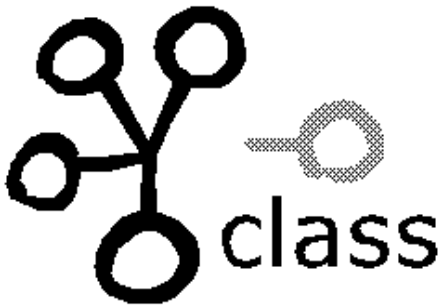
International CLASS Workshop

on

Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems

Proceedings



Edited by

Jan van Kuppevelt, Laila Dybkjær and Niels Ole Bernsen

Copenhagen, Denmark
28-29 June 2002

© 2002

Printed at University of Southern Denmark

PREFACE

We are happy to present the proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, that was held in Copenhagen, Denmark, 28-29 June 2002. The workshop was sponsored by the European CLASS project (http://www.class-tech.org). CLASS was initiated on the request of the European Commission with the purpose of supporting and stimulating collaboration within and among Human Language Technology (HLT) projects, as well as between HLT projects and relevant projects outside Europe.

The workshop was given a special format with the main purpose of bringing into focus both theoretically and practically oriented research that has given rise to innovative and challenging approaches on natural, intelligent and effective interaction in multimodal dialogue systems. In order to reach this goal we planned the workshop to contain a relatively high number of invited contributions in addition to papers solicited via an open Call for Papers. We invited a group of 9 internationally leading researchers with a balanced composition of expertise on the topics of the workshop. We were especially interested in the following topics:

- *Multimodal Signal Processing*

  Models for multimodal signal recognition and synthesis, including combinations of speech (emotional speech and meaningful intonation for speech), text, graphics, music, gesture, face and facial expression, and (embodied) animated or anthropomorphic conversational agents.
- *Multimodal Communication Management*

  Dialogue management models for mixed initiative conversational and user-adaptive natural and multimodal interaction, including models for collaboration and multi-party conversation.
- *Multimodal Miscommunication Management*

  Multimodal strategies for handling or preventing miscommunication, in particular multimodal repair and correction strategies, clarification strategies for ambiguous or conflicting multimodal information, and multimodal grounding and feedback strategies.
- *Multimodal Interpretation and Response Planning*

  Interpretation and response planning on the basis of multimodal dialogue context, including (context-semantic) models for the common representation of multimodal content, as well as innovative concepts/technologies on the relation between multimodal interpretation and generation.
- *Reasoning in Intelligent Multimodal Dialogue Systems*

  Non-monotonic reasoning techniques required for intelligent interaction in various types of multimodal dialogue systems, including techniques needed for multimodal input interpretation, for reasoning about the user(s), and for the coordination and integration of multimodal input and output.
- *Choice and Coordination of Media and Modalities*

  Diagnostic tools and technologies for choosing the appropriate media and input and output modalities for the application and task under consideration, as well as theories and technologies for natural and effective multimodal response presentation.
- *Multimodal Corpora, Tools and Schemes*

  Training corpora, test-suites and benchmarks for multimodal dialogue systems, including corpus tools and schemes for multilevel and multimodal coding and annotation.
- *Architectures for Multimodal Dialogue Systems*

  New architectures for multimodal interpretation and response planning, including issues of reusability and portability, as well as architectures for the next generation of multi-party conversational interfaces to distributed information.
- *Evaluation of Multimodal Dialogue Systems*

  Current practice and problematic issues in the standardisation of subjective and objective multimodal evaluation metrics, including evaluation models allowing for adequate task fulfilment

measurements, comparative judgements across different domain tasks, as well as models showing how evaluation translates into targeted, component-wise improvements of systems and aspects.

The proceedings contain 21 contributions. An online version of the proceedings can be found on the workshop web page (http://www.class-tech.org/events/NMI_workshop2). In addition to 7 invited contributions (2 invited contributions were cancelled) we received 21 paper submissions of which 14 were selected for presentation at the workshop.

