic photos that the system uses are one *Medium Shot*, which features an object to be explained, and two *BackGround* photos, each of which becomes a background picture for an agent who stands in front of it.

Additionally, two conversational agents communicate with each other surrounded by the three pictures. One agent is the guide agent, who is expected to function as empathized ego of the user, and the other is the explainer.

For a user to use this system, he needs to prepare at least three photos beforehand, which should satisfy the virtual settings above. Specifically, the photo preparation templates (Figure 8) are available in order to adjust a photo to each template. For instance, a user has to shoot or select a picture which is suitable as background for the placement of two agents to be natural using BG1 template. Similarly, when selecting a photo for MS, the user need to adjust an explanatory object into the dashed square in MS template.

Figure 8: Photo preparation templates

In order to enhance the user involvement, the back image of the guide agent is also used as shown in BG1 template, which generates an over-the-shoulder shot. Then the back image of the guide agent functions as Empathy Channel.

4.3 System architecture

The outline of the system is shown in Figure 9. The system consists of four main modules: the Content Editor, the Shot Generation Module, the Gesture Generation Module, and the Camerawork Generation Module. Using the Content Editor, users select a scene type, three basic photos for BG1, BG2 and MS for the explanatory scene, and speech for animated agents.

Scene type and three photos are sent to the Shot Generation Module, where the photos, the agents and camerawork are properly arranged in the virtual environment. The utterances to be spoken by the agents are sent to the Gesture Generation Module, where agent gestures are selected and scheduled according to the utterances and the scene setting in the virtual environment. Then, in the Camerawork Generation Module, camerawork is specified for each shot based on the shot transition network constructed from the shot transition model in Section 3. The final output of the system is a set of action command executable by the Haptek player[5]. The details of each process are described in the following subsections.

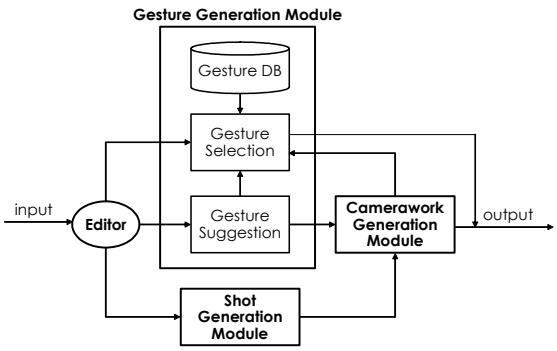


Figure 9: System architecture

[5] http://www.haptek.com/

4.4 Content editor

The users as contents creator specify utterances and photos using the Contents Editor shown in Figure 10.

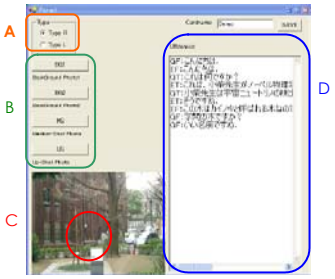


Figure 10: Contents editor interface

The photos and agent utterances are stored as materials constructing a scene. Content editing consists of the following four steps.

First, scene information is specified in (A) in Figure 10. Scene information consists of scene type and placement type. The scene type is either a walk scene or an explanation scene, though our system deals only an explanation scene for the moment.

On the other hand, placement type is either Type R or Type L according to the viewpoint of the guide agent. When the object to be explained (i.e. focused object) is on the right side of the guide agent, the placement type is Type R. When the focused object is supposed to be on the left side of the guide agent, the placement type is Type L.

In picture selection (B), the user selects pictures for the scene. A few photos are selected for a walk scene, and three photos are selected for an explanation scene. BackGround photos (BG1, BG2) are the pictures in which a focused object is not shown. In a medium shot photo (MS), a focused object is shown in a medium distance. There need another photo as UpShot photo (US), which is a zoomed-up picture of a focused object. Therefore, the user can prepare the US separately or might trim the US from the MS, though the US will be of low resolution in that case.

When the system receives the materials, the photos are sent to the Shot Generation Module where all the possible shots are produced from the pictures with character agents. Then the user needs to mark a focused area on the selected MS, which is shown in a preview window (C). Note that a focused object for an explanation scene should not be placed near the edges of the MS photo as described in 4.1.

Utterance information is specified in utterance window (D). First, utterances to be spoken by conversational agents need to be typed in the window. In editing an explanation scene, a speaker tag, which specifies who is the speaker of the utterance, needs to be added to at the beginning of the utterance. A speaker tag "G" is added to the utterances of the guide agent. "E" is added to those of the explainer agent. In addition, "T" tag is added to an utterance which refers to the focused object in it. If the utterance does not refer to the focused object, "F" tag is assigned.

4.5 Shot generation module

In the Shot Generation Module, all the types of shot types defined in Section 3 are generated using setting information in a virtual world and three pictures chosen in the Editor.

Figure 11 shows a virtual environment setting of Type R for an explanation scene. The guide agent (G) and the explainer agent (E) are placed nearly face-to-face. An imaginary line joining the two agents is defined, which properly constrains possible camerawork, and both agents direct 15 degrees away from the imaginary line towards the medium shot. Shot type 1, 2, and 4-7 are produced by camera 1, 2, and 4-7 respectively. Shot type 3 is produced as a zoom-up shot by camera 7.

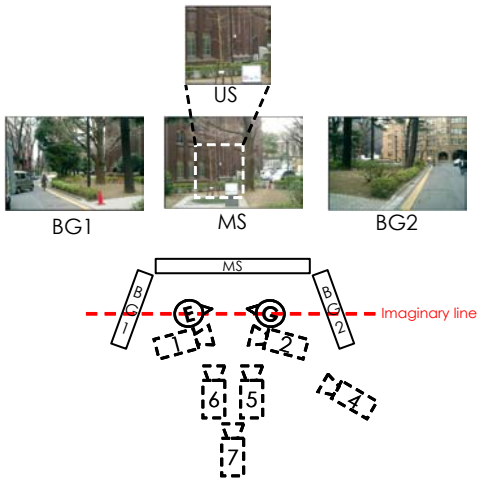


Figure 11: Camerawork generation

In addition, the photos to use for Type 5 and Type 6 are trimmed from the MS as shown in Figure 12.

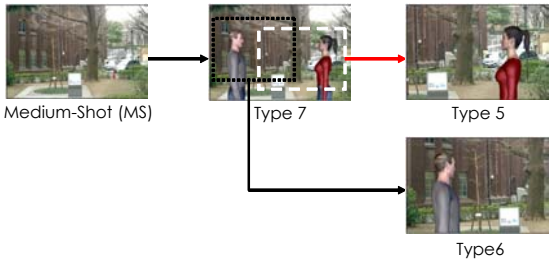


Figure 12: Shot generation from Medium Shot

4.6 Gesture generation

Gesture Generation mechanism consists of two consecutive processes: (1) gesture suggestion in which candidates of gestures are proposed, and (2) gesture selection in which appropriate gesture shapes (e.g., direction of pointing gestures) and gaze direction are determined according to the scene arrangement.

4.6.1 Gesture suggestion

We employ the CAST system (Nakano et al., 2004) as an agent behaviour suggestion mechanism, which outputs the suggestion to the Gesture Selection Module. CAST consists of four main modules: (1) the Agent Behaviour Selection Module (ABS), (2) the Language Tagging Module (LTM), (3) a Text-to-Speech engine (TTS), and (4) a character animation system. When CAST receives a text input, it sends the text to the ABS. The ABS selects appropriate gestures and facial expressions according to linguistic information calculated by the LTM, which uses functions of a Japanese syntactic parser (Kurohashi 1994). Then, the ABS obtains timing information by accessing the TTS, and it calculates a time schedule for the set of agent actions. The output from the ABS is a set of animation instructions that can be interpreted and executed by an animation system. In the proposed system, the timing calculation module is separated from the CAST and used after the gesture selection process.

4.6.2 Gesture Selection

The gesture commands are stored in Gesture Database and are called by the Gesture Selection Module. When the Gesture Selection Module receives suggestions from the Gesture Suggestion Module, it selects appropriate gesture shapes according to a scene setting in the virtual world. At the same time, the Gesture Selection Module receives gaze direction suggestions from Camerawork Generation Module based on the agent placement in the virtual environment setting.

4.7 Camerawork generation

The Camerawork Generation Module produces camerawork for a scene. To produce Cognitive Overlapping and Empathy Channel in CG contents, camerawork for an explanation scene is determined based on the shot transition network shown in Figure 13. Camerawork generation consists of the following three steps.

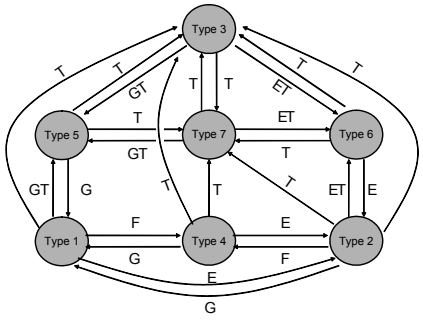


Figure 13: Shot transition network

(1) Determining shot candidates

Shot candidates for each utterance are selected according to who is the next speaker and whether the utterance refers to the focused object or not. Rules for determining shot candidates are shown in Table 2 below.

For example, in the Contents Editor, if the speaker tag "G" (guide agent) and reference tag "T" (referring to a focused object) are assigned to a given utterance, shot type 1, 3, 5, and 7 are selected as shot candidates.

In addition, Camerawork Generation module also manages the time schedule from the Gesture Suggestion Module. For instance, shot is not changed when the duration of an utterance is less than one second. This is because too much shot change makes the contents less comprehensible and increases cognitive burden of the audience. In such a case, a shot type which can be a common candidate for both of the two consecutive utterances is chosen. For example, when utterance A has tag "G" and "F" and utterance B has tag "E" and F", the shot type 4 is continuously used during these two utterances. On the contrary, when the utterance duration is longer than five seconds, a shot is changed not to get the user bored.

Table 2: Shot selection rules

Next speaker	Referring to focused object	Shot type
Guide	F	1, 4
Guide	T	1, 3, 5, 7
Explainer	F	2, 4
Explainer	T	1, 3, 6, 7

(2) Generating shot transition

From the shot type candidates chosen in step (1), this step generates an appropriate shot transition according to the shot transition network. As major transitions of the network produce Cognitive Overlapping, this step generates Cognitive Overlapping camerawork in most of the cases.

(3) Generating eye-gaze

When a transition without Cognitive Overlapping effect is selected, a gaze needs to be generated to produce the shot transition with Empathy Channel, the other device for implementing the User Involvement theory. In this case, a gaze is generated according to the shot transition network and a scene setting given by the user.

5 Conclusion and Future Work

We have so far described our idea on the 'User involvement' as the design theory for establishing natural human-computer interaction environment, and showed our ongoing work on constructing a support system for the immersive CG contents creation, which supports a user who wants to create CG contents with little effort. Since our system has not been fully constructed yet, it still remains unclear what will become obstacles to enhancing the user involvement. However, as long as the system design is based on the framework of the User involvement theory, our system will surely affect and entertain its user. We believe that the virtual settings with Empathy Channel and the shot transition with Cognitive Overlapping and Empathy Channel made the human-computer interaction environment more attractive for the audience of our system.

As our future work, we firstly try to keep constructing our supporting system and then will make a psychological experiment to prove the effect of enhanced user involvement. The experimental results will be shown soon.

Furthermore, since the User involvement theory still remains a design theory that gives a speculative sketch for natural human-computer interaction environment, it is desirable and expected to brush up the theory into that of evaluation.

References

K. Ariyasu, M. Hayashi, H. Sumiyoshi. Automatic Generation of TV Program & Program Relational Contents, *5th Symposium on Intelligent Information Media*, 171-176, 1999.

Mamoru Douke, Masaki Hayashi, Eiji Makino. Automatic Generation of Television News Shows from Given Program Information Using TVML, *Journal of The Institute of Image Information and Television Engineers*, No.7, 1097-1103, 2000.

M. Hayashi, H. Ueda, and T. Kurihara. TVML (TV program Making Language) - Automatic TV Program Generation from Text-based Script, *ACM Multimedia'97 State of the Art Demos*, 1997.
http://www.nhk.or.jp/strl/tvml/index.html

R. N. Kraft. The role of cutting in the evaluation and retention of film, *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Vol. 12, No. 1, 155-163, 1986.

Sadao Kurohashi and Makoto Nagao. A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. *Computational Linugistics*, 20 (4), 507-534, 1994.

Ronald W. Langacker. *Foundations of Cognitive Grammar, vol.1: Theoretical Prerequisites*. Stanford: Stanford University Press, 1987.

Ronald W. Langacker. Reference-Point Constructions. *Cognitive Linguistics* 4: 1-38, 1993.

K. Manos, T. Panayiotopoulos and G. Katsionis. Virtual Director: Visualization of Simple Scenarios, *2nd Hellenic Conference on Artificial Intelligence*, SETN, 2002.

Joseph V. Mascelli, *The five C's of Cinematography: Motion Picture Filming Technique*. Silman-James Press, 1998.

Kiyotaka Miyazaki. The Effects of Human Backimage as Mooring Point in Empathetic Comprehension through Visual Images *(in Japanese). Japan Educational Psychology Association Proceedings*: 35, 1993.

James Monaco. *How to Read a Film: The Art, Technology, Language, History, and Theory of Film and Media*. New York: Oxford University Press, 1978.

Y. Nakano, T. Murayama and T. Nishida. Multimodal Story-based Communication: Integrating a Movie and a Conversational Agent, *IEICE Transactions, Special Issue on Human Communication* (to appear), 2004.

Masashi Okamoto, Yukiko I. Nakano, and Toyoaki Nishida. Toward enhancing user involvement via Empathy Channel in human-computer interface design. In *Proceedings of IMTCI*, 2004.

C.L. Sidner, C. Lee, and N. Lesh. Engagement rules for human-robot collaborative interactions, *IEEE International Conference on Systems, Man & Cybernetics (CSMC),* Vol. 4, 3957-3962, 2003.

Leonard Talmy. Figure and Ground in Complex Sentences. In Joseph H. Greenberg, ed., *Universals of Human Language*, vol.4, *Syntax*, 625-629. Stanford: Stanford University Press, 1978.

Interactive Media for Gently Giving Instructions — Basic idea of watching and teaching users

Takuya KOSAKA*
*Department of Intelligent Interaction Technologies University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573, JAPAN
kosaka,ohta,kameda@image.esys.tsukuba.ac.jp

Yuichi NAKAMURA† Yuichi OHTA* Yoshinari Kameda*
†ACCMS, Kyoto University
Sakyo, Kyoto, 605-8501, JAPAN
yuichi@media.kyoto-u.ac.jp

Abstract

This paper introduces a novel multimedia system for instructing or guiding works. The system observes a user by image and speech recognition, and gives related information or appropriate advices by utilizing pre-recorded video archives. The distinctive feature of our media is that the system quietly observes a user and interrupts the user only when he/she really needs a help, for example, in a situation that the user is at a standstill or asks a question. Otherwise, the system only presents related information that may be useful to the user, and it does not require any responses from the user. In this paper, a method for recognizing a user's status and a method for matching it to the contents in video archives are mainly described.

1 Introduction

Teaching a work has various aspects and needs various ways for it. An ideal way is an experienced human instructor: he/she tells or demonstrates how to do something, gives an advice, answers a question, just carefully watches what a student does or wants to do, or interferes if a student is about to make an irrevocable mistake. On the contrary, when we consider teaching or guiding a work by a conventional multimedia system, the system cannot adjust its behaviors to a student's situation. We can think of, for example, a conventional system that teaches a way of cooking. The system may ask a user to do exactly the same things an instructor does or as shown in a recipe. In other words, it may interfere and/or order the user to do exactly the same thing in stored data.

Moreover, to activate QA function of the system, a user has to *ask a question* explicitly. In this sense, a QA system also forces a user to do something extra that may interrupt his/her work. As seen in those examples, it has not been well considered so far how a multimedia system should care the users and how it should not disturb them.

In this research, we propose a framework for video-based interactive media for gently giving instructions. Unlike conventional electronic instruction manuals, the system observes a user by image and speech recognition, and gives related information or appropriate advices by utilizing pre-recorded video archive. The distinctive feature of the system is that the system quietly observes a user and helps the user only when he/she really needs a help, for example, in a situation that the user is at a standstill or that the user asks a question. When the system recognizes that the user does not need any help, it only presents related information that may be useful to the user, and it does not require any responses from the user.

Currently, our target is a teaching system for assembly works on a desk top, and we are developing an experimental system for the task of assembling toy blocks. Although the system is still under development, we have implemented fundamental functions of the above framework: a method for matching a user's status to the data stored as video manuals, a method for disambiguation, a method for gently giving instructions, etc. We are currently going further toward automating this system and toward combining with question answering.

2 Video-Based Media not too Interfering

Experienced human instructors have a variety of functions for teaching and helping students. Among those functions, we are focusing on the following points:

- Recognition of a user's status is essential. Human instructors carefully watches students and recognizes what they are doing or intending to do, etc. Even if he/she cannot recognize the status, he/she asks the user what the user wants to do, or what the trouble for the user is.
- A user should be allowed to do anything as much as possible. Exactly following instructions given by a teacher is not always a good way for learning, and it is against creativity.
- Overprotecting and without too much interference is harmful to learning.
- Appropriate advice should be given to a user when the user needs a help. For example, one of the most effective ways is question answering.

We are constructing a prototype system for partially realizing such functions. First, let us consider the fundamental processes. They are shown in Figure 1:

- Indexing to instruction videos
- Recognition of objects and the user's actions.
- Status matching between user's current status and indices of videos
- Presenting information relevant for the user's status

Among these, the method of video indexing is basically the same as that of QUEVICO(1), which delineates which information is required for which situation.

For recognizing objects and user's actions, we previously proposed our object tracking system for desktop works(3). However, it needs to be improved for collecting sufficient user's information and object status, and we are currently developing the next system. The system tracks objects held by hands, and recognizes changes on objects such as inflating/deflating and operations such as attaching or splitting objects, etc. Details are introduced in another paper (4).

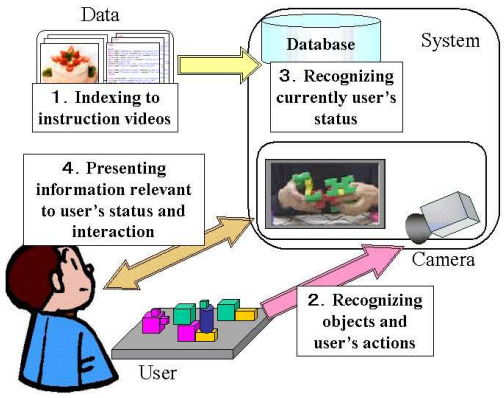


Figure 1: Video-based interactive media

On the other hand, we describe image recognition process of user's status in Section 3 in this paper. We implemented a status matching method, which continuously recognizes the user's current status by comparing the objects and actions of the user with indices of pre-recorded instruction videos. This contributes the flexibility of instruction in the sense that an instruction appropriate to the user's state is given to a user however the user is a beginner, skilled, positive, or negative to the task.

The data presentation according to the user's status is presented in Section 3.2 and 4. The system tries to recognize the user's status by matching the status with those that can potentially occur, and to choose appropriate information for the user. We are currently preparing at least two or three presentations for the same step of works. Those are used differently according to the user's previous state, past records, or current conditions.

The process is quietly performed inside the system, and it does not interfere the user as long as the system can recognize the status. In this case, related information is presented on a screen, and the system does not care whether the user watches it or not. On the contrary, when the user's status is out of a scenario, *i.e.* the video manual or it has too much ambiguity, the system inquires to the user for fixing the problem.

In the following sections, we will focus on the status matching and disambiguation.

3 Recognizing User's Status

3.1 Definitions and Description

For recognizing a user's status, continuous recognition of objects and the user's actions is essential(2),

since direct estimation of the user's intentions is difficult. To simplify the recognition problem, we consider a work as a collection of tasks that are composed of primitive user actions and concerning objects.

The followings define them:

Work: A work is a collection of tasks as shown in Figure 2, and it is the goal of instruction.

Task: A task is a primitive function that is essential for assembly works. A task consists of action(s) and objects that appear in the task. We are currently using two types of task, "move an object" and "attach an object to another". Since other categories such as detaching or reshaping, are surely necessary for usual works. We are currently constructing a system that recognizes such phenomena. The pattern of "attaching task" is shown in

Action: An action is a primitive motion of a user, each of which is assumed to be recognized by image processing. We are currently using "lift an object", "place(put) an object" and "make two objects touched each other". An action is also described by an ID, a name, and concerning objects.

Object: An object is one of the components/parts that is visible and that has a concrete shape. It is described by an ID and the characteristics of visible features, *e.g.* color, shape, texture, etc.

We consider that a task is composed of actions, and relation is defined as an *action pattern*. Figure 3 shows an example for a task "attaching two objects", where O_i represents an object.

Those data are given for each pre-recorded instruction video, and stored as indices. Since the cost of this indexing is not negligible, we expect that our object and action recognition system under development can be also available for this indexing process.

3.2 Status Recognition

Figure 4 shows the overview of the matching process.

(a) The indices of videos is input to the system, and the task graph, object data, are reconstructed from the indices of a video.

(b) When a user does something that can be recognized as an action, the corresponding action and concerning objects are recorded and added to the list of actions.

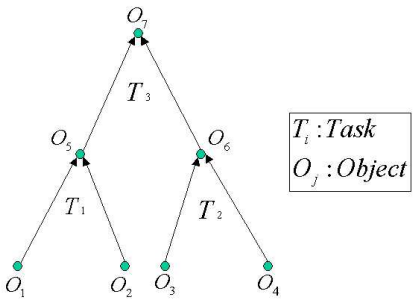


Figure 2: Representation of a work

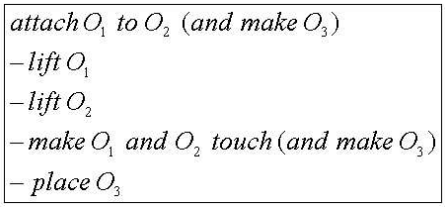


Figure 3: Pattern of attaching task

(c) The user's current status is recognized by comparing (a) and (b).

Step (c) is composed of "partial search" and "whole search" that will be described below.

Partial search: The system searches for a task, *i.e.* a set of consecutive actions, that matches a task in video indices. We use DP matching(5) between a sequence of user actions and a sequence of actions in video data, since a user does not always move exactly the same as recorded. All

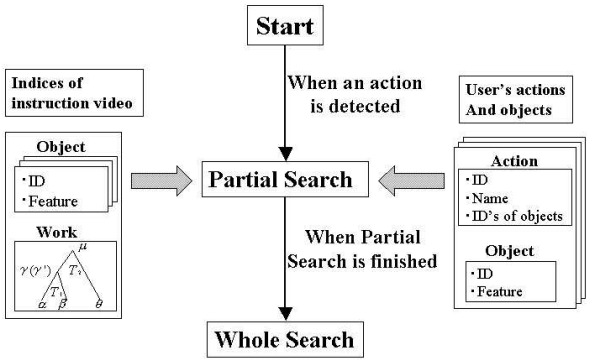


Figure 4: Recognition of user's status

possible matches are searched, and the consistency among tasks are not considered in this step.

Whole search: The possible sequences of tasks are determined by checking the consistency among objects and tasks. We use simple depth-first search for obtaining the possible combinations of tasks. Since the number of candidates suffers from combinatorial explosion, we need further mechanism for disambiguation by interacting with a user. This will be described in the next section.

For the above matching step (c), the following criteria are used.

Similarity between objects $S(O_i, O_j)$: Similarity between objects is calculated based on object features such as color, shape, etc. Currently, we use the color histogram of an object region.

Similarity between actions $S(A_i, A_j)$: Similarity between actions is calculated by the product of "similarity between action names" and "similarity between concerning objects". Suppose an action $A_1(\mathrm{N}_1, O_{11}, O_{12}, ..., O_{1n})$ and $A_2(\mathrm{N}_2, O_{21}, O_{22}, ..., O_{2n})$, where $\mathrm{N_i}$ means an action name and O_{ij} means a concerning object. The similarity between actions S(A_1,A_2) is calculated by the following formula.

$$S(A_1, A_2) = \delta(N_1, N_2) \cdot \{ \prod_{i=1}^{n-m} S(O_{1i}, O_{2i}))^{1/(n-m)} - m * C_{\mathrm{P}} \} \quad (1)$$

$$\delta(X, Y) = \begin{cases} 1 & (X = Y) \\ 0 & (X \neq Y) \end{cases} \quad (2)$$

where, m is the number of objects that are not matched, C_{P} is a constant that represent the penalty value for an unmatched object.

Similarity between tasks: Since it is difficult to recognize a task directly by image processing, the task the user performed is not directly matched to those in the video data. Tasks are recognized through comparing a sequence of actions in the above partial search process.

4 Interacting with Users

4.1 Data presentation to users

Video data relevant to an user's status, such as the explanation of the succeeding tasks, is presented to the user. One important aspect of this framework is that we believe that the system should not interfere the user by the interaction. This requirement is often difficult to satisfy, since the system needs to recognize when the user really needs a help. Before implementing such functions, we consider the way of displaying helpful information based on the user's status.

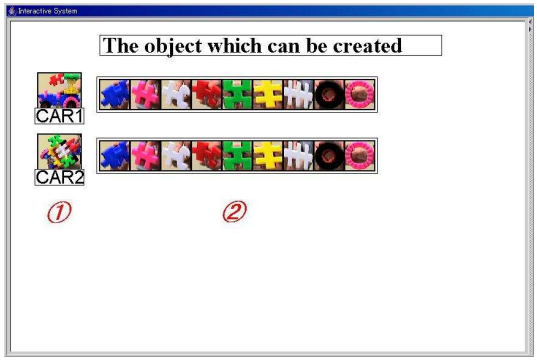


Figure 5: Information display when no object is held,

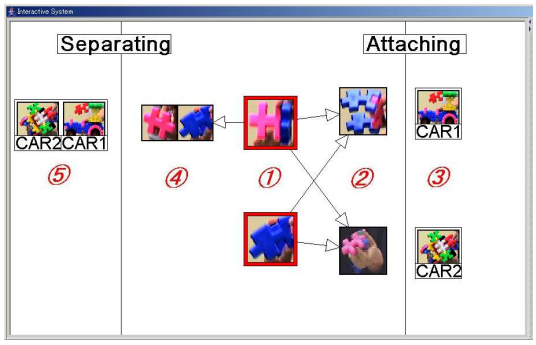


Figure 6: Information display when two objects are held.

For this purpose, the information of objects held by a user is one of the most important keys, *e.g.*, object's name, the number of objects, and their changes. In the near future, we are planning to use the user's behaviors, *e.g.*, holding still an object, or trying to attach objects again and again, etc.

Figure 5 shows an example of displaying information that is given when the user is not holding any object. The system shows two goals that can be reached in this context. The two objects in the leftmost column show the target objects that can be assembled by the object in the right columns.

Figure 6 shows the display when the user holds two objects. It shows the following kinds of information:

- what the held objects are.
- what can be made by joining two objects.

- the target that can be finally made by using those objects.
- what objects can be obtained if a user split the objects into two or more parts.
- the target that can be finally made by using those split objects.

Those displays are shown to a user.

4.2 Inquiry to users

Complete recognition of the user's status is a difficult problem even with the above mechanism. Interpretation of each action and an object is ambiguous, and the ambiguity of combinatorial explosion. We need additional mechanism for efficiently disambiguating the possible situations.

For this purpose, we are preparing a method of inquiring to the user. First, the system checks which portion is much ambiguous. The number of candidate objects or tasks in the videos is one of the indicator of the ambiguity. When the number of candidates are over a threshold value, the system inquires to the user about his/her status. By confirming an object name, a task name, or other things, considerable degree of ambiguity can be reduced. The followings are typical inquiry methods for objects and tasks.

Object: The system shows similar objects and asks the user to choose the correct one. For example, if one object held by a user has many candidates, *i.e.* objects in the video data, the system asks "what is the object in your hand?", or "which object is the same as the object you hold" by presenting objects list as shown in Figure 7. The user, for example, will choose the correct one by touch panel.

Task: The system shows a similar task and asks the user "are you doing this task?", "have you done this?", etc. An snapshot is shown in Figure 8.

The answer from the users are used for choosing the correct correspondence between a real object/task and a object/task in the instruction videos. Then status matching is performed again by using the correct correspondence of that portion.

5 Experiments

For checking the potential of this system, we conducted preliminary experiments. An instruction video is taken for an actual assembly of a toy block

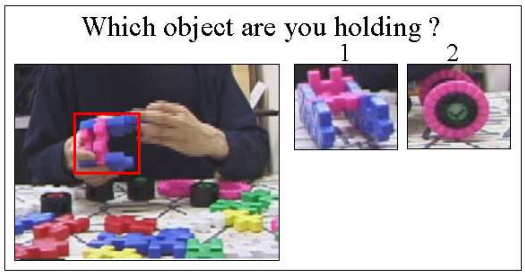


Figure 7: Inquiring about an object

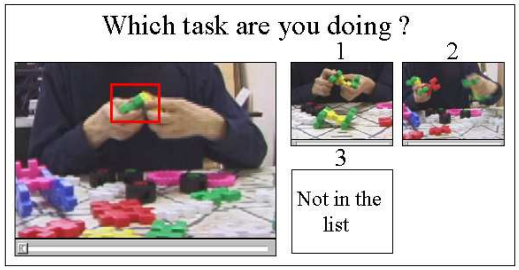


Figure 8: Inquiring about a task

car as shown in Figure 9. The toy block car is composed of 50 blocks and the work consists of 30 tasks. In another video, we also recorded the behaviors of a person who was asked to make the toy car by watching the video. From both videos, objects and motions are manually detected and given to the system. From the video of user's behaviors, about 70 actions are detected.

As a result of the first partial search, 90% of the tasks are correctly detected with false alarm of 95%. And figure 10 shows the precision and the recall rate of this experiment.

Figure 11 shows objects and the number of candidates for some objects detected during the experiment. Num(1) column shows the number of candidates without inquiries for disambiguation. The whole search for current status is not possible at this step, since each portion has too much ambiguity. Therefore, the system executed disambiguation by user inquiry. As object ID9 has 19 interpretation candidates, it is the most ambiguous object. By inquiring to the user, the system obtained the correct matching between the objects. In this case, after obtaining the answer from the user, the improved result was shown at Num(2) in Figure 11. By this disambiguation, the interpretation candidates of object ID7 is also reduced, since the object ID7 and ID9 are used in the same task.

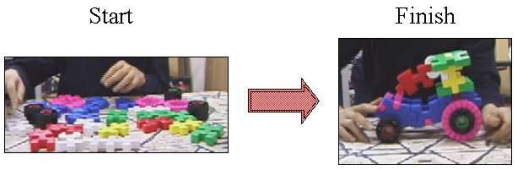


Figure 9: Assembly of a toy block car

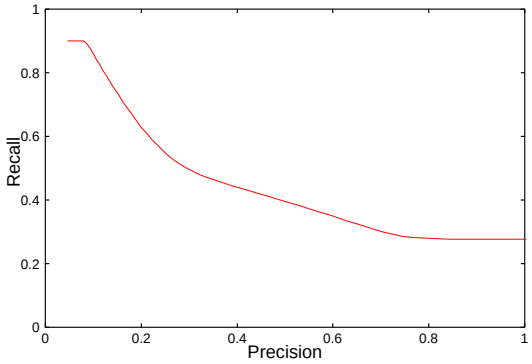


Figure 10: Precision-recall graph

By repeating this type of inquiry, the system eventually gets a small number of candidates that include correct one.

6 Conclusion

In this paper, we introduced our idea of video-based interactive media that gently supports users. We proposed the framework for recognizing user's status and the method for reducing ambiguity by inquiring to users. In our preliminary experiments, this system succeeded in handling a toy-car assembly work in which 50 objects are used and 30 primitive tasks are required.

Our system is, however, still under development. To realize a realtime system, we need further intensive works. Integration of image processing portion is the most urgent topic. Building a good user interface is also an important topic. Currently, user study is ongoing, and we expect that we clarify which help would be comfortable and effective for different people, such as beginners/skilled, positive/negative, child/grown-up, etc.

As future applications, we hope this framework will be extended to more general scenes, such as instructing something in our home, training in a school, etc. For this purpose, we need further investigation

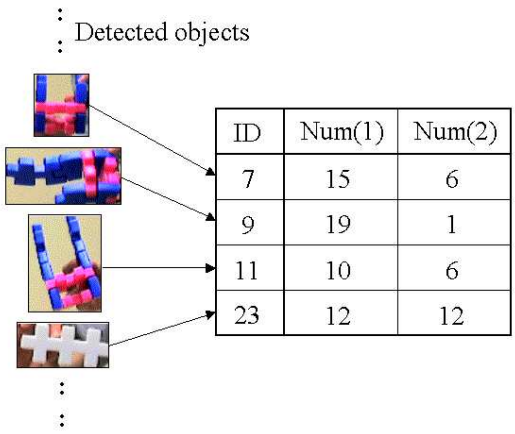


Figure 11: The number of the matched objects

of recognizing human behaviors and objects' states. We need to introduce sensor and ubiquitous computing techniques as well as more advanced computer vision techniques.

References

[1] Hidekatsu IZUNO, Yuichi NAKAMURA and Yuichi OHTA: QUEVICO QA Model for Video-based Interactive Media, Proc. Third International Workshop on Content-Based Multimedia Indexing, pp.413-420, 2003.

[2] Cen RAO, Mubark SHAH: A View-Invariant Representation of Human Action. Control, Automation, Robotics and Vision, 2000.

[3] Masatsugu ITOH, Yuichi NAKAMURA, Yuichi OHTA, "Simple and Robust Tracking of Hands and Objects for Video-based Multimedia Production", IEEE Intn'l Conference on Multi Sensor Fusion and Integration, pp.252-257, 2003.

[4] Yosuke TSUBUKU, Yuichi NAKAMURA, Yuichi OHTA, "Object Tracking and Object Change Detection in Desktop Manipulation for Video-based Interactive Manuals", Pacific-Rim Conference on Multimedia (to appear), 2004.

[5] Seiichi NAKAGAWA. Pattern Information Processing (in Japanese). Maruzen, 1999.

Automatic Content Production for an Autonomous Speaker Agent

Karlo Smid[*†]
[*]Ericsson Nikola Tesla
Krapinska 45, HR-10002 Zagreb
karlo.smid@ericsson.com

Igor S. Pandzic[†]
[†]Faculty of electrical engineering and computing, Zagreb University
Unska 3, HR-10002 Zagreb
Igor.Pandzic@fer.hr

Viktorija Radman[‡]
[‡]Ericsson Nikola Tesla
Krapinska 45, HR-10002 Zagreb
viktorija.radman@ericsson.com

Abstract

We present a graphically embodied animated agent (a virtual speaker) capable of reading plain English text and rendering it in a form of speech accompanied by the appropriate facial gestures. Our system uses a lexical analysis of an English text and statistical models of facial gestures in order to automatically generate the gestures related to the spoken text. It is intended for the automatic creation of the realistically animated virtual speakers, such as newscasters and storytellers and incorporates the characteristics of such speakers captured from the training video clips. Our system is based on a visual text-to-speech system which generates a lip movement synchronized with the generated speech. This is extended to include eye blinks, head and eyebrow motion, and a simple gaze following behavior. The result is a full face animation produced automatically from the plain English text.

1 Introduction

Our intelligent content production system is an extension of the Visual Text-to-Speech (VTTS) system. A classical VTTS system (Pelachaud et al., 1996; Legoff et al., 1997; Lewis et al., 1987; Smid et al., 2002) produces lip movements synchronized with the synthesized speech based on timed phonemes generated by speech synthesis. Normally, it also solves the coarticulation problem (Beskow et al., 1995; Cohen et al., 1993; Lundeberg et al., 1999). A face that only moves the lips, looks extremely unnatural because natural speech always involves facial gestures. However, a VTTS system can only obtain phonetic information from speech synthesis and has no basis for generating realistic gestures. Very often this problem is solved by introducing some partially random gestures triggered by a set of rules (Ostermann et al., 2000). Another solution is recording one or more sequences of facial gestures from real speakers and then playing those tracks during a speech (Pandzic, 2002). These methods produce better visual results than a static talking face, but the movements are generally too simplistic. Yet another approach is to manually insert tags or bookmarks into a text from which the facial gestures or expressions are generated (Pandzic et al., 2002). Obviously, this is time consuming and unsuitable for the fully automatic applications. The Eyes Alive system (Lee et al., 2002) introduces a full statistical model of eye movement based on the known theory of eye movement during speech, as well as precise recordings of eye motion during speech. The system reproduces eye movements that are dynamically correct at the level of each movement and that are globally also statistically correct in terms of frequency of movements, intervals between them and their amplitudes. However, the movements are still unrelated to the underlying speech content, punctuation, accents etc. In natural speech, most gestures are directly related to the lexical structure of speech and have distinct functions (Cassell et al., 2002; Argyle et al., 1976; Collier, 1985; Chovil, 1992). The BEAT system (Cassell et al., 2001) uses linguistic and contextual information contained in a text to control the movements of hands, arms and a face, and the intonation of a voice. The mapping from a text to the facial, intonational and body gestures is contained in a set of rules derived from a state of the art research in nonverbal conversational behavior. That mapping also depends on the knowledge base of an ECA environment. That

knowledge base is populated by a user who animates the ECA so the production of the facial gestures is not automatic. Furthermore, the system does not introduce a full statistical model for the supported gestures so occurrences of the supported nonverbal features could be too predictive. Also, in the current set of supported nonverbal behaviors, head nods and eyes blinks must be included.

We propose a new approach that combines the lexical analysis of input text with the statistical model describing frequencies and amplitudes of facial gestures. The statistical model is obtained by analysing a training data set consisting of several speakers recorded on video and stenographs of their speech. A lexical analysis of the stenograph texts allowed to correlate the lexical characteristics of a text with the corresponding facial gestures and to incorporate this correlation into a statistical model. Using a lexical analysis of input text to trigger this statistical model, a virtual speaker can perform gestures that are not only dynamically correct, but also correspond to the underlying text. Figure 1. depicts training and content production processes.

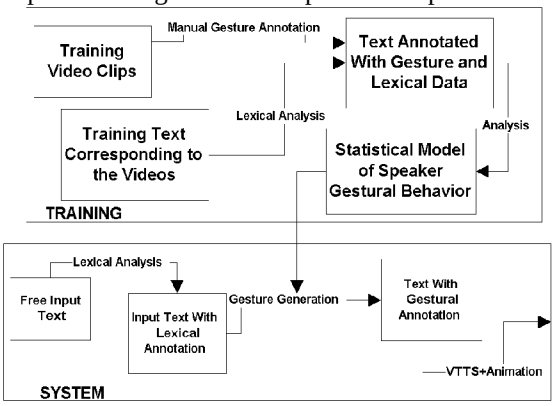


Figure 1: Training and content production processes.

2 Background

A conversation consists of two domains: verbal and nonverbal. These two domains are highly synchronized because they are driven by the same forces: the prosody and lexical structure of the uttered text as well as the emotions and personality of a person that is involved in a conversation (Faigin, 1990). The verbal domain deals with a human voice, while body and facial gestures (head, eyes and eyebrows movement) are part of the nonverbal domain. In this article, our focus is on facial gestures and how they are synchronized and driven by the prosody and lexical structure of uttered text.

Facial gestures are driven by (Ekman et al., 1969):

interactional function of speech: we unconsciously use facial gestures to regulate the flow of speech, accent word or segments, and punctuate speech pauses.

emotions: they are usually expressed with facial gestures.

personality: it can often be read through facial gestures.

performatives: for example, advice and order are two different performatives and they are accompanied with different facial gestures.

In this article we deal with the interactional function of speech. In this context, facial gestures can have several different roles, usually called determinants (Pelachaud et al., 1996). These determinants are:

conversational signals: they correspond to the facial gestures that clarify and support what is being said. These facial gestures are synchronized with accents or emphatic segments. Facial gestures in this category are eyebrow movements, rapid head movements, gaze directions and eye blinks (Ekman, 1979).

punctuators: they correspond to the facial gestures that support pauses; these facial gestures group or separate the sequences of words into discrete unit phrases, thus reducing the ambiguity of speech (Collier, 1985). The examples are specific head motions, blinks or eyebrow actions.

manipulators: they correspond to the biological needs of a face, such as blinking to wet the eyes or random head nods because being completely still is unnatural for humans.

regulators: they control the flow of a conversation. A speaker breaks or looks for an eye contact with a listener. He turns his head towards or away from a listener during a conversation (Duncan, 1972). We have three regulator types: Speaker-State-Signal (displayed at the beginning of a speaking turn), Speaker-Within-Turn (a speaker wants to keep the floor), and Speaker-Continuation-Signal (frequently follows Speaker-Within-Turn). The beginning of themes (an already introduced utterance information) are frequently synchronized by a gaze-away from a listener, and the beginning of rhemes (new utterance information) are frequently synchronized by a gaze-toward a listener.

Since we are currently concentrating on Autonomous Speaker Agent, which is not involved in a conversation but performs a presentation, this work focuses on conversational signals, punctuators and manipulators. All these functions are supported by a fairly broad repertoire of facial gestures. We distinguish three main classes of facial gestures

(Pelachaud et al. 1996):

Head movement
Eyes movement
Eyebrows movement

Within each class we distinguish specific gestures, each characterized by their particular parameters. The parameters that are important for head and eyebrow movements are amplitude and velocity. Those two parameters are in inverted proportion. A movement with a big amplitude is rather slow. Table 1 shows the types of facial gestures as identified during our data analysis (Section 4). This is an extension of the classification proposed in (Graf et al., 2002). We introduce symbols incorporating both a gesture type and a movement direction.

Table 1: The specification of facial gestures.

Head	Nod	^ v > <	An abrupt swing of a head with a similarly abrupt motion back. We have four nod directions: up and down (^), down and up (v), left and right (<) and right and left (>).
	Overshoot nod	~	Nod with an overshoot at the return, i.e. the pattern looks like an 'S' lying on its side.
	Swing	u d L R diag	An abrupt swing of a head without a back motion. Sometimes rotation moves slowly, barely visible, back to the original pose, sometimes it is followed by an abrupt motion back after some delay. Five directions: up (u), down (d), left (l), right (R) and diagonal (diag).
	Reset	reset	Sometimes follows swing movement. Returns head in central position.
Eyes	Movement in various directions		The eyes are always moving. Parameters are: gaze direction, points of fixation, the percentage of eye contact over gaze avoidance, duration of eye contact.
	Blink		Periodic blinks keep the eyes wet. Voluntary blinks support conversational signals and punctuators.
Eyebrows	Raise	^^	Eyebrows go up and down.
	Frown		Eyebrows go down and up.

3. Lexical analysis of English text

The speech analysis module (Radman, 2004) performs the linguistic and contextual analysis of a text written in English language with a goal of enabling the nonverbal (gestures) and verbal (prosody) behavior assignment and scheduling.

Starting from a plain English text, it produces an XML document annotated with tags for each word (Table 4). These tags allow us to distinguish between the newly introduced words, words known from a previous text and punctuation marks. Based on this knowledge, the process, described in Section 5, assigns and schedules the gestures.

The input text is first phrase-parsed because the module needs to know the morphological, syntactic and part-of-speech information. In order to get the morphologic data about the words in a sentence, we have developed a module that classifies words according to the English grammar rules. In the first release we used the Connexor's Machinese Phrase Tagger[1] (MPT). MPT is a commercial tool, its public interface is changing from version to version, and it has much more functionality that we needed. So we have made the simplified version of the morphologic and semantic analyzer extending WordNet 2.0[2] database. In order to determine the correct word type based on the output queried from the extended WordNet 2.0 database, we must pass multiple times through the whole sentence and apply various English grammatical rules. WordNet 2.0 database contains nouns, verbs, adverbs and adjectives, so for other English word types we made our own database using the MySQL engine. Our database contains auxiliary verbs, determiners, pronouns, prepositions and conjunctions. Besides that, one additional table has been created. This table was dynamically filled with new data each time some particular word was not found either in WordNet 2.0 or in our database. After passing through numerous examples, we concluded that 99% of these words were nouns. We can say that our module is learning based on the examples that passed through it. For every query to WordNet 2.0 and our database, we got more than one word type for a particular word. In order to get correct type of a word, we must pass multiple times through the whole text and apply various grammatical rules to it. Here are some of the rules:

1. Every determiner (a, an, the) is followed by an adjective or a noun.

[1] http://www.connexor.com/

[2] http://www.cogsci.princeton.edu/~wn/

2. Every personal pronoun (I, he, you, ...) is followed by a verb, an adverb, or an auxiliary verb.
3. Every possessive pronoun (my, your, hers, ...) is followed by a noun or an adjective.

Table 2 represents an example.

Table 2: An example for the word type classification.

Word	Word type	Lemma
The	POS=DET	
drop	POS=N	drop
in	POS=PREP	
order	POS=N	order
intake	POS=N	intake
that	POS=PRON-DEM	
Ericsson	POS=NNF	
reported	POS=V	report
for	POS=PREP	
the	POS=DET	
third	POS=ADJ	third
quarter	POS=N	quarter
raised	POS=V	raise
a	POS=DET	
great	POS=ADJ	great
deal	POS=N	deal
of	POS=PREP	
concern	POS=N	concern
on	POS=PREP	
the	POS=DET	
market	POS=N	market
.	POS=punct	

Table 3: Parameters of the morpholical analyzer.

Parameter	Meaning	Example
A	Adjective	blue, sweet
ADV	Adverb	quickly, very
CC	Conjunction	Or, and, but
DET	Determiner	the, a, an
N	Noun	boy, dog
PREP	Preposition	in, by, out, from, to,
PRON	Pronoun	I, you, me, your, mine
AUX	Auxiliary verb	had, must
V	Verb	drive, drives, take, takes,
NNF	Word that is not found in WordNet 2.0 and our database	SmartPhone, Unicom
punct	Punctuation mark	. ! ?
token	Another mark	“ : ‘ ( ) %

In the second step we break the paragraphs (UTTERANCE) into clauses (CLAUSE). The largest unit is UTTERANCE, which represents an entire paragraph of input. The next, smaller, unit is CLAUSE, which is held to represent a proposition. In order to detect clauses in an utterance, the module is searching for the punctuation marks and a placement of verb inside a phrase.

Table 4: Input and output of Speech Analysis Module.

Module input	Module output
However, figures presented by Business Unit Systems prompted more positive reactions.	<?xml version="1.0" encoding="UTF-8"?> <UTTERANCE> <CLAUSE> <WORD Text="However" New="Yes"/> <WORD Text=","/> <WORD Text="figures" New="Yes"/> <WORD Text="presented" New="Yes"/> <WORD Text="by"/> <WORD Text="Business"/> <WORD Text="Unit" New="Yes"/> <WORD Text="Systems" New="Yes"/> <WORD Text="prompted" New="Yes"/> <WORD Text="more" New="Yes"/> <WORD Text="positive" New="Yes"/> <WORD Text="reactions" New="Yes"/> <WORD Text="."/> </CLAUSE> </UTTERANCE>

The smallest unit is a word with its new attribute. To determine the newness of each word, we keep track of all previously mentioned words in an utterance. We also use WordNet 2.0 database to identify sets of synonyms. We tagged each noun, verb, adverb or adjective as new if they or their synonyms had not been seen in an utterance before. Other word classes are not considered for the new parameter. Since pronouns need to be tagged as new and WordNet 2.0 does not process them at all, an algorithm is proposed to deal with the pronouns. The logic for this algorithm is based on knowledge and intuition, but that, of course, does not lead us to the universal solution. After going through many Connexor analyzing examples and studying all the pronouns found in them, the following conclusion has been made: every pronoun, substituting a noun that appears after it, or a noun that does not appear in a text at all, needs to be tagged as new. The algorithm is as follows:

Every pronoun, that is not preceded by a noun in a sentence, and is part of the following set: ("any", "anything", "anyone", "anybody", "some", "somebody", "someone", "something", "no", "nobody", "no-one", "nothing", "every", "everybody", "everyone", "everything", "each", "either", "neither", "both", "all", "this", "more", "what", "who", "which", "whom", "whose")

or any pronoun that is a part of the following set: ("I", "you", "he", "she", "it", "we", "they"), gets the new tag assigned. All other pronouns, which do not fulfill the above-stated requirements, are not tagged with new.

4 Statistical model of gestures

In this section we present the statistical model of facial gestures and the methods, tools and datasets used in order to build it.

4.1 Training Data Set

As a training set for our analysis, we chose Ericsson's "5minutes" video clips. Those clips are published by LM Ericsson for internal usage and offer occasional in-depth interviews and reports on major events, news or hot topics from the Telecom industry. They are presented by professional newscasters. We used a footage showing the newscasters (Figure 2). We investigated three female and two male Swedish newscasters.

Figure 2: Tonya's nod with overshoot[3]

4.2 Data Analysis

First, using a video editing tool, we extracted the news casting extracts from the video, according to their stenographs. Then we grouped those news extracts for every observed speaker. Observing those news casting clips, we marked the starting and ending frames for every eye blink, eyebrow raise and head movement (Figure 2). Analyzing those frames, the speakers Mouth-Nose Separation unit (MNS0) value, facial gesture amplitude value, facial gesture type (Table 6) and direction were determined. We used the following algorithm for amplitude values: the values represent the difference between the speaker's nose top position at the end and the beginning of a facial gesture (an eyebrow raise or head movement).

Data values, that were gathered from the video clips, were statistically processed using Microsoft Excel and MatLab 5.3. That means that a number of pie charts (Figure 3) were produced by simply calculating how many times were facial gestures triggered/not triggered by words. Every gesture type has a corresponding pie chart. Amplitude values probabilities (Figure 4) were calculated using the histogram statistical function.

Table 5 presents an example of the gathered raw data for one news extract.

Table 5: An example of the data set gathered during the analysis.

word	52		Three	arraignments
eyes	3		blink::cs	
head		|up;A=2	|d to n |d A=0.25	Id A =0.5
eyebrows	2		raise::cs A=1/4	
pitch	13		+	
lexical	44		new	new
			cs - conversational signal	
			p – punctuator	
			m - manipulator	
		~nod::A1=2:A2=0.5::cs		

The word row contains an analysed news extract separated word-by-word. The eyes, head and eyebrows rows hold data about facial motion that occurred on the corresponding word (separated with the :: symbol), the type of motion and its direction (according to the notation summarised in Table 6), amplitude value (A stands for amplitude) and determinant code values (cs, p, m summarised in Table 5). The head basic motions are mapped to head movements as described in the last column of Table 1. We replayed the newscaster footage to determine the facial gestures type, direction and duration parameters. The last row in Table 5 contains the head movement facial gesture parameters.

The pitch row indicates which words were emphasised by voice intonation. The lexical row holds information about word's newness in the text context (Section 3). The second column in Table 5 represents the number of occurrences of a particular facial gesture and pitch accents, the number of words in the current news extract and the number of words that are new in the text context.