Together with invited contributions, a selected number of extended and updated versions of papers contained in these proceedings will appear in a book to be published by Kluwer Academic Publishers.

We are in particular grateful for the work done by the members of the Program Committee which are leading and outstanding researchers in the field. The authors of papers submitted to the workshop clearly have benefited from their expertise and efforts. The names of the members of the Program Committee are presented on the next page.

Further, we would like to thank Tim Bickmore, Phil Cohen, Ronald Cole, Björn Granström, Dominic Massaro, Candy Sidner, Oliviero Stock, Wolfgang Wahlster and Yorick Wilks for accepting our invitation to serve as invited speaker. Unfortunately, both Wolfgang Wahlster and Yorick Wilks had to cancel their participation to the workshop. We are convinced that the presence of our invited guests will add greatly to the quality and importance of the workshop.

Finally, we want to acknowledge the assistance work done by the NISLab team at University of Southern Denmark, in particular the clerical support provided by Merete Bertelsen and the valuable and direct internet support given by Torben Kruchov Madsen.

We hope you will benefit greatly from these proceedings and your participation in the workshop.

Jan van Kuppevelt (IMS, University of Stuttgart),
Laila Dybkjær (NISLab, University of Southern Denmark) and
Niels Ole Bernsen (NISLab, University of Southern Denmark).

PROGRAM COMMITTEE

Co-Chairs:

- Niels Ole Bernsen (NISLab, University of Southern Denmark)
- Jan van Kuppevelt (IMS, University of Stuttgart)

Reviewers:

- Elisabeth Andre (University of Augsburg)
- Tim Bickmore (MIT Media Lab)
- Louis Boves (Nijmegen University)
- Justine Cassell (MIT Media Lab)
- Phil Cohen (Oregon Graduate Institute)
- Ronald Cole (University of Colorado at Boulder)
- John Dowding (RIACS)
- Laila Dybkjær (NISLab, University of Southern Denmark)
- Björn Granström (KTH, Stockholm)
- Jean-Claude Martin (CNRS/LIMSI, Paris)
- Dominic Massaro (UCSC)
- Catherine Pelachaud (University of Rome "La Sapienza")
- Thomas Rist (DFKI)
- Candy Sidner (MERL, Cambridge, MA)
- Mark Steedman (University of Edinburgh)
- William Swartout (ICT, USC)
- Oliviero Stock (ITC-IRST)
- Yorick Wilks (University of Sheffield)

ORGANIZING COMMITTEE

- Niels Ole Bernsen (NISLab, University of Southern Denmark)
- Laila Dybkjær (NISLab, University of Southern Denmark)
- Jan van Kuppevelt (IMS, University of Stuttgart)

WORKSHOP PROGRAM

FRIDAY June 28

08.00 - 08.45 **Registration**

08.45 - 09.00 **Opening**

09.00 - 09.50 Invited Speaker: Phil Cohen
On the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence

09.50 - 10.15 Nicole Beringer, Sebastian Hans, Katerina Louka and Jie Tang
How to Relate User Satisfaction and System Performance in Multimodal Dialogue Systems? - A Graphical Approach

10.15 - 10.45 **Coffee break**

10.45 - 11.35 Invited Speaker: Oliviero Stock
Intelligent Interactive Information Presentation for Cultural Tourism

11.35 - 12.00 Jan-Torsten Milde
Creating Multimodal, Multilevel Annotated Corpora with TASX

12.00 - 12.40 Short paper presentation session 1

1. Luis Almeida, Ingunn Amdal, Nuno Beires, Malek Boualem, Lou Boves, Els den Os, Pascal Filoche, Rui Gomes, Jan Eikeset Knudsen, Knut Kvale, John Rugelbak, Claude Tallec, Narada Warakagoda
   Implementing and Evaluating a Multimodal Tourist Guide

2. Sorin Dusan and James Flanagan
   An Adaptive Dialogue System Using Multimodal Language Acquisition

3. Carl Burke, Lisa Harper and Dan Loehr
   A Dialogue Architecture for Multimodal Control of Robots

12.40 - 14.00 **Lunch break and poster visit**
- Poster visit from 13.30 to 14.00 -

14.00 - 14.50 Invited Speaker: Tim Bickmore
Phone vs. Face-to-Face with Virtual Persons

14.50 - 15.15 Noelle Carbonell and Suzanne Kieffer
Do Oral Messages Help Visual Exploration?