[3] Published with permission of LM Ericsson

4.3. Determinant Values of Facial Motions

A determinant value for a particular facial motion is determined as follows. If a facial motion occurred on a punctuator mark, then a determinant for that motion was the punctuator (p). If a facial motion accompanied a word that is new in the context of uttered text, then a determinant was the conversational signal (cs). Otherwise, a determinant of a facial motion was the manipulator (m). The raw data tables were populated by manual analysis and measurement. All amplitude values were normalized to MNS0 for the particular speaker. MNS0 is a Facial Animation Parameter Unit (FAPU) in the MPEG-4 Face and Body Animation (FBA) standard (Pandzic et al., 2002). Using MNS0 FAPU our model could be applied to every 3D model of a speaker.

Table 6: Basic facial motions triggered by words.

Facial gesture	Type of motion	Description
Head movement	| up	vertical up
	| up to n	vertical up to neutral (centre) position
	| down	vertical down
	| down to n	vertical down to neutral (centre) position
	-- to left	horizontal left
	-- to right	horizontal right
	-- to n	horizontal to neutral (centre) position
	/ up	diagonal up from left to right[4]
	/ down	diagonal down from right to left4
	\ up	diagonal up from right to left4
	\ down	diagonal down from left to right4
Eyebrows movement	raised s	eyebrows going up to maximal amplitude
	raised e	eyebrows going down to neutral position

[4] From the listener point of view.

In our model, the basic unit which triggers facial gestures is a word. We chose not to subdivide further into syllables or phonemes for simplicity reasons. Since some facial gestures last through two or more words, this level of subdivision seems appropriate.

The raw data (Table 5) for the complete training set was statistically processed in order to build a statistical model of speaker behaviour. A statistical model consists of a number of components, each describing the statistical properties for a particular gesture type in a specific speech context. A speech context can be an old word, a new word or a punctuator. The statistical properties for a gesture type include the probability of occurrence of particular gestures and histograms of amplitude and duration values for each gesture. Figure 3 shows an example of a statistical data component for head gestures in the context of a new word. It is visible that in 51% of occurrences, we have some kind of head movement. For example, the probability of occurrence for rapid head movements is 22%. Further, we have five directions: up (u), down (d), left (L), right (R) and diagonal (diag). In the end, we must determine the amplitude for a rapid movement. Figure 4 shows a linear approximation of the cumulative histogram for the amplitude of a rapid head movement.

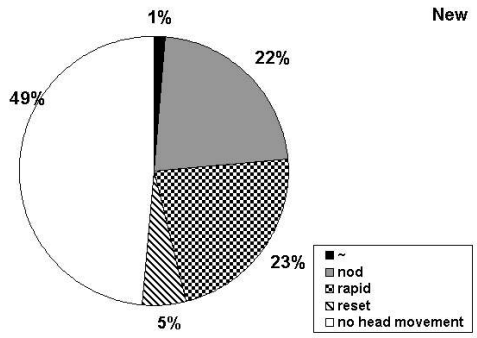


Figure 3: Statistical data for head gestures occurrences in the context of a new word.

Such statistics exist for each gesture type and for each speech context we treated. They are built into the decision tree (Figure 5) that triggers gestures. The process is described in the following section. Note that, in the context of punctuators, only eyes gestures are used, because the statistics show that other gestures do not occur on punctuators.

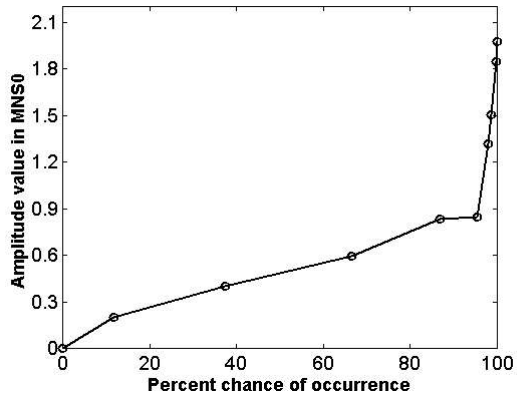


Figure 4: The linear approximation of the cumulative histogram for the amplitude of a rapid head movement.

5. The System

Figure 7 shows the complete Autonomous Speaker Agent system. The input to the system is plain English text. It is processed by lexical analysis (Section 3) which converts it into an XML format with lexical tags (currently describing new/old words and punctuators). The facial gesture module is the core of the system – it actually inserts appropriate gestures into text in the form of special bookmark tags. These bookmark tags (Table 7) are read by the TTS/MPEG-4 Encoding module. While the Microsoft Speech API (SAPI) Text To Speech (TTS)[5] engine generates an audio stream, the SAPI notification mechanism is used to catch the timing of phonemes and bookmarks containing gesture information. Based on this information, an MPEG-4 FBA bitstream is encoded with the appropriate viseme and facial gestures animation. For MPEG-4 FBA bitstream generation, we are using Visage SDK API[6] that uses SAPI 4.0 or 5.1. Visage SDK API uses information provided by the SAPI notification mechanism.

5.1 Facial Gesture Module

The facial gesture module is built upon the statistical model described in the previous section. The statistical model is built into the decision tree illustrated in Figure 5.

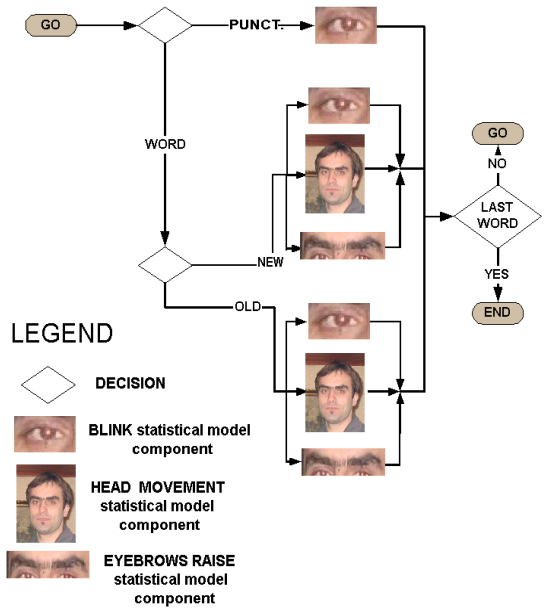


Figure 5: Decision tree with components of the statistical model

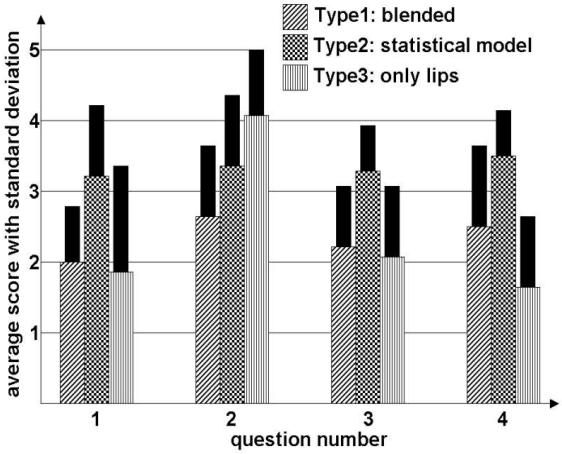


Figure 6: Results of subjective evaluations. Average score and standard deviation.

Let us follow the decision tree (Figure 5). The first branch point classifies the current context as either a word or a punctuation mark. Our data analysis showed that only eye blink facial gesture had occurred on the punctuation marks. Therefore only the blink component of the statistical model is implemented in this context.

The words could be new or old (Section 3) in the context of uttered text - this is the second branch point. All facial gestures occurred in both cases but with different probabilities. Because of that, in each case we have different components for facial gestures parameters. From Figure 5, it is obvious that a word could be accompanied by all three types

[5] Microsoft speech technologies http://www.microsoft.com/speech/ 29/03/2004

[6] Visage Technologies AB http://www.visagetechnologies.com/ 29/03/2004

of facial gestures at the same time. The facial gesture signals (eye blink, head movement, eyebrow raise) are generated separately, based on their statistical component data. They will be blended later in the TTS/MPEG-4 encoding component. The output from the facial gesture module is plain English text accompanied by bookmark pairs for facial gestures.

Table 7: SAPI bookmark codes of facial gestures.

Bookmark code	Facial gesture
\Mrk=1\	conversational signal blink
\Mrk=2\	punctuator blink
\Mrk=300\	eyebrows raise
\Mrk=400\	nod ^
\Mrk=700\	nod V
\Mrk=1000\	nod <
\Mrk=1300\	nod >
\Mrk=9\	rapid reset
\Mrk=1600\	rapid d
\Mrk=1900\	rapid u
\Mrk=2200\	rapid L
\Mrk=2500\	rapid R
\Mrk=2800\	rapid diagonal

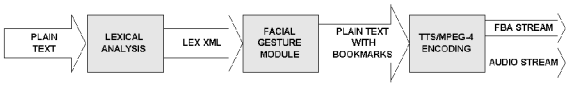


Figure 7: The data flow through the Autonomous Speaker Agent system.

Every facial gesture has a corresponding pair of bookmarks: one bookmark marks the starting moment of a facial gesture and the other marks the ending moment. Table 7 shows values for each bookmark. The head and eyebrows movement bookmark values not only define the type of facial gesture, but also contain the amplitude data of a facial movement. For example, bookmark value 2300 defines the rapid head movement to the left (symbol L) of amplitude 1 MNS0. The function for amplitudes of facial gestures is:

$$A=((Bmk\_value - Bmk\_code)/100) \quad (1)$$

The interval for bookmark values for L is [2200,2500> because the statistical model showed that the maximal amplitude value for facial gesture L was 2.2 MNS0. Head nods and eyebrow raises could last through two or more words. Statistics have shown that the maximum duration of a nod is five words, an eyebrow raise could last through eleven words and the maximal duration for a nod with an overshoot is eight words. We code a nod with an overshoot as two nods: a nod up immediately followed by a nod down. Every nod has its own amplitude distribution.

5.2 TTS/MPEG-4 Encoding Module

TTS/MPEG-4 encoding module, using the bookmark information, encodes an MPEG-4 FBA bitstream with an appropriate viseme and gestures animation. The animation model for head and eyebrow movement facial gestures is based on the trigonometry sine function. That means that our Autonomous Speaker Agent nods his head following the sine function trajectory.
We have implemented a simple model of gaze following, meaning that the eyes of our Autonomous Speaker Agent are moving in the opposite direction of a head movement. This gives an impression of an eye contact with the Autonomous Speaker Agent.

5.3 Results

We conducted a subjective test in order to compare our proposed statistical model to simpler techniques. We synthesized facial animation on our face model using three different methods. In the first (Type 1), head and eye movements were produced playing animation sequence that was recorded by tracking movements of a real professional speaker. In the second (Type 2), we produced a facial animation using the system described in this paper. In the third (Type 3), only the character's lips were animated.
We conducted a subjective test to evaluate the three types of facial animation. The three characters (Type 1, Type 2 and Type 3) were presented in random order to 29 subjects. All three characters presented the same text. The presentation was conducted in the Ericsson Nikola Tesla and all subjects were computer specialists. However, most of the subjects were not familiar with virtual characters, and none of the subjects were authors of the study. The subjects were asked the following questions:

Q1: Did the character on the screen appear interested in (5) or indifferent (1) to you?

Q2: Did the character appear engaged (5) or distracted (1) during the conversation?

Q3: Did the personality of the character look friendly (5) or not (1)?

Q4: Did the face of the character look lively (5) or deadpan (1)?

Q5: In general, how would you describe the character?

Note that higher scores correspond to more positive attributes in a speaker. For questions 1 to 4, the score was graded on a scale of 5 to 1.

Figure 6 summarizes the average score and standard deviation (marked with a black color) for the first four questions. From the figure, we can see that the character of type 2 was graded with the highest average grade for all questions except for the Q2. The reason for that is because type 3 character only moves its lips and its head is static. This gave the audience the impression of engagement in the presentation. A Kruskal-Wallis ANOVA indicated that the three character types had significantly different scores ($p = 0.0000$).
According to general remarks in Q5, the subjects tended to believe the following:

1. Type 1 looked boring and uninteresting, it seemed to have cold personality. Also, implemented facial gestures were not related to the spoken text.
2. Type 2 had a more natural facial gesturing and facial gestures were coarticulated to some extend. Head movements and eye blinks are related to the spoken text. However, eyebrow movements were with unnatural amplitudes and were not related to the spoken text.
3. Type 3 looked irritating, stern and stony. However, it appeared to be concentrated and its lips animation was the best.

6 Conclusion and future work

According to feedback that we have received from the audience, we can conclude that our statistical model of facial gestures can be used in a system that implements a fairly convincing Autonomous Speaker Agent. Furthermore, the implemented decision tree produces better animation than previous techniques. The problem with eyebrow amplitudes can be easily solved by changing some of the parameters in the TTS/MPEG-4 encoding module. Also, with statistical data that we have gathered during our work, we have confirmed some of the conclusions of other papers. We confirmed that, on average, the amplitude of a faster head nod is lesser than the amplitude of a slower nod. Furthermore, we concluded that words, that bring something new in the utterance context, are very often accompanied by some facial gesture.
However, our system is not ready yet for the Turing test. An extension to Embodied Conversational Characters is the logical item for future work, extending the system to support the natural gesturing during a conversation and not only for independent speakers. This will include adapting and extending the statistical model to include more complicated gesturing modes and speech prosody that occur in a conversation. Also, new statistical data should be calculated based on the existing training set data. In this calculation, it must be taken into consideration that words form two higher logical groups: OBJECT and ACTION. Because those groups are labeled with new and old tags, this fact results in new statistical data for facial gestures. Also, coarticulation of facial display occurrences must be taken into consideration during the production of this new statistical model.
Modifying speech prosody (Hiyakumoto et al., 1997; Parent et al., 2002; Silverman et al., 1992) of input text according to statistical prosody data of professional speakers would produce a much more convincing Autonomous Speaker Agent. In order to get more natural head movements, the velocity dynamics (Hadar, 1983) of those movements must be implemented in the TTS/MPEG-4 encoding module. New Visage SDK API works with Microsoft SAPI 5.0 engine. That engine uses XML notation for user defined bookmarks and, because of that, output of the Facial Gesture Module should be adapted to this new notation.

Acknowledgements

This research is partly supported by Ericsson Nikola Tesla (ETK), Zagreb, Croatia and Visage Technologies AB, Linköping, Sweden.

References

M. Argyle and M. Cook. Gaze and mutual gaze. *Cambridge University Press*, 1976.

J. Beskow. Rule-based visual speech synthesis. *EUROSPEECH 4th European Conference on Speech Communication and Technology*, ESCA-4(1):299–302, Madrid, 1995.

J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Embodied Conversational Agents. *The MIT Press Cambridge, Massachusetts London, England*, 2000.

J. Cassell, H. Vilhjálmsson and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. *SIGGRAPH 2001*, ACM :477–486, 2001.

N. Chovil. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163-194, 1992.

M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. *Models and Techniques in Computer Animation*, 139-156, Springer-Verlag, Tokyo, 1993.

G. Collier. Emotional expression. *Hillsdale, N.J.: Lawrence Erlbaum Associates*, 1985.

S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292, Oxford University Press, 1972.

P. Ekman and W. Friesen. The repertoire of nonverbal behavioral categories-Origins, usage, and coding. *Semiotica*, 1:49-98, 1969.

P. Ekman. About brows: Emotional and conversational signals. *Human ethology: Claims and limits of a new discipline*, 169-249, New York: Cambridge University Press, 1979.

G. Faigin. The artist's complete guide to facial expression. *Watson-Guptill Publications, New York*, 1990.

H. P. Graf, E. Cosatto, V. Strom and F. J. Huang. Visual Prosody: Facial Movements Accompanying Speech. *AFGR 2002*, 381-386, 2002.

J. Gratch. Emile: Marshalling Passions in Training and Education. *Fourth International Conference on Autonomous Agents*, ACM Press 325 - 332, 2000.

U. Hadar, T. Steiner, E. Grant and F. C. Rose. Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, 2:35-46, 1983.

L. Hiyakumoto, S. Prevost and J. Cassell. Semantic and Discourse Information for Text-to-Speech Intonation. *ACL Workshop on Concept-to-Speech Generation 1997*, Madrid. 47.-56, 1997.

S. P. Lee, J. B. Badler and N.I. Badler. Eyes Alive. *29th annual conference on Computer graphics and interactive techniques 2002, San Antonio, Texas, USA*, ACM 637 - 644, 2002.

B. Legoff and C. Benoît. A French speaking synthetic head. *ESCA Workshop on Audio-Visual Speech Processing 1997*, Rhodes, Greece, 145-148, 1997.

J.P. Lewis and F.I. Parke. Automated lipsynch and speech synthesis for character animation. *Human Factors in Computing Systems and Graphics Interface 1987*, 143-147, 1987.

M. Lundeberg and J. Beskow. Developing a 3D-agent for the August dialogue system. *AVSP1999, Santa Cruz, USA*, 1999.

J. Ostermann and D. Millen. Talking heads and synthetic speech: An architecture for supporting electronic commerce. *ICME 2000*, 71.-74, 2000.

I.S. Pandzic. Facial Animation Framework for the Web and Mobile Platforms. *Web3D Symposium 2002, Tempe, AZ, USA*, 27 - 34, 2002.

I. S. Pandzic and R. Forchheimer. MPEG-4 Facial Animation - The standard, implementations and applications. *John Wiley & Sons*, 2002.

R. Parent, S. King and O. Fujimura. Issues with Lip Synch Animation: Can You Read My Lips? *Computer Animation 2002, Geneva, Switzerland*, 3–10, 2002.

C. Pelachaud, N. Badler and M. Steedman. Generating Facial Expressions for Speech. *Cognitive, Science*, 20(1), 1-46, 1996.

V. Radman. *Leksicka analiza teksta za automatsku proizvodnju pokreta lica*, Graduate work no. 2472 on Faculty of Electrical Engineering and Computing, University of Zagreb, 2004.

K. Silverman, M. Beckman, J. Pitrelli, M. Osterndorf, C. Wightman, P. Price, J. Pierrehumbert and J. Herschberg. ToBI: A Standard for Labeling English Prosody. *Conference on Spoken Language, 1992, Banff, Canada*, 867-870, 1992.

K. Smid and I.S. Pandzic. A Conversational Virtual Character for the Web. *Computer Animation 2002, Geneva, Switzerland*, 240 - 248, 2002.

An Audiovisual Information Fusion Approach to Analyze the Communication Atmosphere

Tomasz M. Rutkowski*, Koh Kakusho*, and Michihiko Minoh*
*Academic Center for Computing and Media Studies
Kyoto University, Japan
tomek@mm.media.kyoto-u.ac.jp

Victor V. Kryssanov†
†College of Information Science & Engineering
Ritsumeikan University, Japan

Anca Ralescu‡
‡ECECS Department
University of Cincinnati, Ohio, USA

Abstract

This paper discusses the problem of multimodal analysis of human face–to–face communication situations from audiovisual recordings. The main goal of the presented study is to parameterize communicative/interactive events in multimedia (audio and video) for analysis and subsequent edition. Interactive, environmental, and emotional characteristics of the communicators are estimated to define the communication event as a whole. This calls for the adoption and integration of various results obtained in social sciences and multimedia signal processing into one framework – the communication atmosphere analysis and reconciliation – to determine and possibly improve the overall climate of multimedia communication. Experiments have been conducted, their results are discussed, and conclusions are drawn.

1 Introduction

Although common and often used to describe a conversation episode, the term ”communication atmosphere” resists a precise definition. Like many other concepts related to human behavior, and of utmost importance to perform intelligent information processing with social attributes, communication atmosphere is easily felt but hard to define and, therefore, difficult to evaluate.

This study identifies three characteristics of the communication process that can be tracked and that, from an information processing point of view, are essential to create and recreate a meeting's climate (understood as communication atmosphere) from an observer/spectator's point of view.

The communication atmosphere is a qualitative measure somewhat similar to the concept of *affordance* proposed by Gibson (1977), since it evaluates communicative situations from their usability (potential understandability) point of view. The analysis focuses in this paper on qualitative evaluation of visual and auditory attention of the communicators. Norris (2004) defines attention level as *”the degree of clarity of an experience ranging from unconsciousness (total lack of awareness) to focal attention (vivid awareness)*. Observed (recorded) communicative events can then be evaluated, based on attention/awerness levels that they exhibit. Moore et al. (1996) studied the impact of information amount that impinges on students' sensory registers during lectures. The authors reported that students' involvement in the learning process was decreased when their attention was not throughly focused. Attention, hence, plays an important role in selecting sensory information for efficient understanding of communication that is directly related to the communicative involvement of the participants or spectators. In the proposed model of communication atmosphere, three elements (dimensions) of communicative situations are defined, based on interactive (social) features, emotional (mental) characteristics of the communicators, and environmental (physical) features of the place where the event occurs.

There are contemporary studies on communicative activity monitoring, e.g. (Chen, 2003), where the problem of automatic activity evaluation in audio and visual channels for distance learning applications was discussed in detail. The known approaches are limited, however, to communicative interaction evaluation only, without considering other aspects of the communication atmosphere, like environmental and emotional.

There are at least two possible application areas

for the approach described in this paper, which provides for an intelligent analysis of the communication process and the communication atmosphere as well. Indexing of multimedia archives comes as the first possible application, where the expected content of a communication situation is sought, based on criteria specified by the user. The second and more challenging area is the reconciliation of already captured multimedia recordings to modify their atmosphere according to the user's preferences. These applications both require for tracking elements of human communication, which can be used to evaluate the process from a behavioral (interaction) point of view.

Next section presents the three-dimensional communication atmosphere analysis approach. Section 3 elaborates on the proposal for multimedia record editing in the three-dimensional communication space. Examples, experimental results, and a discussion of them concludes the paper.

2 Communication Atmosphere

The problem of monitoring and subsequently modeling the communication atmosphere arises when one has to choose among factors constituting the event and determine relationships between them such as:

- the extent, to which the communicators' behavior (interactions) influences the communications climate;
- the extent, to which the external environment, where the communication happenes (e.g. a room) affects the communication efficiency;
- the extent, to which emotional states of the communicators shapes the final situation of the event.

To address these factors, which ultimately build up the communication atmosphere, and evaluate information streams related to the attention level engendered by the communication process, we propose a threefold "intelligent" approach, in which the problems are analyzed in a three-dimensional space, merging diverse sensory information and making assumptions about typical communication behavior. The dimensions of the communication atmosphere are defined as follows:

Environmental - to describe the communication place or physical space conditions. This dimension is merely to comply with the fact that communication episodes (e.g. meetings) are not conducted in "the vacuum" but in certain places, which can largely create or at least influence the communication climate and can thus enhance or disturb the spectator's attention.

Communicative - to characterize the communicative behavior. As it will be shown later, this dimension is to reflect the communicators' ability to interact and, therefore, it is related to the communication efficiency, which is a qualitative measure of the communication process directly related to the attention level and to the dynamic involvement of the communicators (Kryssanov and Kakusho, 2005).

Emotional - to estimate emotional states of the people. This dimension is to recognize the fact that the behavior related to emotional states of the communicators to a large extent determines the communication atmosphere. It is expected that emotions exposed by the communicating parties are similar, or else emotions shown by the sender are somewhat reflected (after a delay) by the receivers (Dimberg et al., 2000; Pease and Pease, 2004) creating an average emotional state of the situation.

In the case of multimedia content, such as video captured during a meeting or discussion, a separate analysis of the communication atmosphere along the dimensions appears to be easier to perform from the technological point of view. The dimensions can then be dealt with independently, while any possible interdependencies are out of the scope of this paper.

Anthropologist Birdwhistell (1974), in his study of kinetics, discussed nonverbal communication. According to his results, human can recognize about 250,000 facial expressions, which is a huge number for artificial intelligence classifiers that can handle a much smaller number of static facial expressions (Ralescu and Hartani, 1997; Pantie and Rothkrantz, 2000). Mehrabian (1971) found that the verbal component of face–to–face communication is less than 35%, and that over 65% of communication could be conducted non-verbally. Experiments reported in (Pease and Pease, 2004) also demonstrated that in business related encounters, body language accounts for between 60% and 80% of the impact made around the negotiating table.

Following this discussion, our approach will then be focused on only dynamical analysis of nonverbal components of communication. The hypothesis presented in this paper suggests that observation of the nonverbal communication dynamics would allow us to estimate the situation's climate, i.e. the communication atmosphere.

2.1 Environmental Dimension

This dimension is used to describe the state of the environment, in which the communication event takes place. Highly visually or auditory intensive environments affect the overall impression of the observed/perceived communication situation. On the other hand, since the communicators have usually a limited ability to change the environmental features (i.e. the level of external audio or video), this study recognizes the environmental characteristics as a distinct feature set taken for the overall evaluation of communication. Physical features of the environment can be extracted after separation of the recorded information streams into two categories: items related to the communication process and items unrelated to (useless for) it (Rutkowski et al., 2004). The general idea is to split the audio and video streams into background noise and useful signals produced by the communicators.

We will first detect the presence of auditory and visual events that occur in the space but are not related to the communicators' actions (i.e. background audio and video). In the current approach, the analysis of the environmental dimension is performed in two stages:

- noise and non-speech power level difference extraction;
- non-communication-related visual activity (background flow) estimation.

These procedures both can avert the attention of the listeners. The amount of the environmental audio energy (classified as noise) is estimated as a segmental-signal-to-interference-ratio (SIR) since calculation of an integral-signal-to-noise-ratio would not reflect temporal fluctuations of the situation dynamics (e.g. when communicators speak louder or quieter). The segmental auditory SIR, A_{SIR}, is evaluated as:

$$A_{SIR}(m) = 10\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1} s_e^2(n)}{\sum_{n=nM}^{Nm+N-1} s_n^2(n)}, \qquad (1)$$

where $s_n(n) = s_o(n) - s_e(n)$ is the noise estimate after removing from the original signal its denoised version; $s_o(n)$ is the recorded audio signal (with noise); N stands for the number of audio samples; M represents the number of windows, into which the speech utterance is segmented. For visual information, we compare the activity features detected in the communicators' areas with the remaining background. We calculate the amount of visual flow in the signal as an interference–like coefficient, V_{SIR}:

$$V_{SIR}(m) = 10\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1} v_h(n)^2}{\sum_{n=nM}^{Nm+N-1} v_b(n)^2}, \qquad (2)$$

where $v_b(n)$ represent the background visual flow features and $v_h(n)$ are related to the extracted motion features of the active communicators. Both A_{SIR} and V_{SIR} are then summed up and form a single audiovisual SIR measure characterizing the environmental conditions.

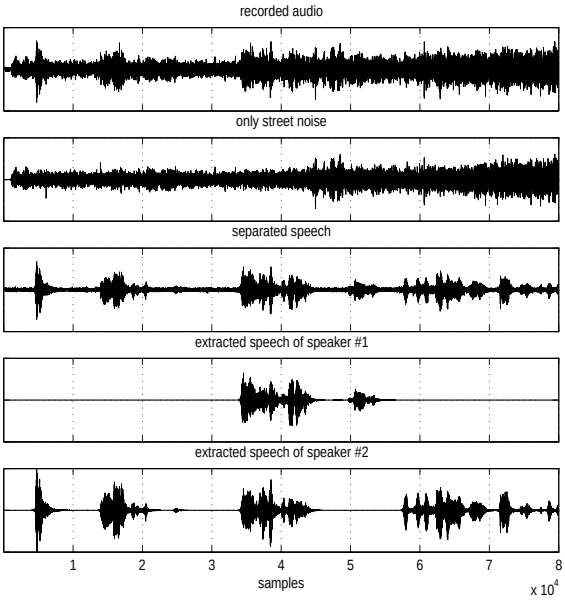


Figure 1: Environmental problem example shows that the wide-band noise might completely cover the speech usable frequencies (compare recorded signal in top box, only interference in second box from the top, enhanced speech in middle box and finally separated activities of two speakers - there is a part in the middle, when both speakers were active).

Figure 1 illustrates the evaluation of the environmental dimension, when the audio and video components of the environment are filtered from the information related to the communication process.

2.2 Emotional Dimension

This dimension is introduced to characterize the inner (cognitive) states of the communicators. Contemporary approaches to emotional state estimation usually deal with a static analysis of facial expressions and seldom consider a dynamic or multimodal analysis. Experiments reported in (Dimberg et al., 2000) and (Pease and Pease, 2004) showed that un-

conscious mind exerts direct control of facial muscles. It was demonstrated that facial emotions presented by a sender (e.g. a smile) were reciprocated by returning a smile by the receiver. The experiments of Dimberg et al. (2000) were conducted using electromyography, so the actual muscle activity of communicators could be captured. Results of those experiments revealed, that the communicators often have no total control over presented facial emotions. These findings suggest that important aspects of "emotional face–to–face" communication could occur on the unconscious level. In our approach we estimate emotional states from available nonverbal features only, which actually reflect (partially) the psychic state of the communicator. Emotions of the communicators can be estimated from their speech only Dellaert et al. (1996), since the video features are hard to consider, owing to the video insufficient resolution and problems with a simple definition of emotions captured from moving faces and bodily expressions. In the current approach, emotions from speech are estimated along three features: the voice fundamental frequency with durational aspects of the stable fundamental frequency periods; the speech harmonic content; and the signal energy expressed as an averaged root mean square (RMS) value in voiced frames (Rutkowski et al., 2004). Primary emotions, such as neutral, sad, angry, happy and joyful, which the communicator might experience during the communication event, are determined using a machine learning algorithm (Vapnik, 1995).

2.3 Communicative Dimension

The third and, probably, most important component of the communication atmosphere analysis is the communicative dimension. It refers to the communicators' audiovisual behavior and to their ability to "properly" interact during the conversation. The synchronization and interaction measures (efficiency-like) developed in the authors' previous research are applied here (Rutkowski et al., 2003, 2004). The communication model used – the hybrid linear/transactional – is linear in short time windows (Adler and Rodman, 2003). The active (in short time windows) communicator - the sender - is supposed to generate more audiovisual flow with breaks, when the receiver responds. The passive (in short time windows) communicator – the receiver – is expected to, on the other hand, react properly, not disturbing (overlapping with) the sender's communication activity. Turn-taking (role changing) between the senders and receivers is a critical assumption in the

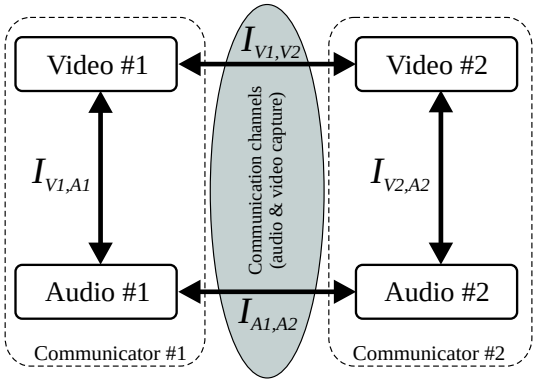


Figure 2: Scheme for the communication synchronicity evaluation. Mutual information estimates I_{A_1,V_1} and I_{A_2,V_2} between audio and visual features streams of localized communicators account for the local synchronization. The estimates I_{A_1,A_2} and I_{V_1,V_2} are to detect crosstalks in the same modality.

hybrid communication model. Only the case of intentional communication, which occurs when all the communicators are willing to interact, is considered here. All the situations when so-called metacommunication (Adler and Rodman, 2003) occurs are out of the scope of this paper. Rutkowski et al. (2003) defined the communication efficiency as a measure that characterizes the behavioral coordination of communicators. This measure describes the communication process from the point of view of the interactive sending and receiving of messages by the communicators, as observed through the audiovisual channel. However, there is no means to evaluate the understanding of the messages by the communicators, but it is assumed that feedback or the receiver's reaction should be presented. This study puts forward a measure of the communication efficiency, as a combination of four mutual information estimates between two visual (V_i), two audio (A_i), and two pairs of audiovisual features ($A_i; V_i$). First, the two mutual information estimates are evaluated for selected regions of interest (ROI), where the communicators may be present and their speech occurs, as follows:

$$I_{A_i,V_i} = \frac{1}{2} \log \frac{|R_{A_i}|\,|R_{V_i}|}{|R_{A_i,V_i}|}, \qquad (3)$$

where $i = 1, 2$, and R_{A_i}, R_{V_i}, R_{A_i,V_i} stand for empirical estimates of the respective covariance matrices of the feature vectors (Rutkowski et al., 2003). Next, the two mutual information estimates indicating simultaneous activity in the same modes (audio and video respectively) are calculated for video streams

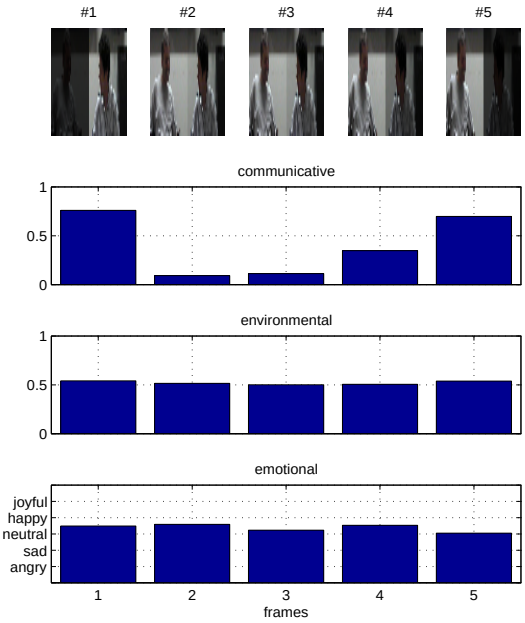


Figure 3: The communication track along the three-dimensions. This short sequence started when right–side communicator was the sender, then during three frames both were active, and finally left side communicator became the sender.

as:

$$I_{V_1,V_2} = \frac{1}{2} \log \frac{|R_{V_1}|\,|R_{V_2}|}{|R_{V_1,V_2}|}, \tag{4}$$

and, analogously, for audio streams:

$$I_{A_1,A_2} = \frac{1}{2} \log \frac{|R_{A_1}|\,|R_{A_2}|}{|R_{A_1,A_2}|}. \tag{5}$$

where R_{A_1,A_2} and R_{V_1,V_2} are the empirical estimates of the respective covariance matrices for unimodal feature sets corresponding to different communicator activities. A conceptual graph illustrating the idea is shown in Figure 2. I_{A_1,V_1} and I_{A_2,V_2} evaluate the local synchronicity between the audio (speech) and visual (mostly facial movements) flows of the observed communicators. It is expected that the sender should have a higher synchronicity reflecting the higher activity. I_{V_1,V_2} and I_{A_1,A_2} are to detect possible crosstalks in same modalities (audio-audio, video-video) of the communicators. The latter pair is also useful to detect possible overlappings in activities, that have a negative impact on the communication quality.

The communicator role (i.e. sender or receiver) can be estimated from the audiovisual mutual information features extracted and monitored over time. It is assumed that a higher synchronization across the audio and video features characterizes the active member - the sender, while the lower synchronization characterizes the receiver. This implies that in efficient communication, synchronized audiovisual behavior of the sender and unsynchronized behavior of the receiver should be observed. From the interactions in observed and recorded audiovisual streams, the communicators' roles are classified as short-time senders and receivers (Adler and Rodman, 2003; Rutkowski et al., 2003). As stated above, the efficient sender-receiver interaction during communication should involve action and feedback. The correlation of feedback with the sender's actions understood as interlaced activities is monitored. The audio-visual synchronicity (the mutual information estimate) is used for determining the roles. The pair of the mutual information estimates for the local synchronization of the senders and the receivers in equation (3) is used to give clues about concurrent individual activities during the communication event, while the unimodal cross-activities estimates from equations (4) and (5) are used to evaluate the interlaced activities for a further classification.

There are many state-of-the-art techniques that attempt to solve the problem of recognition of communicative activities in particular modalities (e.g. audio only or video only), yet human communication involves interlaced verbal and nonverbal clues that constitute efficient or inefficient communication situations. For every communication situation the members are classified as senders, receivers, or in transition. In the presented approach, the interactions between individual participants is modeled. The interaction constitutes a stream of sequences of measurements, which are classified into streams of recognized phases using a machine learning algorithm (Hsu and Lin, 2002). A multistage and a multisensory classification engine based on the linear support vector machines (SVM) approach in one-versus-rest-fashion is used to identify the phases during ongoing communication, based on the mutual information estimates from equations (3), (4), and (5). Results of the communicator phase classification during an ongoing conversation are depicted in the form of shaded windows in Figure 3. The dark shades identify the receiver, while the light shades - the sender. The transient phases are represented with the same shading for both communicators.

The communication efficiency, as proposed in (Rutkowski et al., 2003), is a normalized value calculated from the integration over time of the com-

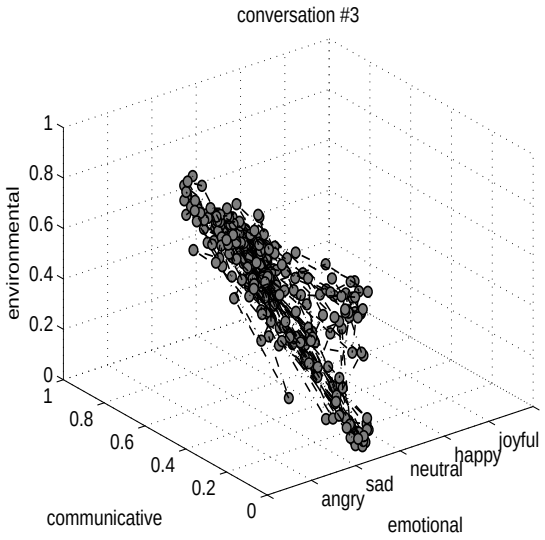


Figure 4: The communication atmosphere space spanned among three dimensions. The trajectory represents a face–to–face communication situation in a relatively quite environment.

municator's role evaluations. It reflects the analysis of the phases in which the communicators are in the process (recognized the sender and receiver, or transient phases), together with the presence of possible crosstalks in audio and video channels (enhanced or unsynchronized receiver activities). This measure estimates the interaction, and its normalized value is set to one for smooth and interlaced events, and zero for completely overlapping communication activities.

The communicative efficiency is estimated as follows:

$$C(t) = \left(1 - \frac{I_{V_1V_2}(t) + I_{A_1A_2}(t)}{2}\right) \cdot \cdot \left|I_{A_1V_1}(t) - I_{A_2V_2}(t)\right|, \quad (6)$$

and it takes values in the range $0 \leq C(t) \leq 1$, since all the mutual information estimate values are limited to $\langle 0, 1 \rangle$.

2.3.1 Tracking the Communication Atmosphere

Communication atmosphere, as defined in this study, is a region in the three-dimensional space (see Figure 4), obtained by independently estimating the environmental impact, the communication efficiency and the communicators' emotions in the ongoing communication process. This measure allows for the communication process evaluation and reconciliation (e.g. movie authoring).

The communication atmosphere definition can be formalized as follows:

$$A(t) = A\left(E(t), C(t), M(t)\right), \quad (7)$$

where $A(t)$ represents the communication atmosphere evaluation at a time t, which is a function of the environmental estimate E, the communicative estimate C (also the communication efficiency measure), and the communicators' emotion estimate M.

The estimate $A(t)$ is to characterize communication atmosphere at a given time. Less sensitive measures are the average values for a given time window:

$$A_{avg(t_a,t_b)} = \frac{1}{|t_b - t_a|} \sum_{t=t_a}^{t_b} A(t), \quad (8)$$

where $t_a > t_b$, and short time functions:

$$A_{t_a,t_b} = \{A(t_a), \ldots, A(t_b)\}, \quad (9)$$

The short time trajectories A_{t_a,t_b} in the three-dimensional space can later be used as inputs for situation classifiers or three-dimensional atmosphere models, which represent a communicative semantics modeling problem, in which multiple sequentially structured simultaneous communication processes are in a dialogical relationship. In such models, the particular focus should be put on beats and deixis (Norris, 2004), as lower-level action structure the foregrounding and backgrounding of the higher-level actions that participants are simultaneously engaged in . The problem of classification of communicative semantics–related stages is a subject for our future research, which will explore the ultimate relation between the communication atmosphere and communicative situation affordances or communicative situation norms as proposed by Stamper (1996) and modeled by Kryssanov and Kakusho (2005). The plot of short time functions A_{t_a,t_b} is presented in Figure 4 as a three-dimensional trajectory, and in Figure 3 – as the three separated bar plots for every dimension showing a vivid independence in time among the chosen dimensions.

2.3.2 The Atmosphere Reconciliation

The communication atmosphere of a recorded meeting can be reconciled according to the user preferences or wishes after the proposed analysis is accomplished. Once the communication atmosphere features are estimated, it is possible to manipulate their values and appropriately postprocess the recorded multimedia content to obtain the desired values. It

is possible to independently manipulate characteristics of the environmental dimension by increasing or decreasing the auditory or visual presence of the environment. The information separation discussed in previous sections can be used in the opposite way to add or remove the environmental components. In a similar way, it is possible to edit emotional features of the communicators' auditory activities in order to change the emotional component of the overall climate of recorded communication. The communicative dimension can also be reconciled by adjusting the occurrences of communicators' interactions (e.g. by adding or removing silent breaks) in time. An example of the original and reconciled communication atmosphere tracks (only for environmental and emotional dimensions in this case) is presented in Figure 5.

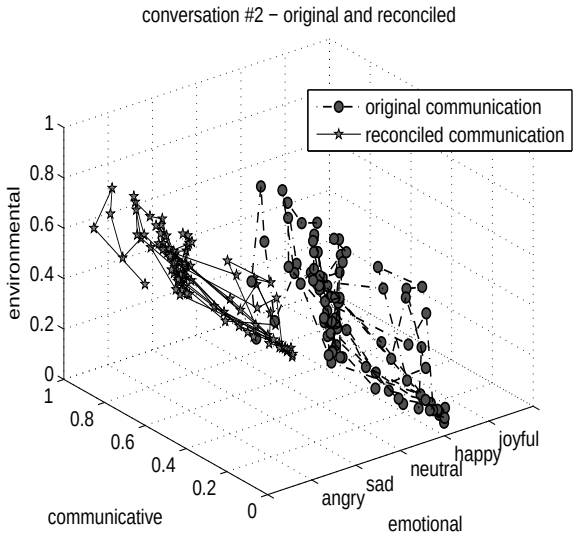


Figure 5: Communication atmosphere reconciled: The presented trajectory of a face–to–face communication was reconciled in the environmental (subtraction of environmental noise) and emotional (shifting of the emotional features of the communicators' voices) dimensions.

3 Experimental Results

The approach presented in this study was tested in two experiments. In the first, two sets of cameras and microphones where used to capture ongoing communication events. Two pairs of synchronized video and stereophonic audio streams were recorded. In the second experiment, we utilized a single high definition digital video camera (HDV) with a stereo microphone. This setup is similar to usual video recordings broadcasted in television channels. Both setups allowed capture of facial areas with higher resolution, which are highly synchronized with speech waveforms (Rutkowski et al., 2003, 2004). In both experiments conducted in laboratory controlled environments, the subjects (communicators) were asked to talk freely in face–to–face situations. We focused on the interlaced communication analysis, so that the subjects were asked to make a conversation with frequent turn taking (the discussion style). Such instruction given to subjects had a side effect of increased attention, which had positive impact on our assumption of intentional communication analysis. The experiments were conducted to validate the thesis, that the separate analysis of the three dimensions related to communication can be performed and allows for comprehensively describing the process as a whole (feature independence). For the communicative dimension, estimation of the communication efficiency is based on mutual information in multimedia streams. The track of the integrated communication efficiency value over an ongoing person-to-person communication event is shown in Figure 3. The normalized values close to one indicate the moments when the interaction was "proper", crosstalks in audio and visual channels (the local, negative synchronization) did not occur, and there was only single active person at each time (the lighter area around the person in the top box). Low values close to zero correspond to the transitional situations, or when both parties are active at the same time. The communicative situation analysis in this three-dimensional space allows for tracking of communication events and later classifying from obtained trajectories (compere the shapes of the trajectories of two independent events with different communicators but with similar discussion topics shown in Figures 4 and 5). The reconciliation procedure was performed over the the recorded communication after the analysis. In the current approach any or all the communication atmosphere dimensions can be considered. An example of a manipulated atmosphere trajectories before and after reconciliation is presented in Figure 5, where the environmental and mental dimensions were reconciled.

4 Conclusions

This study proposed a three dimensional communication atmosphere space to analyze and edit multimedia recorded communication events. It can be said that the concept of atmosphere analysis and its implementation was the missing link in contemporary stud-

ies dealing with communications situation modeling. The experimental show that it is indeed possible to estimate and later modify communicative events by changing the underlaying "affordances" qualitatively, based on behavioral analysis of the communicators. The idea of mutual information evaluation in multimodal sensory data streams makes it possible to identify the participants and to classify them according their role, evaluate their emotional states together with environmental interferences. Furthermore, separate adaptation of the proposed atmosphere–related dimensions, permits us to reconcile the event's climate. In the current study, the three dimensions of the communication atmosphere were considered independent. Interdependencies between these dimensions will be, as we stated before, addressed in the authors' future work.

References

R. B. Adler and G. Rodman. *Undestanding Human Communication*. Oxford University Press, 2003.

R.L. Birdwhistell. *Human Communication: Theoretical Explorations*, chapter The language of the body: the natural environment of words, pages 203–220. Lawrence Erlbaum Associates Publishers, Hillsdale, NJ, USA, 1974.

M. Chen. Visualizing the pulse of a classroom. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, pages 555–561. ACM Press, 2003.

F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceedings of The Fourth International Conference on Spoken Language Processing, ICSLP'96*, volume 3, pages 1970–1973, Philadelphia, PA, 1996.

U. Dimberg, E. Thunberg, and K. Elmhed. Unconscious facial reactions to emotional expressions. *Psychological Science*, 11(1):149–182, 2000.

J. J. Gibson. The theory of affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting and Knowing*. Erlbaum, Hillsdale, NJ, 1977.

C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

V.V. Kryssanov and K. Kakusho. From semiotics of hypermedia to physics of semiosis: A view from system theory. *Semiotica*, 2005. (in press).

M. Mehrabian. *Silent Messages*. Wadsworth, Belmont, California, 1971.

D. M. Moore, J. K. Burton, and R. J. Myers. Multiple-channel communication: The theoretical and research foundations of multimedia. In David Jonassen, editor, *Handbook of Research for Educational Communications and Technology*, pages 851–875. Prentice Hall International, 1996.

S. Norris. *Analyzing Multimodal Interaction - A Methodological Framework*. Routledge, 2004.

M. Pantie and L. J. M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, December 2000.

A. Pease and B. Pease. *The definitive book of body language - How to read others' thoughts by their gestures*. Pease International, 2004.

A. Ralescu and R. Hartani. Fuzzy modeling based approach to facial expressions understanding. *Journal of Advanced Computational Intelligence*, 1(1): 45–61, October 1997.

T. M. Rutkowski, K. Kakusho, V. V. Kryssanov, and M. Minoh. Evaluation of the communication atmosphere. In *Proceedings of 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES 2004, Part I*, volume 3215 of *Lecture Notes in Computer Science*, pages 364–370, Wellington, New Zealand, September 20–25 2004. Springer-Verlag Heidelberg.

T. M. Rutkowski, S. Seki, Y. Yamakata, K. Kakusho, and M. Minoh. Toward the human communication efficiency monitoring from captured audio and video media in real environments. In *Proceedings of 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems KES 2003, Part II*, volume 2774 of *Lecture Notes in Computer Science*, pages 1093–1100, Oxford, UK, September 3–5 2003. Springer-Verlag Heidelberg.

R. Stamper. Signs, information, norms and systems. In B. Holmqvist, P.B. Andersen, H. Klein, and R. Posner, editors, *In Signs of Work*, pages 349–399. De Gruyter, Berlin, 1996.

V. Vapnik. *The Nature of Statistical Learnig Theory*. Springer Verlag, 1995.

Conversational Locomotion of Virtual Characters

Soh Masuko[*†] Junichi Hoshino[†]
[*†] [†]Systems & Information Engineering, University of Tsukuba
[*†] [†]Tsukuba Ibaraki, Japan
[*†]masuko@edu.esys.tsukuba.ac.jp [†]jhoshino@esys.tsukuba.ac.jp

Abstract

Generating composite human motion such as locomotion and gestures is important for interactive applications, such as interactive storytelling and computer games. In interactive story environment, characters do not just stand in one position, they should be able to compose gestures and locomotion based on the discourse of the story and other object locations in the scene. Thus in this paper, we propose a conversational locomotion for virtual characters. We construct a conversational locomotion network for a virtual environment. The optimal walking path is calculated by a multi-pass searching algorithm which use node activation from the story locations and conversation units. The character also locally adjusts its position so that it does not occlude the referenced object from user's sight. We have applied our technique to the interactive 3D movie system, the composite motion of the character's locomotion and conversation thus strengthens the immersion in the story environment.

1 Introduction

In our daily life, human do many composite actions simultaneously. Walking and talking is one of the typical composite human actions. Composing locomotion and gestures is also important for applications such as interactive movies and games. In the interactive story environment, characters do not just stand in one positions, they should be able to compose gestures and locomotion based on the series of story locations and surrounding objects.

Figure 1. An example of conversational locomotion. The character generates composite locomotion and gestures using the story locations and local conversations.

The proper location and timing of the character is influenced by various contexts such as the connection of scene locations and the current environment. The apparent size of objects and the detail of the explanation affects how much closer the character should move. Connection of the scene locations also affects the current position. When the character refers to some objects during a conversation, and the objects are far from the character's current position, it is time consuming to make the character approach the object every time. However, when the referenced object is close to the next scene location, it is reasonable that the character moves closer to that object.

In this paper, we propose the conversational locomotion for virtual characters. This is done by calculating the optimal locomotion path which is influenced by both the conversation and the story location, the characters then generate composite walking and conversation actions. The character also locally adjusts its position not to occlude the referenced object from user's sight. Figure 1 shows a typical example of conversational locomotion. In this scenario, the character first explains that the book is in the bookshelf over there. In the next scene, the user asks about the specific contents of the book and the character moves closer to it to then explain more about the book.

2 Related Works

Locomotion generation has been previously researched in the computer graphics field with many

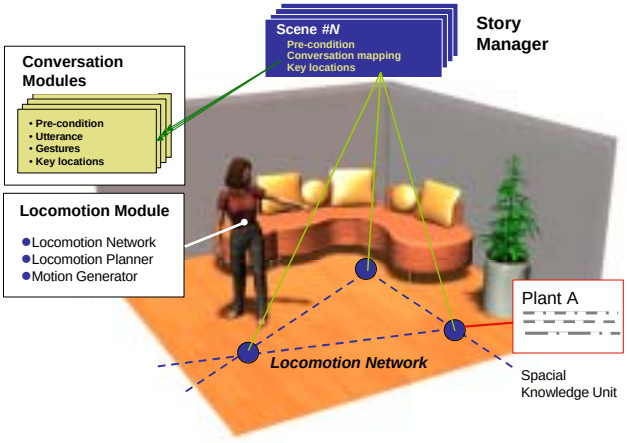
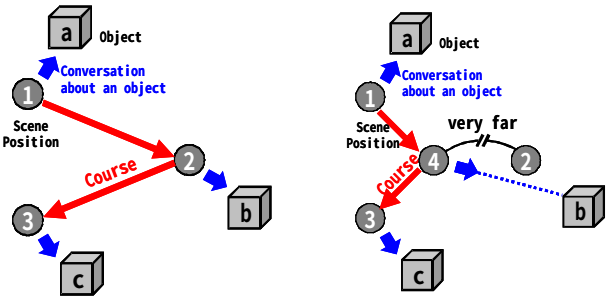


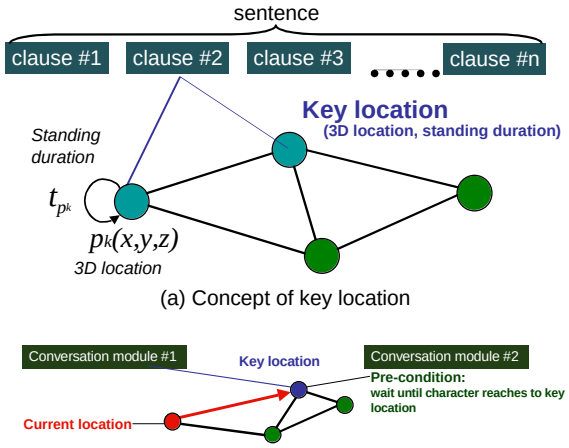
Figure 2. Overview of the conversional locomotion architecture.



a) When location (2) is near

b) When location (2) is far

Figure 3. A concept of conversational locomotion using simple example. How close the character moves to the object_b depends on the scene locations and apparent size of the referenced object.