15.15 - 15.45 **Tea break**

15.45 - 16.35 Invited Speaker: Candy Sidner
Engagement between Humans and Robots for Hosting Activities

16.35 - 17.00 G.T. Healey and Mike Thirlwell
Analysing Multi-Modal Communication: Repair-Based Measures of Communicative Co-ordination

18.30 - 19.30 **Reception**

SATURDAY June 29

09.00 - 09.50 Invited Speaker: Ron Cole
Perceptive Animated Interfaces: The Next Generation of Interactive Learning Tools

09.50 - 10.15 T. Darrell, J. Fisher and K. Wilson
Geometric and Statistical Approaches to Audiovisual Segmentation for Untethered Interaction

10.15 - 10.45 Coffee break

10.45 - 11.35 Invited Speaker: Björn Granström
Effective Interaction with Talking Animated Agents in Dialogue Systems

11.35 - 12.00 Dirk Heylen, Ivo van Es, Anton Nijholt, Betsy van Dijk
Experimenting with the Gaze of a Conversational Agent

12.00 - 12.40 Short paper presentation session 2

1. Brady Clark, Elisabeth Owen Bratt, Stanley Peters, Heather Pon-Barry, Zack Thomsen-Gray and Pucktada Treeratpituk
   A General Purpose Architecture for Intelligent Tutoring Systems

2. Dave Raggett
   Task-Based Multimodal Dialogs

3. Norbert Reithinger, Christoph Lauer, and Laurent Romary
   MIAMM - Multidimensional Information Access using Multiple Modalities

12.40 - 14.00 Lunch break and poster visit
- Poster visit from 13.30 to 14.00 -

14.00 - 14.50 Invited Speaker: Dominic Massaro
The Psychology and Technology of Talking Heads in Human-Machine Interaction

14.50 - 15.15 Tea break

15.15 - 16.45 Panel discussion
Co-chairs: Niels Ole Bernsen and Oliviero Stock
Panellists: Tim Bickmore, Phil Cohen, Ron Cole, Björn Granström, Dominic Massaro, Candy Sidner

16.45 - 17.00 Closing

TABLE OF CONTENTS

Implementing and Evaluating a Multimodal Tourist Guide
Luis Almeida, Ingunn Amdal, Nuno Beires, Malek Boualem, Lou Boves, Els den Os, Pascal Filoche, Rui Gomes, Jan Eikeset Knudsen, KnutKvale, John Rugelbak, Claude Tallec, Narada Warakagoda ……. 1

How to Relate User Satisfaction and System Performance in Multimodal Dialogue Systems? - A Graphical Approach
Nicole Beringer, Sebastian Hans, Katerina Louka and Jie Tang 8

Phone vs. Face-to-Face with Virtual Persons
Timothy Bickmore and Justine Cassell [Invited Contribution] ……..…… 15

A Dialogue Architecture for Multimodal Control of Robots
Carl Burke, Lisa Harper and Dan Loehr ……..……. 23

Do Oral Messages Help Visual Exploration?
Noelle Carbonell and Suzanne Kieffer ……. 27

MIND: A Semantics-based Multimodal Interpretation Framework for Conversational Systems
Joyce Chai, Shimei Pan and Michelle X. Zhou …… 37

A General Purpose Architecture for Intelligent Tutoring Systems
Brady Clark, Elisabeth Owen Bratt, Stanley Peters, Heather Pon-Barry, Zack Thomsen-Gray and Pucktada Treeratpituk …… 47

Perceptive Animated Interfaces: The Next Generation of Interactive Learning Tools
Ron Cole [Invited Contribution] …….. 51

On the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence
Andrea Corradini and Philip R. Cohen [Invited Contribution] …… 52

Geometric and Statistical Approaches to Audiovisual Segmentation for Untethered Interaction
T. Darrell, J. Fisher and K. Wilson …… 62

An Adaptive Dialogue System Using Multimodal Language Acquisition
Sorin Dusan and James Flanagan …… 72

Effective Interaction with Talking Animated Agents in Dialogue Systems
Björn Granström and David House [Invited Contribution] …….. 76

Analysing Multi-Modal Communication: Repair-Based Measures of Communicative Co-ordination
Patrick G.T. Healey and Mike Thirlwell ……... 83

Experimenting with the Gaze of a Conversational Agent
Dirk Heylen, Ivo van Es, Anton Nijholt, Betsy van Dijk …….. 93

FORM: An Extensible, Kinematically-based Gesture Annotation Scheme
Craig Martell ……. 101

The Psychology and Technology of Talking Heads in Human-Machine Interaction
Dominic W. Massaro [Invited Contribution] …….. 106

Creating Multimodal, Multilevel Annotated Corpora with TASX
Jan-Torsten Milde ……. 120

Task-Based Multimodal Dialogs
Dave Raggett ……. 127

MIAMM - Multidimensional Information Access using Multiple Modalities
Norbert Reithinger, Christoph Lauer, and Laurent Romary ……. 137

Engagement between Humans and Robots for Hosting Activities
Candace L. Sidner [Invited Contribution] …… 141

Intelligent Interactive Information Presentation for Cultural Tourism
Oliviero Stock and Massimo Zancanaro [Invited Contribution] …….. 152

Implementing and evaluating a multimodal and multilingual tourist guide

Luis Almeida* (1), Ingunn Amdal (2), Nuno Beires (1), Malek Boualem (3), Lou Boves (4), Els den Os (5), Pascal Filoche (3), Rui Gomes (1), Jan Eikeset Knudsen (2), Knut Kvale (2), John Rugelbak (2), Claude Tallec (3), Narada Warakagoda (2)
* Authors in alphabetic order

(1) Portugal Telecom Inovação,
(2) Telenor R&D,
(3) France Télécom R&D,
(4) University of Nijmegen,
(5) Max Planck Institute for Psycholinguistics

E-Mail: els.denos@mpi.nl

Abstract

This paper presents the EURESCOM[1] project MUST, (MUltimodal, multilingual information Services for small mobile Terminals). The project started in February 2001 and will last till the end of 2002. Based on existing technologies and platforms a multimodal demonstrator (the MUST tourist guide to Paris) has been implemented. This demonstrator uses speech and pen (pointing) for input, and speech, text, and graphics for output. In addition a multilingual Question/ Answering system has been integrated to handle out of domain requests. The paper focuses on the implementation of the demonstrator. The real-time demonstrator was used for evaluations performed by usability experts. The results of this evaluation are also discussed.

Introduction

For Telecom Operators and Service Providers it is essential to stimulate the widest possible use of the future UMTS networks. Wide usage presupposes that services fulfil at least two requirements: customers must have the feeling that the service offers more or better functionality than existing alternatives, and the service must have a easy and natural interface. Especially the latter requirement is difficult to fulfil with the interaction capabilities of the small lightweight mobile handsets. Terminals that combine speech and pen at the input side, and text, graphics, and audio at the output side in a small form factor, promise to offer a platform for the design of multimodal interfaces that should overcome the usability problems. However, the combination of multiple input and output modes in a single session appears to pose new technological and human factors problems of its own. The research departments of three Telecom Operators collaborate with two academic institutes in the EURESCOM project MUST (Boves & den Os, 2002)[1]. The main aims of MUST are:

1. Getting hands-on experience by integrating existing speech and language technologies into an experimental multimodal interface to a realistic real-time demonstrator in order to get a better understanding of the issues that will be important for future multimodal and multilingual services in the mobile networks accessed from small terminals.
2. Use this demonstrator to conduct human factor experiments with naive non-professional users to evaluate the multimodal interaction.