(a) Concept of key location

(b) Decide timing of next utterance

Figure 4. Association of a clause and a key location, where locomotion is synchronized with a timing of utterance.

researchers dealing with generating walking motion patterns for various situations. Noser et al. [1995][1996] presented a navigation model for animation characters using synthetic visions. Bandi and Thalmann [1998] discretized a synthetic environment into a 3D uniform grid to search paths for autonomous characters. Whilst Rose et al. [1998] introduced the framework of "verbs and adverbs" to interpolate example motions with a combination of radial basis functions.

Rose et al. [1996] also generated seamless transitions between motion clips using spacetime constraints [Cohen 1992]. Gleicher [1998] then simplified the spacetime problem for motion re-targeting. Lee and Shin [1999] presented a hierarchical displacement mapping technique based on the multi-level B-spline approximation.

Kovar et al. [2002] introduced a motion graph to represent the transitions between poses of the captured motion data: A node of this graph represents a pose, and two nodes are connected by a directed edge if they can be followed from one to the other. Lee et al. [2002] represented motion data with a graph structure, and provided a user interface for interactive applications. Arikan and Pullen and Bregler [2002] developed a method for enhancing roughly-keyframed animation with captured motion data. Li et al. [2002] developed a two-level statistical model by combining low-level linear dynamic systems with a high-level Markov process.

Generating gestures from natural languages was developed by Cassell [1994][2000].Rickel [2002][1999] constructed a conversational system for numerous participants for team training in virtual environments. The actions of the characters are controlled using symbolic action descriptions in many layers. Towns [1997] mentioned the relationship of locomotion and conversation. When the referenced object is far from agent's current position, the agent moves near the objects in the 2D screen space until the distance from the object is under the pre-decided distance.

3 Synchronization of Conversation and Locomotion

3.1 Overview of the architecture

Figure 2 shows the conversational locomotion architecture. The system has a locomotion module, conversation modules, and a story manager. A story consists of a set of scene units and controls the discourse of the conversation. A scene unit has a pre-

condition, scene location, and links to a collection of possible conversation modules. Proper scene units are selected using the pre-conditions, such as environment change and the history of the user's utterance. When a story unit is selected, the possible conversation units applied in the scenes are activated.

Conversation modules have pre-conditions, utterance, corresponding gestures and key locations. Locomotion module dynamically plans locomotion paths and generates walking motion pattern based on key locations.

3.2 Key location control

To compose locomotion and conversation, we need to decide the character's location and the timing of walking during the conversation. The locations of the actors are influenced by where the scenes are talking place and the content of the conversations. Figure 3 shows a typical example of locomotion planning during a conversation. Assume that the actor should move from node1 to node3, it is reasonable that the actor stops at node4 if the referenced object is visible enough.

Locomotion and conversation is composed by considering following three types of location constraint:

1) Scene location: The scene location corresponds to where the actions and conversations are taking place. To begin a conversational scene, the actor should be at a proper location.
2) Interpersonal location: The character changes relative locations from the other actors during conversation. For example, when the character begins to talk it needs to approach the other participants. When the character tries to explain something, visibility of the referenced object is also considered to decide interpersonal location.
3) Referent location: This is the relative locations of the character and the referenced object.

These location constraints are used as key locations K_n in the conversational locomotion planning (Figure 4). The key location consists of a position in the floor coordinate system, and a standing duration t_k at key location. In most scenes, the proper standing position of the character has a degree of freedom. Key location has a several candidate positions with different activation values.

The standing duration of the key location can be dynamically changed by the key location control rules in the conversation units. For example, the initial standing duration can be used to decide how long the character can talk with the user at that particular position. When conversation with the user ends, the conversation units set the standing duration to zero that then causes the character to move onto the next scene location.

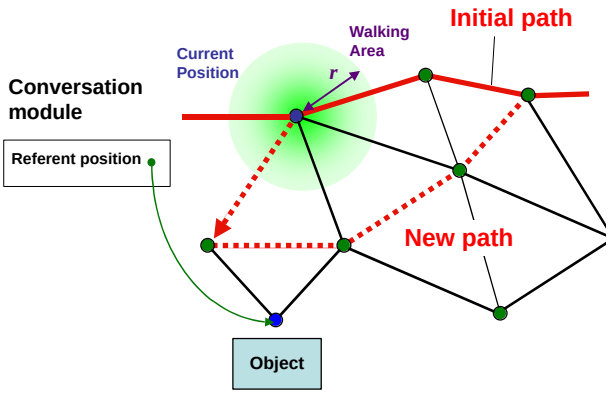


Figure 5. Key location specification. Key location associates 3D location and standing durations to a clause in utterance to control character's position and timing of walking.

3.3 Conversational locomotion network

Conversational locomotion is controlled by activating the locomotion network using story locations and conversation units. An optimal locomotion path is selected by calculating the optimal locomotion path with the maximum activation.

Locomotion nodes k represents a point on the floor coordinate systems (u_k, v_k). Characters can walk apart from locomotion nodes for local position adjustment. The locomotion network N_k=(G_k, lengthk) consists of directed graph G_k=(K, E).Where the edges representing distance between the nodes are represented as $length_k$(e). The initial locomotion network is constructed by sampling the possible standing locations. The candidate node positions are story locations and before objects referenced in conversation units. To increase the possible locations, we randomly sample the possible walking space.

Timing of locomotion is controlled by associating a key location at a proper clause in utterance. For example, the referent location can be associated with clauses including the referenced object. There are several methods for associating the key location to a clause. When the number of conversational modules are limited manual association may be easy. Even if the key location specification is pre-determined, the actual character motion is dynamically changed depending on the story locations and the order of the conversations.

The timing of utterance is also synchronized to the character locomotion. As in figure 4 (b), the pre-condition of conversation unit is used to wait until the character moves to the proper positions.

4 Conversational Locomotion Planning

Key locations are activated using scene locations and activation rules in the conversation modules.The locomotion path is dynamically selected by using a multi-pass searching algorithm that calculates the maximum activated path. When the status of activation is changed by the conversation units the locomotion path is re-calculated.

4.1 Multiple pass searching

Multiple key locations are set with a different activation value. By selecting N-best key locations, the possible locomotion segments between key locations are selected. The total activation along the locomotion segments is calculated. Locomotion segments between candidate key locations are obtained by Dijkstra method [Dijkstra 1959]. As in figure 6, we calculate candidate locomotion segments such as P_{00}- P_{01} and P_{00}- P_{02} ,to obtain the total activation value.

4.2 Activation functions

In addition to the scene locations, we use the apparent object size and walking size to locally control locomotion. Figure 7 (a) and (b) shows the activation function used in this system.

1) Apparent object size: $A(P_{s,n})$

When the apparent size of the referenced object is small, the character should move closer until it becomes large enough. We determine the activation function as figure 7 (a). The referent object is approximated by a sphere, and the view angle from user's eye position is calculated. Note that the approximated object size corresponds to the object area referred to in the conversation. When the character refers to a small area of a big object, the approximated object size is small. Orientation constraints are also integrated by forming activation distribution to a specific direction.

2) Walking distance: $D(P_{s,n}, P_{s+1,m})$

When the walking distance from the current location of the character is longer, the character tries to avoid this longer path. We determine activation function as figure 7 (b).

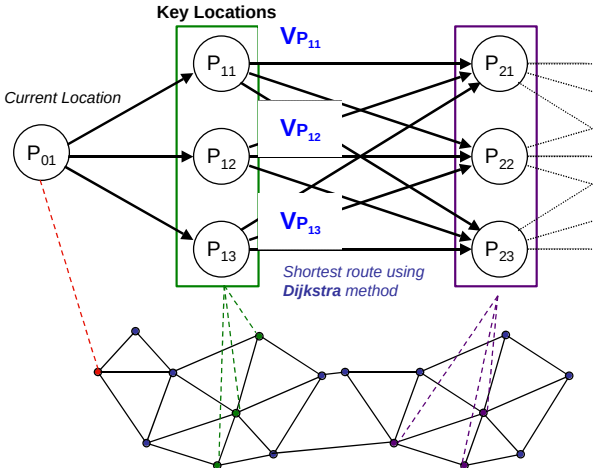
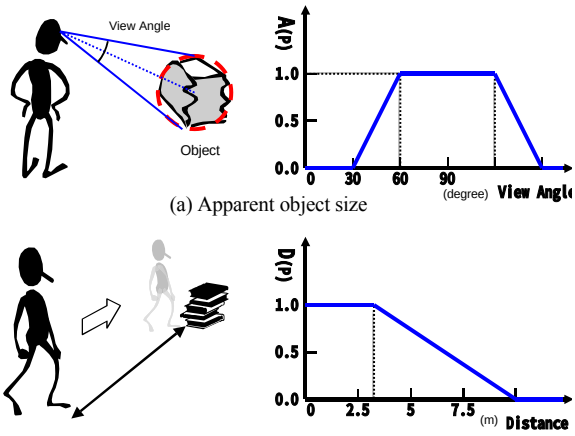


Figure 6. Key location and multiple pass searching. N-best key location is selected to search maximum activated pass.



(a) Apparent object size

(b) Distance from the object

Figure 7. Scene constraints and activation value. The apparent object size and distance from the object causes trade-off.

The total activation values are calculated along locomotion segments.

$$V(P_{0,0}, P_{1,n_1}, \ldots, P_{s,n_s})$$

$$= \sum_{t=1}^{s} \left\{ w(t) \cdot \left[\alpha A(P_{t,n_t}) + (1-\alpha) D(P_{t-1,n_{t-1}}, P_{t,n_t}) \right] \right\}$$

where w(t) is a weighting value. w(t) is used to control the number of key locations that the character should consider.

Another type of activation function is easily integrated in this framework. For example, the access control of the character to a specific area can be represented. By setting the negative activation value

to the specific locations, the character will avoid entering that place.

4.3 Local position adjustment

The actor's position is locally adjusted not to obscure the user's sight of the referenced object. Figure 8 shows the concept of local position adjustment. As described in 4.2, we approximate the referent area by using a sphere. The viewing area of the user is calculated from the 3D location of user's eye and the referent object sphere. When the character approaches the object, it stops at the intersection of the view area and the edges of locomotion networks.

5 Result

We have applied our composite motion generation technique to interactive 3D movie applications. In the interactive story environment, actors do not just stand at one position. If the actor changes its responses by user's input, the actor should then decide when and where to move.

Figure 9 shows the prototype of an interactive 3D movie system. The system uses real-time gesture and speech recognition to interact with 3D characters. We capture the user's gesture by video camera, and use silhouette and motion template matching for real-time gesture recognition. In the current system, the animation rendering module works in 8~10 frames/sec depending on the number of polygons in the scene. In the following experiments, the user's interaction was done in real-time, but the demonstration movie was rendered offline for display purposes.

5.1 Composition of locomotion and conversation

Figure 10 shows the result of conversational locomotion. To synchronize locomotion and conversation, we need to decide the location and timing of locomotion during conversation. Figure 10 (a) shows the snapshot of without locomotion planning, whilst (b) shows the snapshot of locomotion planning. We can see that the actors try to select closer positions when explaining detail.

5.2 Synchronization with user's motion

Figure. 10 also shows the result of cooperative motion of the character and user. The user's view is controlled using set of user's locomotion rules. The actor's attention is controlled by the attention module to dynamically switch the walking direction and

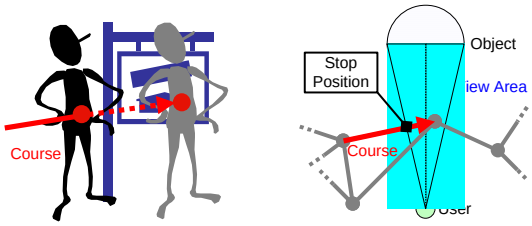


Figure 8. The local position adjustment using the user's relative position and the referenced object.

Figure 9. The prototype of the interactive 3D movie system.

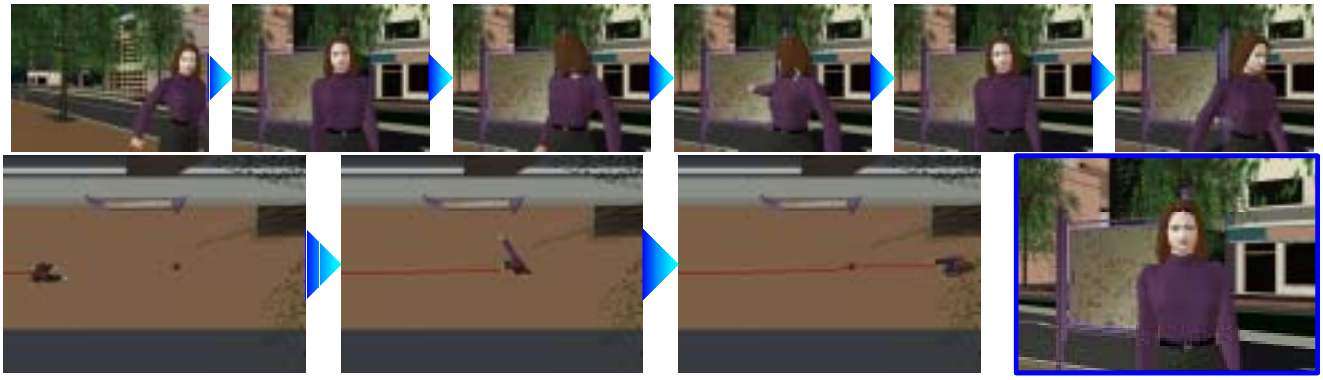
the user's direction. Such attention generates the character awareness of the user's existence.
Eye contact is important to generate a feeling that the character is aware of the user. When the character is walking and talking simultaneously, it dynamically changes attention to the walking direction and the user's direction. We apply attention control rules to direct the attention of the characters.

5.3 Limitations

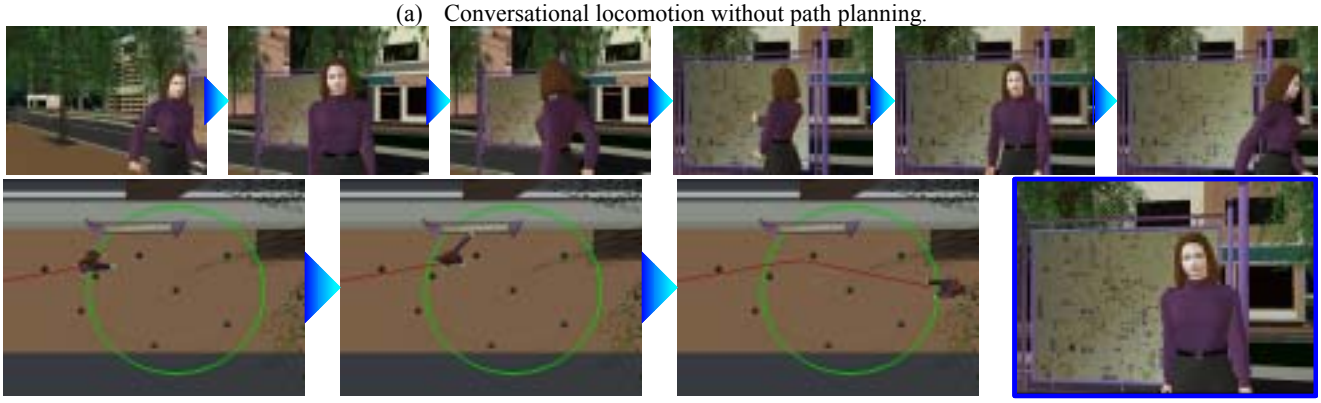
The current limitation of this system may be we do not have a function to detect the user's walking actions. Proper combination with user's locomotion devices may be useful for increasing the reality of the conversational locomotion. The other limitation is that we use simple keyword matching to select the conversation module. A more sophisticated language analysis and answer selection would be useful.

5 Conclusion

In this paper, we proposed a conversational locomotion model for virtual characters. By calculating the optimal locomotion path influenced by conversation



(a) Conversational locomotion without path planning.



(b) Conversational locomotion with path planning .

(c) Controlling locomotion and attention.

Figure 10. Snapshots from a conversational locomotion sequence.

and story locations, the characters generate composite walking and conversation actions. The character also locally adjusts the position considering the visibility of the object from the user. We have produced an interactive animation sequence using our conversational locomotion model.

References

BANDI, S. AND THALMANN, D. 1998. Space discretization for e□cient human navigation. ComputerGraphicsForum17,3, 195–206.

CASSELL,J., PELACHAUD, C., BADLER, N., STEEDMAN, M., ACHORN, B., BECKET, T., DOUVILLE, B., PREVOST, S., STONE.M., 1994. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In Proceedings of ACM SIGGRAPH '94, 413–420

CASSELL, J., BICKMORE, T., CAMPBELL, L., VILHJ´ALMSSON, H., AND YAN H.,. Conversation as a system framework: Designing embodied conversational agents. 2000. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, Embodied Conversational Agents. MIT Press, Cambridge, MA

COHEN, M. F. 1992. Interactive spacetime control for animation. ComputerGraphics(Proc.SIGGRAPH'92)26, 293–302.

DIJKSTRA, E. W. 1959. A note on two problems in connection with graphs. NumerischeMathematik1, 269–271.

GLEICHER, M. 1998. Retargetting motion to new characters. ComputerGraphics(Proc.SIGGRAPH'98)32, 33–42. Gleicher, M. 2001.

KOVAR, L., GLEICHER, M., AND PIGHIN, F. 2002. Motion graphs. ACMTransactionsonGraphics(Proc.SIGGRAPH2002)21,3, 473–482.

LEE, J. AND SHIN, S. Y. 1999. A hierarchical approach to interactive motion editing for human-like figures. Computer Graphics (Proc.SIGGRAPH'99)33, 395–408.

LESTER,J., VOERMAN, J., TOWNS, S., CALLAWAY, C.. 1999. Deictic believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. Applied Artificial Intelligence, 13:383–414

LI, Y., WANG, T., AND SHUM, H.-Y. 2002. Motion texture: A two-level statistical model for character motion synthesis. ACM Transactions on Graphics(Proc.SIGGRAPH2002)21,3, 465–472.

NOSER, H., PANDZIC, I. S., CAPIN,T.K., THALMANN, N. M., AND THALMANN, D. 1996. Playing games through the virtual life network. In Proc.Alife'96.

NOSER, H., RENAULT, O., THALMANN, D., AND THALMANN, N. M. 1995. Navigation for digital actors based on synthetic vision, memory, and learning. ComputersandGraphics19,1, 7–19.

PULLEN, K. AND BREGLER, C. 2002. Motion capture assisted animation: Texturing and synthesis. ACM Transactions on Graphics (Proc.SIGGRAPH2002)21,3, 501–508.

RICKEL, J., JOHNSON, W., 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. Applied Artificial Intelligence, 13:343–382

RICKEL, J., JOHNSON,J L.. 1999. Virtual humans for team training in virtual reality. In Proceedings of the Ninth International Conference on Artificial Intelligence in Education, pages 578–585. IOS Press

ROSE, C., COHEN, M. F., AND BODENHEIMER, B. 1998. Verbs and adverbs: Multidimensional motion interpolation. IEEE Computer Graphics and Applications 18,5, 32–40.

TOWNS, S. VOERMAN, J. ,CALLAWAY,C. LESTER, J. 1997. Coherent gestures, locomotion, and speech in life-like pedagogical agents, Proceedings of the 3rd international conference on Intelligent user interfaces. pp. 13-20

Awareness of Perceived World and Conversational Engagement by Conversational Agents

Yukiko I. Nakano*
*Japan Science and Technology Agency (JST)
2-5-1 Atago, Minato-ku, Tokyo 105-6218, Japan
nakano@kc.t.u-tokyo.ac.jp

Toyoaki Nishida†
†Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
nishida@i.kyoto-u.ac.jp

Abstract

In face-to-face communication, conversation is affected by what is existing and taking place within the environment. With the goal of improving communicative capability of humanoid systems, this paper proposes conversational agents that are aware of a perceived world, and use the perceptual information to enforce the involvement in conversation. First, we review previous studies on nonverbal engagement behaviors in face-to-face and human-artifact interaction. Based on the discussion, we implement some engagement functions into a conversational agent embodied in a story-based communication environment where multimodal recognition and generation techniques support user-agent communication.

1 Introduction

In face-to-face conversation, it rarely happens that two people are talking in an empty place, but conversations are usually affected by what is existing and taking place within a environment. Conversation is embedded in the environment, and established by linking linguistic messages to the perceived world (Clark, 2003). Aiming at improving the naturalness and the reality of conversation with conversational agents, this paper discusses agents' ability to recognize situational information and exploit the information to get the user involved in the conversation.

As a similar and more well-defined problem, this paper focuses on "engagement" in conversation: the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake (Sidner et al., 2003). Grounding is a device contributing to engagement. It is a process of ensuring that what has been said is understood and shared between the conversational participants (Clark, 1996). According to (Clark, 1996; Clark & Schaefer, 1989), attention to the speaker is the most basic form of signaling understanding by the addressee. In addition, analyzing direction giving dialogue in which a map is shared between conversational participants, (Nakano et al., 2003) found that attention to a shared reference serves as evidence of understanding when it co-occurs with the speaker's direction giving utterances with the map manipulation.

Therefore, not only communicative verbal/nonverbal behaviors directed towards a conversational partner, but also some kinds of nonverbal behaviors directing towards a perceived world would be useful in establishing engagement. To make conversational agents capable of engagement in a situated conversation, the agents should be able to be aware of the user's attention in the environment, and establish a communication channel by connecting the user's attention with the linguistic context of the conversation.

However, previous research on conversational agents has mainly focused on the relationship between a user and a system, and little has been studied about the agents' ability to link verbal messages to a perceived world. With the goal of improving agents' engagement capability, this paper discusses two types of engagement behaviors: engagement cues to a conversational partner, and those to an environment. We also discuss how to implement conversational agents that can recognize and generate these two types of engagement behaviors.

In the following sections, first, we describe nonverbal cues for engagement observed in face-to-face, and human-artifact communication. Then, we discuss how to utilize nonverbal cues and situational context in order to improve engagement capability of conversational agents. Based on the discussion,

we propose a design of conversational agents capable of recognizing/displaying engagement cues to the conversational partner as well as to the environment. Finally future directions are discussed.

2 Aspects of engagement

In our review of conversational engagement, we found two types of engagement behaviors: engagement cues directed to a conversational partner, and those to a conversational environment. The first type has been discussed in a number of studies claiming that nonverbal behaviors are used to maintain an interaction by signifying that interlocutors have access to each other's communicative actions, such as eye gaze and head nod (Argyle & Cook, 1976; Clark, 1996; Clark & Schaefer, 1989; Duncan, 1972, 1974; Kendon, 1967; Rosenfeld & Hancks, 1980).

Note that not only nonverbal signals directed to conversational partners, but also those to a shared environment serve as a signal of engagement. Whittaker (Whittaker, 2003; Whittaker & O'Conaill, 1993) claimed that sharing the same physical environment is important when tasks require complex reference to, and joint manipulations of, physical objects. In a shared environment, speakers and listeners can achieve joint attention to an object or event introduced in a conversation. These studies suggest that attention to a shared reference indicates that the interlocutor is engaged in a task as well as a conversation.

In the following subsections, we will review studies in these two types of engagement cues with respect to face-to-face communication as well as human-artifact communication.

2.1 Engagement cues in face-to-face conversation

2.1.1 Engagement cues to conversational partner

In his precise analysis of human greetings, Kendon (1967) described a sequence of nonverbal cues in greeting interactions. For example, before any greeting can begin, conversational participants must sight each other, and identify the other as someone they wish to greet. Then, they change the orientation and begin to approach the other. Once they start approaching, both/either of them may avert their gaze because keeping looking at directly may be a threat for the partner. Before starting a close salutation (and after approaching), they smile, hold a head position, and sometimes perform a "palm presentation gesture" that the palm of the hand is oriented toward the other.

According to (Argyle & Cook, 1976), eye gaze plays an important role in engagement in a conversation. Speakers look up at grammatical pauses to obtain feedback from the listener. Listeners, on the other hand, look at speakers to display that they are attending to the speaker, and to see the speaker's facial expressions and directions of eye gaze. Goodwin (1981) reported that speakers need the listener's gaze to start speaking, suggesting that eye gaze is an important engagement signal in conversation. He also gave interesting examples showing how a speaker directs a listener's attention using gestures. He claimed that gesture allows the speaker to redirect the listener's attention without disturbing the conversation (e.g., saying, "Look!") (Goodwin, 1986).

Head nods have a similar function to verbal acknowledgements such as "uh huh", "I see". A number of studies of face-to-face communication have mentioned that listeners return feedbacks as to whether conversation is on the right track, by giving visual evidence in the form of head nods and attention (Argyle & Cook, 1976; Duncan, 1974; Kendon, 1967).

(Duncan, 1974) observed nonverbal signals exchanged in turn-taking. In releasing a turn, the current speaker stops gesticulation, and drops in a paralinguistic pitch and/or intensity. In taking a turn, the next speaker shifts head direction away from the partner to the environment, and starts gesticulation.

2.1.2 Engagement cues to conversation environment

While engagement cues to the conversational partner are effective in establishing a communication channel with the partner, engagement cues directed to an environment are useful to connect linguistic messages with a physical perceived situation. Clark (2003) proposed "Directing-to" and "Placing-for" as conversational device to establish joint attention by indicating focused objects and events in a situated conversation.

Directing-to: Speaker's signal that **directs** addressee's attention **to** object *o*.
Placing-for: Speaker's signal that **places** object *o* **for** addressee's attention.

As directing-to behaviors, he listed up various kinds of non-verbal behaviors using different body parts as shown in Table 1. Pointing is a typical form of directing-to, but there are many other types of directing-to acts. For example, gazing at an object has a directing-to function. More importantly, directing-to acts are combined with verbal behaviors to redirect the addressee's attention. Demonstrative pronouns, such as "this", "that", "these", and

Table 1: Examples of directing-to behaviors

Instrument	Index i	Example
Finger	Pointing at *o*	That is the book I want.
Arm	Sweeping *o*	All this is yours.
Head	Nodding at *o*	She was standing there
Finger	Tapping on *o*	This is the book I want
Foot	Tapping on *o*	[of carpet samples] I like this best

"those", most frequently co-occur with directing-to nonverbal behaviors. For example; saying, "Look at that beautiful flower!" while pointing at a flower is a combination of verbal and nonverbal directing-to behaviors.

Placing-for is different from directing-to in a point that it works by moving an object into the addressee's attention. Most of the placing-for acts are concerned with manipulating objects. For example, customers at a drug store place soaps and shampoos on the counter to make the clerk identify these items as what they want to buy (Clark, 2003). More intriguingly, the same device is used in grounding what speakers are doing. If the clerk moved the items to a new area on the counter to ring them up, that behavior signals that the clerk understands the customer's intention to buy these items. Dillenbourg et al. (1996) reported similar phenomena as "multimodal grounding" in a computer mediated communication.

Directing-to and placing-for are used in indicating the focused objects in the conversation. Both of these techniques are used to connect a message with a perceived world described in the message, and allow the addressee to access and perceive the message.

2.2 Engagement cues in human-artifact communication

Research on Embodied Conversational Agents focuses on communication capability of animated agents, and has a goal of implementing agents that can generate non-verbal behaviors such as head nod, gaze towards user and away, and gestures (Cassell et al., 2001). The goal of this approach is to improve naturalness of human-computer interaction by implementing face-to-face conversational protocols, which are mainly concerned with engagement cues to a user, into animated agents.

In contrast, Steve (Rickel & Johnson, 2000) cohabits 3D virtual worlds with people and other agents, so it addressed immersive aspects of dialogue in virtual worlds. This project has been extended to the Mission Rehearsal Experience (MRE) project (Swartout et al., 2001), and a multimodal dialogue model was proposed to control a multiparty conversation in immersive virtual world (Traum & Rickel, 2002). This model consists of multiple layers, such as contact, attention, and conversation. On the basis of this model, agents' engagement behaviors can be generated.

Aiming at improving agents' ability to interpret user's engagement cues directed to an environment, (Nakano et al., 2003) proposed nonverbal grounding process model between a user and an agent. Analyzing eye gaze, head nods and attentional focus in the context of a direction-giving task in face-to-face conversation, they investigated joint attention with respect to grounding, which is a kind of engagement behaviors. They found that listener's attention to a joint shared reference serves as evidence of understanding in the grounding process. Based on the results, they implemented an embodied conversational agent relying on both verbal and nonverbal signals to establish common ground in human-computer interaction.

Research in communication robots has discussed similar issues in terms of social intelligence in robots (Fong et al., 2003). Dautenhahn et al., (2002) proposed that socially embedded agents are required not only to be embodied in an environment in the sense that mutual perturbation relationship is established between agent and environment, but also to be embedded in a collection of other agents and at least partially aware of the structures of interactions in a given social domain. They also claimed that an important ability of socially intelligent robots is to track, identify, and interpret visual interactive behaviors.

Based on analyses of face-to-face conversations, (Sidner et al., 2003) proposed a humanoid robot designed to mimic human conversational gaze behavior in collaboration conversation. They analyzed usage of eye gaze in human face-to-face conversation, and proposed some engagement rules to maintain a conversation with a human user. Moreover, (Sidner et al., 2003) proposed communicative capabilities required for collaborative robots: (a) Conversation management, including abilities of turn

taking, interpreting the intentions of conversational participants, and updating the state of the conversation; (b) Collaboration behavior to accomplish the goal for the conversation (e.g., determining what the agent should say and do); (c) Engagement behaviors consisting of initiating, maintaining the interaction, and disengaging from the interaction.

3 Immersive conversational environment

Based on the discussion about engagement behaviors in the previous section, we implemented a conversational agent that recognizes user's gaze direction as a nonverbal engagement cue, and responds to it in form of agent engagement behaviors. The conversational agent displays engagement cues to the user, such as greetings and turn-taking nonverbal signals, as well as those towards the environment, such as directing-to behaviors and joint attention.

As a platform system, we are developing a conversational environment, IPOC, using a panoramic picture as a background. Figure 1 shows a snapshot of IPOC. In IPOC, a conversational agent talks about objects and events shown in the background.

Figure 1: IPOC agent snapshot

3.1 Overview of the Interaction Control Component

In this section, we describe the Interaction Control Component (ICC) in IPOC system. The ICC architecture is shown in Figure 2. The ICC interprets inputs from a speech recognizer and a head tracking

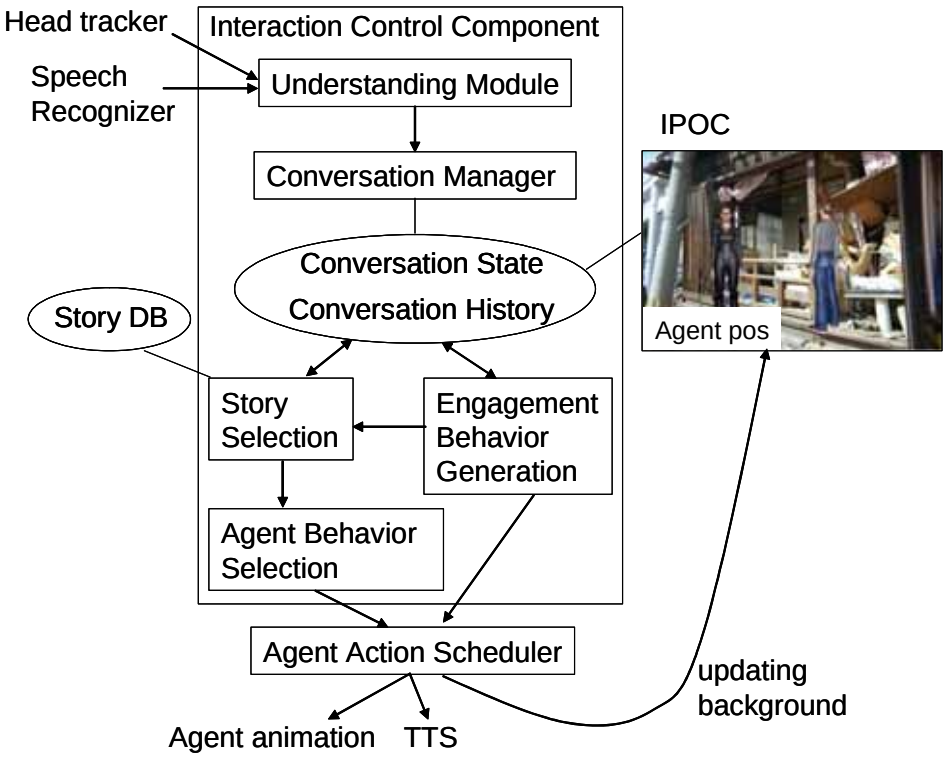


Figure 2: Architecture of the Interaction Control Component

system, and selects an appropriate story according to the conversational and perceptual situation. It decides verbal and nonverbal behaviors to be performed by conversational agents. The output of the ICC is a narrative text on which nonverbal actions are annotated. Then, the text is input to the Agent Action Scheduler, where a time schedule for agent animations is calculated by accessing a text-to-speech engine.

Finally, according to the time schedule, agent animations and speech sound are produced in a synchronized way. Agent animations are produced by Haptek animation system and speech sound is synthesized by HITACHI HitVoice.

3.2 Perceiving user's actions

Input devices: Available input devices are a speech recognizer and a head tracking system (Sato et al., 2004). The head tracking system calculates user's head position and rotation every one 30th of a second, and the average of data in a 0.3-second window is used as a recognition result. Using the head pose recognition results, the system estimates user's gaze direction on a display. In the current implementation, the whole display is split into six areas, and the system judges which of there areas the user's gaze falls in. Although this method does not recognize accurate gaze direction, we think it is still very useful for perceiving user's attention without using a motion capture.

Understanding Module (UM): The UM interprets the user's verbal and nonverbal input. As a verbal input, the UM understands a limited number of sentences recognized by a speech recognition system. For example, the UM understands user's request of telling a story about a specific object, like "tell me about the collapsed house." The user's gaze direction recognized by a head tracking system is interpreted as user's directing-to behavior. By adding simple annotation to the background, the system can identify objects that the user is looking at (directing-to) on the background.

Conversation Manager (CM): The CM maintains the history and the current state of the conversation. The conversation history is a list of story IDs that are already told. The conversation state is represented as a set of values including current story ID, agent current position and rotation, and user's gaze direction. The CM updates the conversation state when the system receives user's input (e.g., change of user's focus of attention), or generates agent actions such as telling a story and confirming user's intention.

3.3 Presenting a story with nonverbal behaviors

Story Selection Module (SSM): When a keyword or a sentence is input to the SSM, it retrieves short paragraphs from the story DB using a text searching algorithm (Kiyota et al., 2002), and the agent presents them as a story. The keywords are determined by the CM, which looks up the current conversation state.

Agent Behavior Selection Module (ABS): The ABS generates agent animations for presenting a story to the user. Agent gestures emphasizing important concepts in a story are automatically determined and scheduled using an agent behavior generation system CAST (Nakano et al., 2004), which selects agent gestures based on linguistic information in a Japanese text.

3.4 Generating agents' engagement behaviors

The Engagement Behavior Generation Module (EBG) determines engagement behaviors according to the conversation state including user's attentional information. As this study focuses on nonverbal engagement behaviors, the current system does not much consider verbal engagement behaviors.

First, the system classifies user's eye gaze into four types: (a) attention to the agent, (b) to a joint reference, (c) to other objects in the background, and (d) disengaging. When the user's attentional state is in (a)-(c), the system assumes that the user is engaged in the conversation with the agent. Agent's engagement behaviors in each case are determined as follows.

(a) Attention to the agent

(a-1) When the user looks at the agent before starting an interaction with the agent, the system interprets the user's gaze as an engagement behavior signaling that the user would like to start a conversation with the agent. In this case, the EBS generates verbal and nonverbal greeting behaviors to start a conversation. In addition, we implemented some kinds of conversation initialization cues (Kendon, 1967) which have been mentioned in Section 2.1.

(a-2) When the user pays attention to the agent during a story, the system judges that the user is engaged in the agent's narrative. To maintain this relationship, the EBS generates agent's glancing at the user, and demonstrates that the agent checks the user's status. In addition, the system confirms user's understanding if it necessary. Nakano et al.,

(2004) reported that listener's keeping staring at the speaker is a signal of not-understanding, and in such a case, the speaker needs to give additional explanation to the listener about ungrounded matters.

(a-3) When the user keeps gazing at the agent after a story is ended, the EBS requests the SSM to start a new story which is related to the previous one. Then, the SSM sends back information about a new focused object to the EBS. By using this information, the EBS generates directing-to behaviors in order to redirect the user's attention to the new focused object smoothly. Figure 3 shows an example of directing-to behavior by the agent. In this example, the agent redirects the user's attention to the roof of the house by directing her head and upper torso towards the house and pointing at it at the beginning of a story about a collapsed house.

Figure 3: Agent directing-to behavior

(b) Attention to a joint reference
(b-1) In IPOC, the user's attention to a reference can be possible while the agent telling a story. In this case, the EBS generates agent's looking at the focused object to establish joint attention with the user.

(c) Attention to other objects
(c-1) If the user looks at an object which is not the current focused object, the EBS generates gestures more aggressively to draw back user's attention to the agent.

(c-2) After finishing a story, the user has a chance to choose a new story by looking at an object in the background. If the CM finds that the user is interested in an object that is different from the previous focused object, the EBS requests the SSM to select a story related to the object. When a story is chosen, it is presented by the agent while the EBS generates agent's gaze at the new focused object to establish joint attention. From user's point of view, it looks as if the agent follows the user's directing-to behavior, and tells a story about her/his focused area.

(d) Disengaging
(d-1) If the user completely looks away from the display, the system assumes that the user is disengaged from the interaction with the agent. So, the system asks the user whether s/he wants to stop the conversation and generates farewell greetings.

Functions of these agent's engagement behaviors can be classified into four types: initiation of a conversation, maintaining a conversation, transition of topic, and disengaging from an interaction. The mapping between the functions and types of engagement behaviors is shown in Table 2.

Table 2: Functions of engagement behaviors performed by an agent

Function	Type
Initiation	(a-1)
Maintaining	(a-2), (b-1), (c-1)
Transition	(a-3), (c-2)
Disengagement	(d-1)

5 Conclusion and Future Work

This paper proposed a conversational agent capable of recognizing/displaying nonverbal engagement behaviors. First, we reviewed previous studies addressing engagement behaviors in face-to-face interaction and those in human-artifact interaction. Based on the discussion, we have proposed a design of a conversational agent having engagement capability. We implemented a prototype system of a story-based immersive communication environment, IPOC. In IPOC, conversational agents tell stories about a situation shown in a background picture, and interact with users while recognizing/displaying nonverbal engagement cues.

We admit that a lot of works need to be done to improve the current mock-up system. First, it is definitely necessary to improve the engagement behavior recognition algorithm because the current system only employs a simple gaze estimation using head pose information. It would be useful to recognize head nod, which is another very useful nonverbal engagement cue.

Second, although we focused on nonverbal engagement behaviors, engagement itself can be accomplished verbally as well as nonverbally. Research on verbal engagement and the combination of verbal and nonverbal engagement behaviors would be an important future work.

Moreover, as a study of human interface design, evaluating effectiveness of agents' engagement capability in human-computer interaction will be indispensable. It would be necessary to examine whether the implemented behaviors actually look human-like, and contribute to better engagement and natural interaction.

Finally, while this paper has mainly focused on conversational virtual agents, it would be interesting to apply agents' engagement functions to robots. As robots inhabit a physical world and can change the world by manipulating objects, capability of engagement may be effective to improve robot's communication capability.

Acknowledgements

We would like to express special thanks to Prof. Sato and Mr. Oka, who kindly provided us their head tracking technology for this study. Thanks to Masashi Okamoto, Kazunori Okamoto, and Toshihiro Murayama in helping me conduct this study. This study was supported by Mission-oriented Program I in RISTEX, which was established under the joint auspices of the Japan Science and Technology Corporation (JST) and the Japan Atomic Energy Research Institute (JAERI).

References

Argyle, M., & Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge: Cambridge University Press.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2001). More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. *Knowledge-Based Systems, 14 (2001)*, pp. 55-64.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H. (2003). Pointing and Placing. In Kita, S. (Ed.), *Pointing. Where language, culture, and cognition meet*. NJ: Hillsdale NJ: Erlbaum.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*, pp. 259-294.

Dautenhahn, K., Ogden, B., & Quick, T. (2002). From Embodied to Socially Embedded Agents - Implications for Interaction-Aware Robots. *Cognitive Systems Research 3(3), Special issue on Situated and Embodied Cognition*, pp. 397-428.

Dillenbourg, P., Traum, D. R., & Schneider, D. (1996) *Grounding in Multi-modal Task-Oriented Collaboration*. In *Proceedings of EuroAI&Education Conference*.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), pp. 283-292.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society, 3*, pp. 161-180.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A Survey of Socially Interactive Robots,. *Robotics and Autonomous Systems, 42*((3-4)), pp. 143-166.

Goodwin, C. (1981). Achieving Mutual Orientation at Turn Beginning, *Conversational Organization: Interaction between speakers and hearers* (pp. 55-89). New York: Academic Press.

Goodwin, C. (1986). Gestures as a resource for the organization of mutual orientation. *Semiotica, 62*(1/2), pp. 29-49.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26*, pp. 1-47.

Kiyota, Y., Kurohashi, S., & Kido, F. (2002) *"Dialog Navigator" : A Questions Answering System based on Large Text Knowledge Base*. In *Proceedings of The 19th International Conference on Computational Linguistics (COLING 2002)*, (Taipei), pp. 460-466.

Nakano, Y. I., Okamoto, M., Kawahara, D., Li, Q., & Nishida, T. (2004) *Converting Text into Agent Animations: Assigning Gestures to Text*. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004), Companion Volume*, (Boston), pp. 153-156.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003) *Towards a Model of Face-to-Face Grounding*. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL03)*, (Sapporo, Japan), pp. 553-561.

Rickel, J., & Johnson, W. L. (2000). Task-Oriented Collaboration with Embodied Agents in Virtual Worlds. In Cassell, J. & Sullivan, J. & Prevost, S. & Churchill, E. (Eds.), *Embodied Conversational Agents* (pp. 95-122). Cambridge, MA: MIT Press.

Rosenfeld, H. M., & Hancks, M. (1980). The Nonverbal Context of Verbal Listener Responses. In Key, M. R. (Ed.), *The Relationship of Verbal and Nonverbal Communication* (pp. 194-206). New York: Mouton Publishers.

Sato, Y., Oka, K., Nakanishi, Y., & H. Koike. (2004) *Video-based tracking of user's motion and its use for augmented desk interface*. In *Proceedings of IEEE Int'l Conf. Automatic*

Face and Gesture Recognition (FG 2004), pp. 805-809.

Sidner, C. L., Lee, C., & Lesh, N. (2003) *Engagement Rules for Human-Robot Collaborative Interactions*. In *Proceedings of IEEE International Conference on Systems, Man & Cybernetics (CSMC), Vol. 4*, pp. 3957-3962.

Swartout, W., Hill, R., Gratch, J., Johnson, W. L., Kyriakakis, C., Labore, K., Lindheim, R., Marsella, S., Miraglia, D., Moore, B., Morie, J., Rickel, J., Thiebaux, M., Tuch, L., Whitney, R., & Douglas, J. (2001) *Toward the holodeck: Integrating graphics, sound, character and story*. In *Proceedings of 5th International Conference on Autonomous Agents*.

Traum, D., & Rickel, J. (2002) *Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds*. In *Proceedings of the first International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pp. 766-773.

Whittaker, S. (2003). Theories and Methods in Mediated Communication. In Graesser, A. (Ed.), *The Handbook of Discourse Processes*: MIT Press.

Whittaker, S., & O'Conaill, B. (1993) E*valuating videoconferencing.* In P*roceedings of* C*ompanion Proceedings of CHI Human Factors in Computing Systems,* AC**M Press.**

Eye Movement as an Indicator of Users' Involvement with Embodied Interfaces at the Low Level

Chunling Ma[⋆]
[⋆]Dept. of Information and Communication Eng.
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
macl@miv.t.u-tokyo.ac.jp

Helmut Prendinger[†]
[†]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
helmut@nii.ac.jp

Mitsuru Ishizuka[⋆]
[‡]Dept. of Information and Communication Eng.
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
ishizuka@miv.t.u-tokyo.ac.jp

Abstract

In this paper, we motivate an approach to evaluating the utility of animated interface agents that is based on human eye movements rather than questionnaires. An eye tracker is employed to obtain quantitative evidence of a user's focus of attention. The salient feature of our evaluation strategy is that it allows us to measure important properties of the user's interaction experience on a moment-by-moment basis. We describe an empirical study in which we compare attending behavior of participants watching the presentation of an apartment by three types of media: an animated agent, a text box, and speech only. Users' eye movements may also shed light on their *involvement* in following a presentation.

1 Introduction

Animated interface agents have attracted considerable interest and attention in recent years, mainly for their ability to emulate human–human communication styles that is expected to improve the intuitiveness and effectiveness of user interfaces (see, e.g. André et al. (1996) for early work in this area). Following this user interface paradigm, a considerable number of animated agent (or character) based systems have been developed, ranging from information presentation and online sales to personal assistance, entertainment, and tutoring (Cassell et al., 2000; Prendinger and Ishizuka, 2004b). While significant progress has been made in individual aspects of animated agents, such as their graphical appearance or quality of synthetic voice, evidence of their positive impact on human–computer interaction is still rare. The most well-known evaluation studies have been directed towards showing the 'persona effect', stating that animated agents can have a positive effect on the dimensions of motivation, entertainment, and perceived task difficulty (Lester et al., 1997; van Mulken et al., 1998).

A common feature of most evaluations of interface agents is that they are based on questionnaires and focus on the user's experience with the systems hosting them, including questions about their believability, likeability, engagingness, utility, and ability to attract attention. However, as Dehn and van Mulken (2000) pointed out, the broad variety of realizations of animated agents and interaction scenarios complicates their comparison. More importantly, subtle aspects of the interaction, such as whether users pay attention to the agent or not, cannot be deduced reliably from self-reports (Nisbett and Wilson, 1977).

Furthermore, the concept of a user's *involvement* when interacting with computers has recently been discussed in the areas of Social Intelligence Design and Conversational Informatics (Nishida, 2001; Prendinger and Ishizuka, 2004a). While the term "involvement" embraces a wide range of concepts, including immersion and engagement (over a possibly extended time span), we want to consider involvement *at the low level* – whether the user is attending to designated objects of the interface – which seems to be an important prerequisite for any 'higher level' notion of involvement.

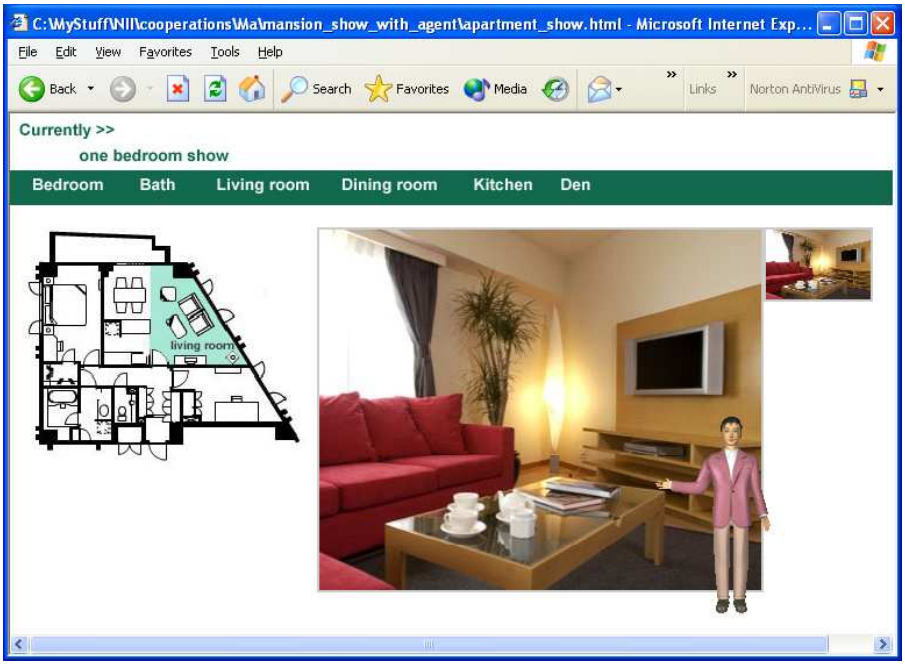


Figure 1: An animated agent presents the living room.

In this paper, we want to propose a different approach to evaluating animated agents, one that is based on eye movement behavior of users interacting with the interface. Although gaze point and focus of attention are not necessarily always identical, a user's eye movement data provide rich evidence of the user's visual and (overt) attentional processes (Duchowski, 2003). The movements of the human eye can be used to answer questions such as:

- Is the user paying attention to the interface agent?
- To which part of the agent (face or body) is the user attending to?
- Can the agent's verbal or gestural behavior direct the user's focus of attention?

Hence, eye movement data can offer valuable information relevant to the utility of animated agents and the usability of interfaces employing those agents. The tracking of eye movements lends itself to reliably capturing the moment-to-moment experience of interface users, which is hard to assess by using post-experiment questionnaires.

We will track and analyze eye movements while users are following the web page based presentation of different rooms of an apartment. Three types of presentations will be contrasted: (i) An animated interface agent presents the apartment using speech and gestures (see Fig. 1), (ii) the apartment is presented by means of a text-box and read out by speech, or (iii) the presentation is given by speech only.

The rest of the paper is organized as follows. The next section overviews work related to using eye movement as an evaluation method for user interfaces. The core part of the paper (Sect. 3) is devoted to the description of an experiment that provides both spatial and temporal analyses of users' eye movements during a presentation. The paper is rounded off by conclusions.

2 Related Work

This section reports on work that employs eye movements in the context of user interfaces. Eye movement data have been analyzed for two purposes, *diagnostic* and *interactive*. In the diagnostic use, eye movement data provide evidence of the user's attention and can be investigated to evaluate the usability of interfaces (Faraday and Sutcliffe, 1996; Goldberg and Kotval, 1999). In the interactive use, a system responds to the observed eye movements and can thus be seen as an input modality (Duchowski, 2003).

Goldberg and Kotval (1999) performed an analysis of eye movements in order to assess the usability of an interface for a simple drawing tool. Comparing a 'good' interface with well-organized tool buttons to a 'poor' interface with a randomly organized set of tool

buttons, the authors could show that the good interface resulted in shorter scan paths that cover smaller areas. The measure of interest in their study is efficient scanning behavior, i.e. a short scan path to the target object. While this measure might not have high priority in our application domain, the merit of this study is to have introduced a systematic classification of different measures based on (temporal) scan paths rather than on cumulative (spatial) fixation areas. The temporal succession of transitions between different areas of attention is particularly relevant to investigate the effect of deictic references of animated agents to interface objects.

Faraday and Sutcliffe (1996) investigated attentional processing and comprehension of multimedia presentations. Core findings of the authors relevant to our domain can be summarized along the following dimensions:

Shifts of attention.

- A moving interface object induces a shift of attention to the object in motion.
- Attention is re-oriented when the presentation scene shifts.
- Labelling a presentation object produces fixation shifts between the object and the label.

Locked attention. A viewer's attention is locked when a moving object is processed, so that other presentation objects which are concurrently changed are not attended to.

Auditory language processing and attention. Comprehension of objects being presented visually with a spoken comment is increased only if both media types produce a single unified proposition.

The last mentioned item has also been investigated by Cooper (1974) who reports that people who simultaneously listen to speech and a visual object featuring elements that are semantically related to the spoken information tend to focus on the elements that are most closely related to the meaning of the currently heard spoken language (see also Duchowski (2003, p. 167)).

Witkowski et al. (2001) employ eye-tracking technology in order to assess user attention while interacting with an animated interface agent based online sales kiosk. In this setting, the interface agent provides help to the user and presents a product (a selection of wines). The authors conjecture that the agent will direct the attention of the users to the item of interest (help buttons, pictures of wines), following the agent's verbal comments. However, the results of their study do not support this hypothesis. In the experiment, a character agent controlled by the Microsoft Agent package (Microsoft, 1998) has been chosen with the text balloon enabled that depicts the text that is currently being spoken. The results reveal that users mostly focus on reading the text, rather than attending to the agent or to the product. In our study, we thus decided to disable the text balloon in order to avoid this problem. For the time that users were looking at the agent (face, gesture, body), the face was focussed on the most. In general, Witkowski et al. (2001) observed that interface agents do attract the attention of users. Similar results have been obtained by Takeuchi and Naito (1995) who compared an interface featuring either a (facial) agent or an arrow.

Besides its diagnostic role, eye movement data have also been used as an additional input modality to character-based intelligent systems. For instance, Qu et al. (2004) consider a user's focus of attention (among others) to decide an appropriate response in the context of educational software, and Nakano et al. (2003) investigate attentional focus (among others) in the setting of a direction-giving task.

3 Method

3.1 Experimental Design

A presentation of an apartment located in Tokyo has been prepared using a web page based interface (Mansions, 2004). The apartment consists of six rooms: living room, bedroom, dining room, den, kitchen, and bathroom. Views of each room are shown during the presentation, including pictures of some part of the room and close-up pictures of e.g. a door handle or sofa. Three versions of the apartment show have been designed for the experiment:

- *Agent (& speech) version.* A character called "Kosaku" presents the apartment using synthetic speech and deictic gestures (see Fig. 1). The character is controlled by a version of MPML (Prendinger et al., 2004).
- *Text (& speech) version.* The presentation content of each scene is displayed by a text box and read out by Microsoft Reader.
- *Voice (only) version.* Synthetic speech is the only medium used to comment on the apartment.

The main purpose of programming the Text and Voice versions was to provide interfaces that represent conceivable presentation types and can be compared to the Agent version in terms of the user's eye movements. The same type and speed of (synthetic) voice was used in all versions.

3.2 Subjects

Fifteen subjects (3 female, 12 male), all students or staff from the University of Tokyo, participated in the study (5 in each version). Their age ranged from 24 to 33 (mean 28.75 years). They were recruited through flyers and received 1,000 Yen for participation. In some cases the calibration process of the eye tracker was not successful due to reflections of contact lenses. Those subjects were excluded from the experiment beforehand.

3.3 Apparatus

The presentation of the apartment was hosted on a computer with a 17 inch (42.5 cm) monitor (the main monitor). A second computer was used to control the eye tracking system, a NAC Image Technology Eye-mark Recorder model EMR-8B (NAC, 2004). The eye mark recorder is shown in Fig. 2 and the experimental setup is shown in Fig. 3.

Figure 2: NAC EMR-8B eye tracker.

The EMR eye tracker uses two cameras directed toward the subject's left and right eye, respectively, to detect their movements by simultaneously measuring the center of the pupil and the position of the reflection image of the IR LED on the cornea. A third camera is faced outwards, in the direction of the subject's visual field, including the main monitor. The system has a sampling rate of 60 Hz. The subject's head posture was maintained with a chin rest, with the eyes at a distance of 24 inch (60 cm) from the main monitor. A digital video recorder that captured the data from the third camera was connected to the computer that processed the eye movements.

The subjects were also connected to a bio-signal encoder that provides skin conductance (SC) and heart rate (HR) sensors. The bio-signal part of the experiment will not be reported here.

3.4 Procedure

Subjects were first briefed about the experiment. They were told that an apartment will be shown to them, and that they would be asked general questions about the apartment afterwards. They were also instructed to watch the demonstration carefully since they should be able to report features of the apartment to others.

The subjects were first put on the cap with the eye tracker. Calibration was performed by instructing the subject to fixate nine points in the screen area. After that, the subjects were shown the presentation that lasted for 8 minutes. Finally, the subjects were freed from the tracking equipment, and asked to fill out a questionnaire in order to report on their perception of the interface and to answer some content-related questions concerning the presented material.

3.5 Data Analysis

For analysis, the recorded video data of a presentation were first divided into individual scenes. A scene is a presentation unit where a referring entity (agent, text box, or voice) describes a reference object (an item of the room). Only the Agent and Text versions feature a visible referring entity. In Fig. 1, the scene consists of the agent performing a hand gesture to its right and introducing the living room. In order to be able to compare the three versions, scenes where the agent or text box moves from one location were left out. For each scene (41 in total), the following four screen area categories were defined:

- The area of a (visible) referring entity is either the smallest rectangle demarcating the agent or the text box (the agent area is further subdivided into face and body areas).
- The area of the reference object is the smallest rectangle demarcating the object currently described.
- The layout area (a designated, permanent reference object) is the field on the screen that displays the layout of the room.
- Other screen areas.

A program has been written that first maps eye-tracking data to xy-coordinates of the video sequence, and then counts the gaze points in each of the four categories.

When eye movements are relatively steady for a short period in one area, they are called *fixations* whereas rapid shifts from one area to another are called *saccades* (Duchowski, 2003). During a saccade, no visual processing takes place. If a cluster of

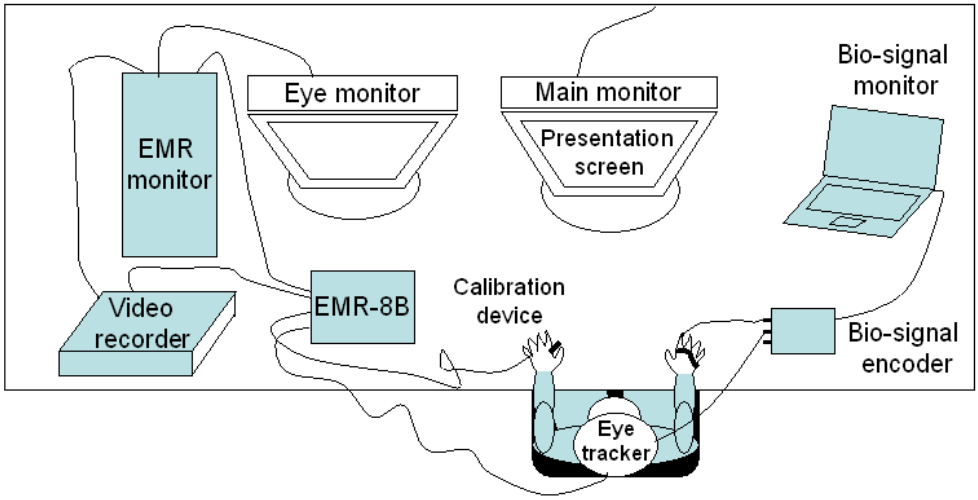


Figure 3: Experimental setup.

gaze points has less than 6 entries, it was categorized as part of a saccade (Goldberg and Kotval, 1999). All data accounted for in the analysis are derived from the activity of subjects' left eyes.

3.6 Results

The core of our results was distilled from analyzing subjects' eye movements. The level of statistical significance was set to 5%.

Focus of Attention Hypothesis. The ability of the interface to direct a subject's focus of attention to reference objects has been tested in two ways, spatial and spatio-temporal. The *spatial* analysis counts the gaze points that fall within areas of interest, specifically the reference object area and the layout area. Except for the introductory episode, the layout is not explicitly referred to during the presentation although it may serve as an orientation aid for users. The hypothesis is tested by restriction to those scenes where the referring entity (agent, text, voice) refers to some item of the apartment. An between–subjects analysis of variance (ANOVA) showed that users focus on the reference objects more in the Voice version than in either of the Agent or the Text version ($F(2,9) = 8.2$; $p = 0.009$). The mean values are indicated in Fig. 4. The result for the map area, while not statistically significant, shows a tendency toward a similar distribution of gaze points ($F(2,9) = 2.8$; $p = 0.11$). (For a comparison between gaze points in the agent and text box areas, see the Locked Attention Hypothesis below.)

Those results suggest that gaze points are not randomly distributed across the screen area but depend

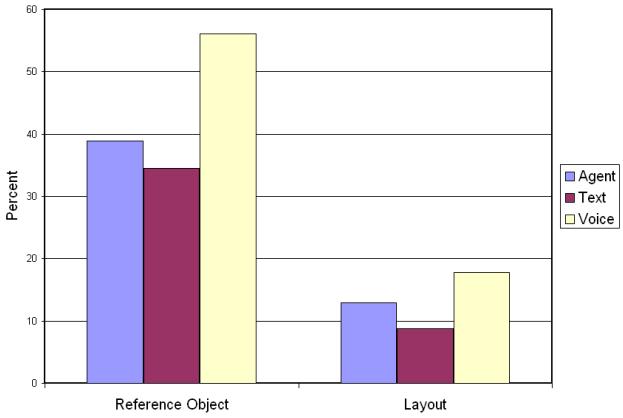


Figure 4: Impact of Agent vs. Text vs. Voice version on gaze points in reference object area and layout area.

on the presence or absence of a visible presentation medium. When an agent or a text box is present, users' attentional focus is more evenly shared between the presentation medium and the presented material.

Locked Attention Hypothesis. This hypothesis compares the portions that subjects focus on the agent (face or body) or the text box, which reveals text line by line. The mean for the agent is 18% of the total number of gaze points, and the mean for the text box is 32%. The t-test (one-tailed, assuming unequal variances) showed that subjects look significantly more often at the text box ($t(6) = -2.47$; $p = 0.03$).

This result can be seen as evidence that users spend considerable time for processing an object that gradually reveals new information. Locked attention can

Figure 5: "To your left is the layout of the apartment. As you can see, the apartment includes: bedroom, living room, dining room, den, kitchen and bathroom."

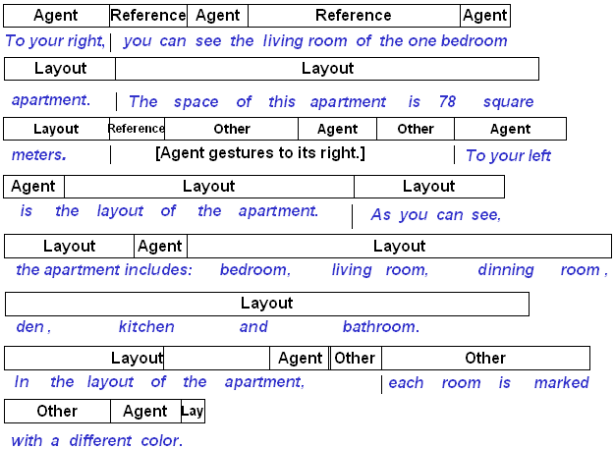


Figure 6: Example of attentional shifts in the introductory episode of the presentation.

prevent users from attending to other salient information (Faraday and Sutcliffe, 1996).

Shift of Attention Hypothesis. While a spatial analysis can indicate where attention is spent, it cannot reveal the nature of how users traverse the interface when watching a presentation. In order to address those more complex aspects of intelligent interfaces, we performed a (preliminary) *spatio-temporal* analysis of eye movement data. Figure 5 depicts a screen shot of the original view (taken by the outward directed camera of the EMR-8B system) of a subject in the Agent version. The dark colored dots are gaze points drawn by our program. The numbers have been added to the screen shot by hand. The frames around the agent (face, body) and the layout have been re-drawn for clarity. When the agent speaks the sentence in Fig. 5, the subject's focus of attention is first on the agent's face, next on the layout area, then it traverses back to the agent's face, and finally shifts to the layout area.

A more detailed description of one subject's attentional shifts is shown in Fig. 6. The rectangles above the sentences of the introductory episode of the apartment presentation indicate the focus of the subject's attention. The surface structure of the sentences is synchronized with attentional focus. Observe that the subject initially shifts attention between the agent and the living room (the reference object), and when the agent says "The space of this apartment is 78 square meters", the subject focuses on the layout that depicts the size of the apartment. In the following, the subject partly attends to the agent's gesture, and after some occasional shifts to other areas, fixates on the layout. When the agent explains how the rooms are marked, the subject is apparently not attending to the layout during the utterance of the sentence.

The attentional shifts in the example of Fig. 6 suggest that users can perceive animated agents to possess a certain degree of competence, such as directing the user to locations of interest. Even more importantly, it demonstrates how a user re-directs attentional focus back to the agent after being directed to a reference object, which supports the interpretation of users expecting agents to provide them conversational cues and other meaningful information.

As a first attempt to provide a systematic spatio-temporal analysis of eye movements for intelligent embodied interfaces, we propose a Instructor–Reference–Instructor (IRI) triple as a basic unit for evaluation. An IRI denotes a situation where the user first attends to an instructor, a referring entity like an agent or a text box, then focuses on a reference object, and afterwards shifts attention back to the instructor. IRIs appear to be important interaction patterns in conversation, including direction-giving tasks (Nakano et al., 2003), and strong indicators of the instructor being conceived of as a social actor.

As a preliminary evaluation, we compared the number of IRIs of the Agent and Text versions for the episode displayed in Fig. 6 (plus one sentence). Here, both the living room and the layout qualify as reference objects. Figure 6, e.g., has 4 IRIs. The t-test on the small sample was not significant ($t(5) = 1.75$; $p = 0.07$). The means are: Agent (4.34) and Text (2). While this outcome indicates a tendency, further analysis with more episodes is needed to support the hypothesis that animated agents trigger conversational behavior in users.

Agent Face–Body Hypothesis. This hypothesis has been tested by summarizing gaze points that are contained in either the agent face or the agent body region. It could be shown that subjects were looking mostly at the agent's face (mean = 83.1%; stdev = 6.8), which supports the hypothesis that users interact socially with interface agents.

Questionnaire. The questionnaire contained two types of questions, one focusing on the subjects' general impression of the presentation, the other on the subjects' ability to recall shown items. In the first set of questions, subjects were asked (i) whether they would want to live in the apartment, (ii) whether they would recommend the apartment to a friend, and (iii) whether they thought the presentation helped them in their decision to rent the apartment. A 5 point Likert scale was used, ranging from "1" (strongly agree) to "5" (strongly disagree). The intention of questions (i) and (ii) was to investigate the effect of the presentation type on the users' perception of the apartment, but there were no results of statistical significance. An ANOVA of the third question, however, showed that subjects judged the Voice version to be more helpful that either of the other versions ($F(2,12) = 8.9$; $p = 0.004$). The means are: Agent (2.2), Text (2.8), and Voice (1.2).

The second set of questions (eight in total) asked subjects for details of the presentation, such as "What could you see from the window in the living room?". Answers could be chosen from three options. The percentage of correct answers was 81.25% for the Agent version, 80% for the Text version, and 87.5% for the Voice version.

The results obtained from the questionnaire indicate that a presentation given by a disembodied voice is superior to an embodied agent or text together with underlying speech. This outcome supports the interpretation of agents carrying the risk of distracting users from the material being presented (van Mulken et al., 1998). On the other hand, agents might provide a more enjoyable experience to the user, but that dimension was not tested in the present study.

4 Conclusions

It is often argued that animated agents are endowed with *embodied intelligence* – they are able to employ human-like verbal and gestural behavior to behave naturally toward users (Cassell et al., 2000). However, little quantitative evidence exists that users also interact naturally with embodied agents in terms of involuntary indicators of interactivity such as attentional focus, which is an important prerequisite for their utility as virtual interaction partners. The same is true for the question to what extent users are involved in their interaction with embodied agents.

This paper has introduced a novel method for evaluating the interaction of users with animated interface agents, which is based on tracking users' eye movements. In terms of involvement in the interaction, this method allows us to evaluate whether users are *involved at the low level* and hence focus of the intended interface objects.

Primarily, it was demonstrated that the attentional focus hypothesized from gaze points constitutes a rich source of information about users' actual interaction behavior with computer interfaces. Both cumulative and temporal analyses of attentional focus revealed that users interact with animated agents in an essentially natural way. They follow the verbal and non-verbal navigational directives of the agent and mostly look at the agent's face. Unlike a textual interface (one revealing text line by line) that seems to capture users' attention to a high degree, users seem to attend to the visual appearance of the agent in a balanced way, with shifts to and from the object currently being presented. Although this result does not offer an interpretation as distinct as gaze behavior in grounding during face-to-face interaction (Nakano et al., 2003), it can provide valuable insights into the usability of the interface.

Besides an extended investigation of the obtained user gaze point data for spatio-temporal analysis, future work will also include the definition of comprehensive temporal measures of analysis for character-based interactive interfaces. A further interesting future direction is to track and analyze users' pupil dilating that has been shown as an index for confusion and surprise (Umemuro and Yamashita, 2003) and for affective interest (Hess, 1972).

5 Acknowledgements

We would like to thank Yukiko I. Nakano for her valuable suggestions on how to annotate the data, and Jin YingZi for helping with the analysis. This research was supported by the Research Grant (FY1999–FY2003) for the Future Program of the Japan Society for the Promotion of Science (JSPS).

References

Elisbeth André, Jochen Müller, and Thomas Rist. The PPP Persona: A multipurpose animated presenta-

tion agent. In *Proceedings Advanced Visual Interfaces (AVI-96)*, pages 245–247. ACM Press, 1996.

Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors. *Embodied Conversational Agents*. The MIT Press, Cambridge, MA, 2000.

Roger M. Cooper. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, (6):84–107, 1974.

Doris M. Dehn and Susanne van Mulken. The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies*, (52):1–22, 2000.

Andrew T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, London, UK, 2003.

Pete Faraday and Alistair Sutcliffe. An empirical study of attending and comprehending multimedia presentations. In *Proceedings of ACM Multimedia 96*, pages 265–275, Boston MA, 1996.

Joseph H. Goldberg and Xerxes P. Kotval. Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, 24:631–645, 1999.

Eckhard H. Hess. Pupillometrics: A method of studying mental, emotional and sensory processes. In N.S. Greenfield and R.A. Sternbach, editors, *Handbook of Psychophysiology*, pages 491–531. Holt, Rinehart & Winston, New York, 1972.

James C. Lester, Sharolyn A. Converse, Susan E. Kahler, S. Todd Barlow, Brian A. Stone, and Ravinder S. Bhogal. The Persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI-97*, pages 359–366. ACM Press, 1997.

Tokyo Mansions, 2004. URL: `http://www.moveandstay.com`.

Microsoft. *Developing for Microsoft Agent*. Microsoft Press, Redmond, WA, 1998.

NAC. Image Technology, 2004. URL: `http://eyemark.jp`.

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of Association for Computational Linguistics (ACL-03)*, pages 553–561, 2003.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977.

Toyoaki Nishida. Social intelligence design – an overview. In *New Frontiers in Artificial Intelligence. Joint JSAI 2001 Workshop Post-Proceedings*, pages 3–10. Springer, 2001.

Helmut Prendinger, Sylvain Descamps, and Mitsuru Ishizuka. MPML: A markup language for controlling the behavior of life-like characters. *Journal of Visual Languages and Computing*, 15(2):183–203, 2004.

Helmut Prendinger and Mitsuru Ishizuka. Introducing the cast for social computing: Life-like characters. In H. Prendinger and M. Ishizuka, editors, *Life-Like Characters. Tools, Affective Functions, and Applications*, Cognitive Technologies, pages 3–16. Springer Verlag, Berlin Heidelberg, 2004a.

Helmut Prendinger and Mitsuru Ishizuka, editors. *Life-Like Characters. Tools, Affective Functions, and Applications*. Cognitive Technologies. Springer Verlag, Berlin Heidelberg, 2004b.

Lei Qu, Ning Wang, and W. Lewis Johnson. Pedagogical agents that interact with learners. In *AAMAS-04 Workshop on Balanced Perception and Action in ECAs*, 2004.

Akikazu Takeuchi and Taketo Naito. Situated facial displays: Towards social interaction. In *Proceedings CHI 95 Conference*, pages 450–455, New York, 1995. ACM Press.

Hiroyuki Umemuro and Jun Yamashita. Detection of user's confusion and surprise based on pupil dilation. *The Japanese Journal of Ergonomics*, 39(4): 153–161, 2003.

Susanne van Mulken, Elisabeth André, and Jochen Müller. The Persona Effect: How substantial is it? In *Proceedings Human Computer Interaction (HCI-98)*, pages 53–66, Berlin, 1998. Springer.

Mark Witkowski, Yasmine Arafa, and Oscar de Bruijn. Evaluating user reaction to character agent mediated displays using eye-tracking technology. In *Proceedings AISB-01 Symposium on Information Agents for Electronic Commerce*, pages 79–87, 2001.

Informing the Design of Embodied Conversational Agents by Analyzing Multimodal Politeness Behaviors in Human-Human Communication

Matthias Rehm and Elisabeth André
*Multimedia Concepts and Applications
Faculty of Applied Informatics
University of Augsburg
{rehm|andre}@multimedia-werkstatt.org

Abstract

In order to build embodied conversational agents that are able to communicate with the user in a more natural manner, the consideration of social aspects seems inevitable. One aspect of social interaction is the use of politeness strategies. In this paper, we report on a corpus study we conducted in order to shed light on the co-occurrence of gestures and verbal politeness strategies in face threatening situations. The results of the study will be used to inform rules for the selection of gestures in an ECA.

1 Introduction

Embodied conversational agents (ECAs) are becoming more and more realistic in their appearance and their animations. But supplying an interface agent with a body also poses great challenges to the design of appropriate interactions because the user will expect - at least in part - humanlike verbal and non-verbal conversational behaviors of such an agent. In the long run, it is therefore inevitable to enrich ECAs with social competencies to render their interactions with the user more natural and entertaining. One aspect of social interaction is the use of politeness strategies as they are described in detail in Brown and Levinson's (1987) seminal work. People maintain positive (self image) and negative face (wants and desires), which are continuously threatened during interactions, e.g., by commands or criticism on one's behavior. Such speech acts are called face threatening acts (FTAs). People try to redress or mitigate such undesirable acts, e.g., by referring to the good looks of the addressee before asking her for a favor.

Previous work has concentrated for the most part on the linguistic aspects of FTAs, i.e., on verbal means to deliver and redress FTAs. But FTAs are often multi-modal. Dressing up a threat in a joke usually only works if the speaker shows in his whole appearance (facial expression, body posture) that he is telling a joke. Otherwise the threat might be even more severe than it is (see Fig. 1 for an example).

Due to the sparse literature on the use of non-verbal communicative behaviors of politeness, we collected our own corpus based on staged conversations between humans. To trigger the use of politeness strategies, we had to make sure that the communicative situation was inherently face-threatening for the participants. We therefore decided to record scenarios where an audience had to provide criticism to the speaker. The recorded video material was annotated and analyzed in order to identify frequently occurring combinations of gestures and verbal politeness strategies.

2 Related Work

Research on non-verbal communicative behaviors, such as gestures or facial expressions, provides a good impression of the relevance of multi-modal aspects of communication, e.g. (Allwood, 2002), (Kendon, 1986), (Knapp and Hall, 1997), (Pease, 1993), and reveals a bunch of implicit information about the role of gestures and facial expressions in delivering and redressing face threats. However, there is hardly any work that explores the relationship between multi-modal means of communication and face threats. An exception is an empirical study by Trees and Manusov (1998) who found that non-verbal behaviors, such as pleasant facial expressions and more direct body orientation may help to mitigate face threats evoked by criticism. Bavelas et al. (1995) provide a classification of gestures some of which

can be directly mapped onto Brown and Levinson's strategies of politeness. Shared information gestures mark material that is part of the interlocutors common ground. Citing gestures refer to previous contributions of the addressee and aim at conveying the impression that the interlocutors share a common opinion. Elliptical gestures mark incomplete information that the addressee should augment for him- or herself and may take on a similar function as off-record strategies. Seeking agreement gestures directly correspond to Brown and Levinson's approval oriented strategies. Turn open gestures can be regarded as attempts to satisfy the addressee's desire for autonomy. Linguistic means to deliver FTAs have partly become part of the grammar and Bavelas classification of gestures suggests that there might be similar principled and standardized connection between nonverbal means of communication and politeness strategies.

Walker et al. (1997) have presented one of the first approaches to implement politeness strategies as a means to more flexible dialogue control. They summarize the available strategies into four main categories: (1) direct, (2) approval oriented, (3) autonomy oriented, (4) off record. In direct strategies, no redress is used, the speaker just expresses his wishes. Approval oriented strategies are related to the positive face needs of the addressee, using means to approve of her self-image. Autonomy oriented strategies on the other hand are related to the negative face wants of the addressee, trying to take care of her want to act autonomously. Off record strategies at last are the most vague and indirect form to address someone, demanding an active inference on the side of the addressee to understand the speaker. Depending on variables such as social distance and power, and a culture-specific rating of the speech act, a speaker chooses an appropriate strategy to deliver a face threatening act (FTA), e.g. (i) I really enjoyed your talk but you should be more coherent vs. (ii) The talk should be more coherent. In (i) the speaker compliments the addressee on her talk before delivering his critic, thus employing an approval oriented strategy. In (ii), an autonomy oriented strategy is used in impersonalizing the criticism. The speaker neither refers to the addressee nor to himself. Johnson et al. (2004) describe the value of politeness in a tutoring system. Examining the interactions between a real tutor and his students, they came up with a set of templates to generate appropriate utterances depending on the current situation. One interesting modification of the original theory by Brown and Levinson (1987) was to select approval and autonomy oriented strategies based on the type of the expected face threat (and

Figure 1: Critic of the Popidol show wrapping his criticism in a joke.

not just on its weight). André et. al (2004) augmented the model of Brown and Levinson with an emotional layer. The emotion of the addressee as it is observed by the speaker plays a crucial role in determining an appropriate strategy. Bickmore and Cassell (2000) describe how smalltalk is utilized to build up common ground between an embodied conversational agent and the user based on an extension of Brown and Levinson's theory of politeness. Nakano et al. (2003) study how people use non-verbal signals, such as eye gaze and head nods, to provide common ground in the context of direction-giving tasks. Even though their work relies on a sophisticated model of gestural communication, they did not investigate how the use of gestures may help to mitigate the face threat for the user. Porayska-Pomsta and Mellish (2004) make use of Brown and Levinson's model in order to motivate linguistic variations of a natural language generator. Prendinger and Ishizuka (2001) consider Brown and Levinson's social variables distance and power in order to control emotional displays of agents. For instance, if the social distance between an agent and its conversational partner is high, the agent would not show anger to the full extent. This behavior can be interpreted as an attempt to reduce the face threat for the conversational partner.

Summing up it may be said that the implementation of politeness behaviors in an ECA mainly focused on verbal aspects so far.

3 The Augsburg SEMMEL corpus

Since there is hardly any research into the multimodal aspects of human politeness strategies, we decided to acquire our own multi-modal corpus for an empirical grounding of the intended system. We explored two alternatives. Our first approach was to rely

on video recordings from the German version of the TV show Popidol (see Fig. 1). In this show, a number of candidates present a song. A jury comments on the performances and the viewers vote for the candidates. After some weeks, the popidol for the season emerges. The advantage of this corpus lies in the fact that the phenomena we are interested in are a major ingredient of the show. Furthermore, the TV personalities were experienced speakers that make use of expressive gestures and facial expressions. On the other hand, their behavior is certainly not representative of ordinary people. Furthermore, the corpus did not provide enough examples of multi-modal politeness behaviors since there was little criticism towards the end of the show and the gestures and facial expressions of the jury were not always visible. Although this corpus gave us interesting insights in the combined use of verbal and non-verbal politeness behavior, the limitations of the corpus only allowed for anecdotal evidence. Thus, we decided to collect a new corpus based on staged scenarios with a group of students.

3.1 Collecting the SEMMEL-Corpus

We devised a scenario that forced the participants to use their (unconscious) knowledge of politeness strategies by confronting them with an inherently face threatening situation. Criticizing someone is a prototypical example of such a situation. Therefore, we chose seminar talks with subsequent discussion to provide for a more or less ”natural” situation for the participants. The focus was on the criticism given by the audience to the speakers on their performance. Students were divided into two groups: audience and speakers. The speakers were asked to give a five minute talk about one of their hobbies. This topic was chosen to keep the necessary preparatory work for the talk at a minimum and to ensure that the audience had enough knowledge on the topic to easily criticize the speaker.

The initial explanation for this setup that was given to the participants one week before the experiment was our need to collect a corpus of non-verbal communicative behavior. This explanation also accounted for the two cameras we were using, one videotaping the speaker, the other one the audience. The initial explanation was detailed on the day of the experiment. The speakers were informed about the real setup to prevent them from reacting in an unwanted way to the critic or the criticism. The audience was told that we were interested in the reaction of the speaker to (potentially unjustified) criticism. In order to ensure that we would collect enough examples of relevant communicative acts, each member of the audience was instructed to criticize the speaker on three different dimensions and received a list of issues that had to be brought up during the discussion: (i) formal aspects, e.g. too many/too few slides, (ii) content, e.g. snowboarding is far too dangerous, and (iii) personal, e.g. the speaker was to nervous. After the experiment, the participants were informed about the actual objective of the data collection.

12 students in their first and second year participated in this data collection, three male, nine female. Four of them (two male, two female) prepared a talk on their hobby and were criticized by four audience members immediately after their presentation (see Fig. 2). The audience for each talk was constituted randomly from the remaining eight students ensuring that each of them participated two times as an audience member and met one of the other audience members only twice. We tried to hold the social variables distance and power constant and made sure that the speakers and the audience were not from the same year. The resulting SEMMEL-corpus (Strategy Extraction for MultiModal Eca controL) contains 66 different acts of criticism, i.e., 16.5 on average per talk. An act of criticism covers one of the aspects mentioned above and is always delivered with a mix of strategies and co-occuring gestures. Up to now, roughly half of our material has been annotated containing 125 combinations of strategies and gestures.

3.2 Annotating the SEMMEL-Corpus

The collected material was annotated using ANVIL (Kipp, 2003). Fig. 2 shows a screenshot of the ANVIL system along with annotations of our corpus. Focusing on the interaction of verbal and non-verbal behavior in the use of politeness strategies, the SEMMEL coding scheme features four main layers:

1. trl: The transliteration, i.e., the words spoken.
2. affective facial expression: Facial expressions that can be labeled with an emotion.
3. gesture: The hand gestures of the speaker visible in the video.
4. strategy: The politeness strategies employed by the speaker.

Focusing on the use of gestures as a non-verbal means to redress face threats, facial expressions are not annotated at the moment. In the coding scheme, facial expressions may be annotated using the affective tags available in APML (Carolis et al., 2002).

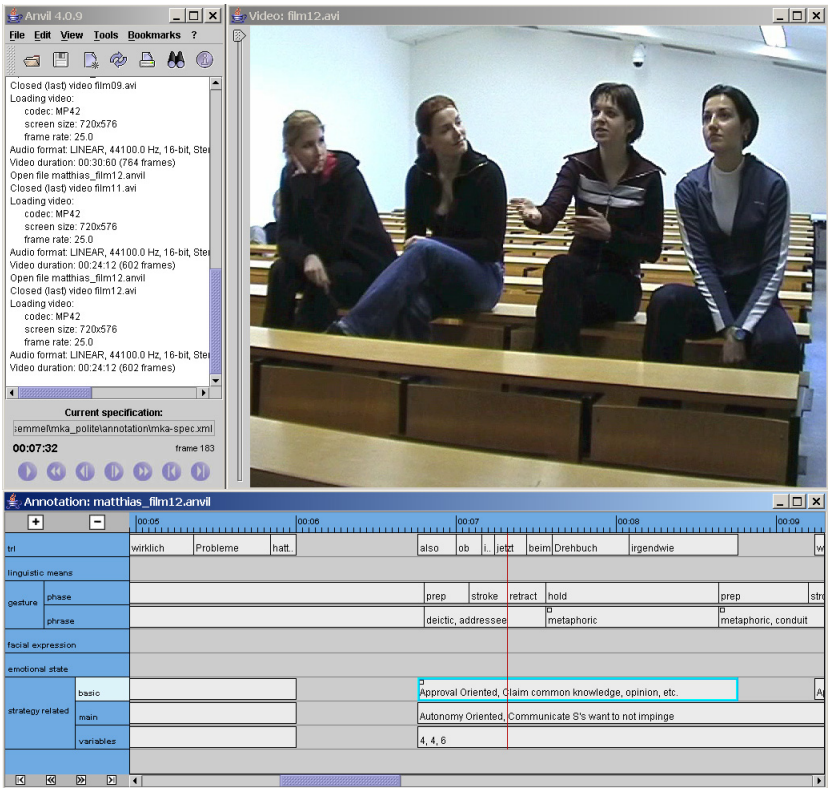


Figure 2: Snapshot from the ANVIL annotation system. Above the video is displayed, below the annotation board.

Category	Strategy	Verbal means
Direct	Direct	State the threat directly
Approval Oriented	Convey interest	Compliments, intensifying adjectives
	Claim in-group membership	Address forms, slang, elliptical utterances
	Claim common knowledge	White lies, use of ”sort of”, ”in a way”, jokes
	Indicate knowledge about wants	State to regard addressee’s wants
	Claim reflexivity	Inclusive ”we”, give/ask for reasons
	Claim reciprocity	State that addressee’s owns speaker a favor
	Fulfil wants	State sympathy
Autonomy Oriented	Make minimal assumptions	Hedges ”I think”, ”kind of”
	Give option not to act	Subjunctive, use of ”perhaps”
	Minimize threat	Euphemisms, use of ”a little”, ”just”
	Communicate want not to impinge	Avoidance of ”you” and ”I”, state threat as general rule
	Indebting	
Off Record	Violate relevance maxim	Associations, hints
	Violate quantity maxim	Exaggerations like ”always”
	Violate quality maxim	Irony, rhetorical questions
	Violate manner maxim	Ambiguity, elliptical utterances

Table 1: Types of strategies used for coding the SEMMEL corpus and examples of verbal means to realize these strategies.

In this vein, we will get well defined categories of facial expressions that can be used later for generation purposes in a straight forward way.

The coding of gestures follows Kipp's approach (Kipp, 2003) which is based on McNeill's guidelines (McNeill, 1992). Accordingly, two different parts of a gesture are distinguished: the gesture phase and the gesture phrase.

- Track gesture.phase: This is a primary track, which means that it is directly related to the video. Although gestures are mostly co-verbal, i.e., they accompany speech and add additional meaning to it by visualizing aspects of the mentioned referents, only the stroke of the gesture has verbal-nonverbal synchronization constraints. Thus it does not suffice to bind the gesture only to the transliteration layer but to the video itself. The most prominent phases of a gesture are preparation, stroke, and retraction. Generally, the hands are brought from a resting position into the gesture space during preparation. The stroke is the phase of the gesture, that carries/visualizes its meaning. Afterwards, the hands are brought back to a resting position during the retraction phase.

- Track gesture.phrase: The gesture phrase denotes the type of the gesture. It is realized as a secondary track which means it is related to another track of the coding scheme, in this case to gesture.phase. Thus, the gesture phases specify the time dimension of the gesture in regard to the video whereas the gesture phrase gives the interpretation of this specific gesture. McNeill (1992) distinguishes roughly between adaptor, beat, emblem, deictic, iconic, and metaphoric gestures. Adaptors comprise every hand movement to other parts of the body like scratching one's nose. Beats are rhythmic gestures that may emphasize certain propositions made verbally or that link different parts of an utterance. Emblems are gestures that are meaningful in themselves, i.e., without any utterance. An example is the American "OK"-emblem, where the thumb and first finger are in contact at the tips while the other fingers are extended. Deictic gestures identify referents in the gesture space. The referents can be concrete like the addressee or they can be abstract like pointing to the left and the right while uttering the words "the good and the bad". In this case the good and the bad are identified in the gesture space and it becomes possible to refer back to them later on by pointing to the corresponding position. Iconic gestures depict spatial or shape-oriented aspects of a referent, e.g., by using two fingers to indicate someone walking while uttering "he went down the street". Metaphoric gestures at last are more difficult in that they visualize abstract concepts by the use of metaphors, e.g. using a box gesture to visualize "a story". This is the conduit metaphor that makes use of the idea of a container in this case a container holding information.

The coding of strategies uses a simplified version of Brown and Levinson's hierarchy distinguishing between seven different approval oriented, five different autonomy oriented, and four different off record strategies (see Table 1).

- Track strategy.basic: Every strategy that is employed by the speaker is coded and bound to the words in the transliteration track that give rise for this interpretation. For each category of strategies (direct, approval oriented, autonomy oriented, off record), the coder has to decide for a specific type (see Table 1).

- Track strategy.main: Because a single utterance contains nearly always a mix of strategies, a track is added for coding the main strategy used in a specific utterance. The same elements as in the basic track are used (see Table 1), but the elements in this track are not bound to the transliteration but to the basic track.

- Track.variables: Brown and Levinson introduce the contextual variables social distance, power relation, and ranking of the imposition to calculate the weight of the face threat that is redressed by the strategy. This track is bound to strategy.main assuming that neither of the variables changes during a single utterance.

3.3 Analyzing the SEMMEL corpus

The first part of our analysis concentrated on the distribution of the four basic categories of politeness strategies. Remarkably is the high number of autonomy oriented strategies. From the 125 strategy/gesture combinations, 61% include autonomy oriented strategies, 18% Off record, and 15% Approval oriented strategies. By opting for autonomy oriented strategies, the critics try to leave the choice of action on the side of the addressee. Thus, the criticism is wrapped into some kind of suggestion for the addressee on how to improve the talk. We put this

Strategy	Freq.
Make minimal assumptions	0.22
Give option not to act	0.21
Minimize threat	0.22
Communicate want not to impinge	0.34

Table 2: Frequency of autonomy oriented strategies.

result down to the nature of the power relationship between the speaker and the audience. Since both the speaker and the critics were students, the critics obviously did not feel like being in the position of judging on the performance of their colleagues.

Out of the five autonomy oriented strategies, only four can be found in the corpus (see Table 2). Apart from the communicative strategy ”Communicate want not to impinge” which relies mainly on the impersonalization of the speech act (reflected by the avoidance of pronouns, such as ”you” and ”I”) and which is used in 34% of the time, the use of the other strategies is equally distributed around 22%. Most communicative acts that correspond to the category ”Make minimal assumptions” employ hedging verb phrases, such as ”I think”, ”I guess”, or ”I suppose”. In case of the strategy ”Give option not to act”, the subjunctive is widely used along with words, such as ”perhaps”. The strategy ”Minimize threat” employs minimizing expressions, such as ”a little”.

Out of the approval oriented strategies only the ”claim reflexivity” strategy was used regularly (47% of the time). This strategy was realized by giving reasons for the criticism and thus trying to explain to the addressee why the criticism is necessary. Although all off record strategies identified by Brown and Levinson (1987) can be found in the corpus only one is used regularly: violate manner maxim. To realize this strategy, the critics usually employed elliptical utterances.

Furthermore, we were interested in the distribution of gesture types. Out of the six gesture types that were annotated, only two are exceptional in the frequency of their use: beats and emblems (see Fig. 3). Whereas emblems can be rarely observed (3%), beats are the most frequently used gestures (26%). Emblematic gestures are self-sufficient in that they can be interpreted without any accompanying utterance. Thus, it is not astonishing to find them rarely as coverbal gestures. Beats are rhythmic gestures that emphasize words in an utterance or relate different parts of an utterance. But they might also connect different parts of an utterance thus indicating that the turn has not yet ended. Thus, the extensive use of beats might

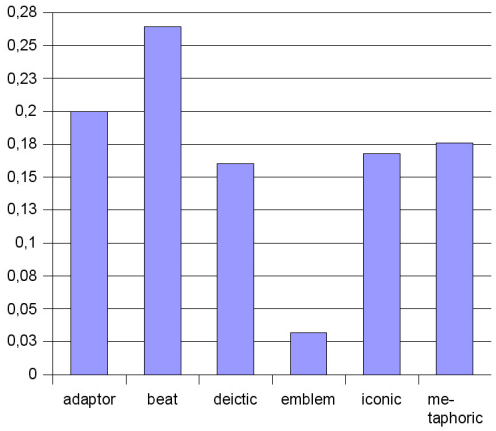


Figure 3: Distribution of gestures in the SEMMEL corpus.

be an artifact of the experimental setting because the critics had to ”invent” an act of criticism that was not their own on the fly and thus the beat gesture might be an outward sign of this process indicating that the turn has not yet ended. As noted by McNeill (1992), the number of beats depends among other things on the discourse context. He observed about 25% beats in narrative contexts which roughly corresponds to our findings versus 54% beats in extra narrative contexts.

Overall, we did not notice great differences in the distribution of deictic, iconic and metaphoric gestures. However, when analyzing their co-occurrence with politeness strategies, two general tendencies may be observed (see Fig. 4). First, adaptors are used considerably while employing autonomy oriented strategies (26%). They are used least frequently with off record strategies (5%). Off record strategies are the most ambiguous and vague means to deliver a face threat. Given that adaptors often indicate that people are nervous, the more frequent use of adaptors in autonomy oriented strategies seems plausible because the criticism is delivered more openly resulting in more stress for the speaker.

Second, there is a difference in the use of gestures of the abstract (metaphoric) and gestures of the concrete (iconic and deictic). Nearly all deictic gestures that occurred in our setting referred to the addressee or concrete locations in the space (76.8%). 50% of all gestures used with the off record strategies were metaphoric in nature vs. 14% for iconic and deictic gestures. In contrast to this, 50% of the gestures employed with the direct strategies, and 49% of the gestures employed with the approval oriented strategies were iconic and deictic in nature. The same is true

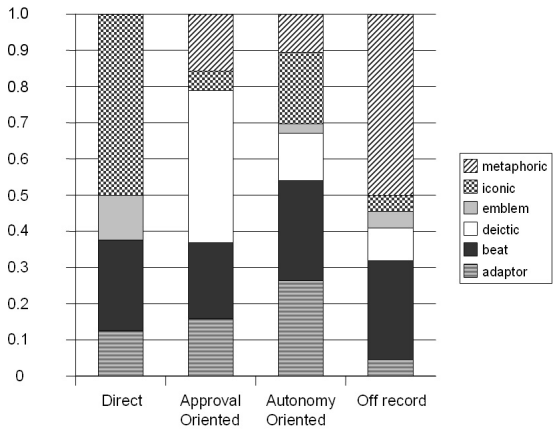


Figure 4: Co-occurrence of gestures and politeness strategies in the SEMMEL corpus.

to a lesser degree for the autonomy oriented strategies. In this case, 33% were gestures of the concrete and only 11% metaphoric gestures. Thus, the more abstract, vague and ambiguous the strategies become, the more abstract and vague the primarily employed gesture type becomes.

These results confirm the assumptions that not only linguistic regularities can be found in the use of politeness strategies, but that also non-verbal behaviors like gestures play a principled role in the realization of strategies. Metaphoric gestures relate to abstract concepts and illustrate an aspect of a referent in the utterance by the aid of a metaphor. The best known metaphoric gesture is the conduit metaphor where the hands form a kind of container that symbolizes the concept of a story or narrative. Most of the time, metaphoric gestures contain iconic as well as abstract parts. Why are metaphoric gestures found foremost with off record strategies? In contrast to direct strategies which do not consider the loss of face of the addressee and in contrast to approval and autonomy oriented strategies where the direct criticism is redressed but still visible, off record strategies just hint at what the speaker intends to deliver as a message, leaving the addressee at a loss to inference the speaker's intention. Being vague and ambiguous does not leave much ground for concrete gestures that refer to aspects of concrete and direct referents. Thus, metaphoric gestures are the first choice for co-verbal gestures while employing off record strategies. The contrary argument holds for the other types of strategies and the gestures of the concrete. For example, employing a direct strategy, one of the critics said: "'... some pictures of the instruments, especially of this cornet[iconic] that you mentioned"'[1]. The direct referent cornet is iconically visualized by outlining the shape. The left hand is raised like holding the cornet, the index finger of the right hand is extended and the hand describes a circle. In the off record case the speaker might try to give only association clues, such as another critic who used an elliptical utterance: "'not so clearly to identify ... so of the structure[metaphoric] ... structure you have somehow"'[2]. Here the verbal information is accompanied by a gesture which comes in the form of the conduit metaphor. The left and right hand indicate holding something like a box.

4 Conclusions and future work

In this paper, we presented the results of a corpus study we devised to shed light on the question of how face threats are mitigated by multi-modal communicative acts. Unlike earlier work on politeness behaviors, we focus on how politeness is expressed by means of gestures. The results we presented are preliminary because up to now roughly half of the material has been annotated. But we are confident that the found tendencies will scale up to the whole corpus. The results indicate that gestures are indeed used to strengthen the effect of verbal acts of politeness. In particular, vagueness as a means of politeness is not only reflected by verbal utterances, but also by gestures. Iconic and deictic gestures were overwhelmingly used in more direct criticism while there was a high frequency of metaphoric gestures in off record strategies. Obviously, our subjects did not attempt at compensating for the vagueness of their speech by using more concrete gestures. Interestingly, McNeill (1992, pp. 93) noticed a high number of sequence-related iconics and deictics in narrative contexts while metaphorics appear more frequently in extra-narrative contexts. The question arises of whether the critics rather referred to the story line of the presentation in the case of direct criticism while indirect criticism rather addresses the meta narrative structure level. We will investigate this question in a further study.

The results gained from our studies may serve as guidelines for the formulation of non-verbal strategies of politeness for an ECA. We illustrate this by the BEAT system presented by Cassell et al. (2001). BEAT suggests non-verbal gestures based on a linguistic and contextual analysis of typed text. Since

[1]Original utterance: "'... ein paar Bilder der Instrumente, also gerade dieses Horn[iconic] dass du angesprochen hast"'

[2]Original utterance: "'nicht so klar erkennen ... so von der Struktur[metaphoric] ... Struktur habt ihr euch irgendwie"'

non-verbal behaviors are generated independently of each other, the system may end up with a set of incompatible gestures. The set of proposed gestures is therefore reduced to those gestures that are actually realized by the animation module. The findings of our studies may inform both the generation of gestures and the filtering process of the BEAT system. For instance, deictic gestures may be given a higher priority than iconic gestures when suggesting non-verbal behaviors for approval oriented strategies. On the other hand, they may be filtered out with a higher probability when realizing off record strategies. Currently, we are preparing an empirical study to compare the effect of two kinds of ECA on the user's perception of the interaction: an ECA that reflects the degree of vagueness both by speech and gestures versus an ECA that behaves inconsistently in that respect.

References

Jens Allwood. Bodily Communication Dimensions of Expression and Content. In Björn Granström, David House, and Inger Karlsson, editors, *Multimodality in Language and Speech Systems*, pages 7–26. Kluwer Academic Publishers, Dordrecht, Boston, London, 2002.

Elisabeth Andre, Matthias Rehm, Wolfgang Minker, and Daniel Bühler. Endowing Spoken Language Dialogue Systems with Social Intelligence. In Elisabeth Andre, Laila Dybkjaer, Wolfgang Minker, and Paul Heisterkamp, editors, *Affective Dialogue Systems*, pages 178–187. Springer, Berlin, 2004.

Janet B. Bavelas, Nicole Chovil, L. Coates, and L. Roe. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21:394–405, 1995.

Timothy Bickmore and Justine Cassell. Small Talk and Conversational Storytelling in Embodied Interface Agents. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*, pages 87–92, 2000.

Penelope Brown and Stephen C. Levinson. *Politeness — Some universals in language usage*. Cambridge University Press, Cambridge, 1987.

Berardina De Carolis, Valeria Carofiglio, Massimo Bilvi, and Catherine Pelachaud. APML, a Markup Language for Believable Behavior Generation. In Z. Ruttkay and C. Pelachaud, editors, *Workshop AAMAS02: Embodied conversational agents — let's specify and evaluate them!*, 2002.

Justine Cassell, Hannes Vilhjalmsson, and Timothy Bickmore. BEAT: The Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH '01*, pages 477–486, Los Angeles, CA, 2001.

W. Lewis Johnson, Paolo Rizza, Wauter Bosma, Sander Kole, Mattijs Ghijsen, and Hermin van Welbergen. Generating Socially Appropriate Tutorial Dialog. In Elisabeth Andre, Laila Dybkjaer, Wolfgang Minker, and Paul Heisterkamp, editors, *Affective Dialogue Systems*, pages 254–264. Springer, Berlin, 2004.

Adam Kendon. Some reasons for studying gestures. *Semiotica*, 62(1-2):3–28, 1986.

Michael Kipp. *Gesture Generation by Imitation — From Human Behavior to Computer Character Animation*. PhD thesis, Universität des Saarlandes, Saarbrücken, 2003.

Mark L. Knapp and Judith A. Hall. *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, Fort Worth, 4. edition, 1997.

David McNeill. *Hand and Mind — What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, London, 1992.

Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a Model of Face-to-face Grounding. In *Proceedings of the Association for Computational Linguistics*, Sapporo, Japan, July 1–12 2003.

Allan Pease. *Body language — How to read others' thought by their gestures*. Sheldon Press, London, 20th edition, 1993.

Kaska Porayska-Pomsta and Chris Mellish. Modelling Politeness in Natural Language Generation. In *Proceedings of INLG*, 2004.

Helmut Prendinger and Mitsuru Ishizuka. Social Role Awareness in Animated Agents. In *Proceedings of Agents '01, Montreal, Canada*, pages 270–277, 2001.

April R. Trees and Valerie Manusov. Managing Face Concerns in Criticism — Integrating Nonverbal Behaviors as a Dimension of Politeness in Female Friendship Dyads. *Human Communication Research*, 24(4):564–583, 1998.

Marilyn A. Walker, Janet E. Cahn, and Stephen J. Whittaker. Improvising Linguistic Style: Social and Affective Bases for Agent Personality. In *Proceedings of AAMAS'97*, 1997.