Multimodal interaction has been studied for several years, see e.g. (Oviatt, 1999 and Oviatt et al, 2000). Most papers on user studies report experiments that were carried out with Wizard-of-Oz systems and professional users who manipulated objects on large terminal screens (Kehler et al., 1998, Martin et al., 1998, and Wahlster et al., 2001). For the Telecom Operators these studies

[1] Updated information from the MUST-project can be found at http://www.eurescom.de/public/projects/P1100-series/p1104/default.asp

are of interest in so far that they indicate some of the general principles of multimodal interaction. However, Telcos can only start to consider developing multimodal services if these can be built on standard architectures and off-the-shelf components, that work in real-time and that can be accessed from small mobile terminals by non-professional users. Therefore, the MUST project is focused on a user study with a real-time demonstrator of what could become a real service.
In addition, a large part of the existing literature is based on experiments that address issues such as the preference for specific modes for error repair and comparisons of several combinations of modes (including unimodal interaction). In MUST we concentrate on gathering knowledge about behaviour of untrained users interacting with one –carefully designed- multimodal system that is virtually impossible to use without combining speech and pen for input.
In this paper we first present the functionality of the demonstrator service that served as the backbone of the MUST project. Then we describe the architecture, and the user interface. Finally, we present the results of an expert evaluation of the first operational version of the demonstrator.

1 The functionality of the demonstrator

Multimodal interaction comes in several forms that imply different functionalities for the user. In MUST we decided to investigate the most powerful approach, i.e. simultaneous coordinated multimodal interaction[2]. We want to provide Telecom Operators with information on what this type of interaction implies in terms of implementation effort and on how users will appreciate this new way of interaction.
Only some of the services that one might want to develop for the mobile Internet networks lend itself naturally to the use of simultaneous coordinated interaction combining speech and text input. A necessary requirement for such a service is the need to talk about objects that can be identified by pointing at them on the screen. One family of services where pointing and speaking can be complementary is when a user is required to talk about objects on a map. This probably explains why multimodal map services have been so popular in the research community (Oviatt, et al, 2000; Martin et al., 1998). Tourist guides that are organised around detailed maps of small sections of a city are an example of this family of services. Therefore, we decided to model the MUST demonstrator service after this metaphor. Paris was selected as the object city.
Thus, the MUST Guide to Paris is organized in the form of small sections of the town around "Points of Interests" (POI's), such as the Eiffel tower, the Arc de Triumph, etc. These POI's are the major entry point for navigation. The maps show not only the street plan, but also pictorial representations of major buildings, monuments, etc. When the user selects one of the POI's, a detailed map of the surroundings of that object is displayed on the screen of the terminal (cf. Fig. 2). Many map sections will contain additional objects that might be of interest to the visitor. By pointing at these objects on the screen they become the topic of the conversation, and the user can ask questions about these objects, for example "What is this building?", and "What are the opening hours?". The user can also ask more general questions about the section of the city that is displayed, such as "What restaurants are in this neighbourhood?' The latter question will add icons for restaurants to the display, that can be turned into the topic of conversation by pointing and asking questions, for example about the type of food that is offered, the price range, and opening hours. The information returned by the system is rendered in the form of text, graphics (maps, and pictures of hotels and restaurants), and text-to-speech synthesis.
For mobile network operators a substantial part of access to services comes from roaming customers. It is well-known that most people prefer to use their native language, especially when using speech recognisers, that are known to degrade in performance for non-native speech. Therefore, information services offered in the mobile networks must be multilingual, so as to allow every customer to use the preferred language. The MUST demonstrator is developed for Norwegian, Portuguese, French and English.

Users will be allowed to ask questions about POI's for which the answers are not in the database of the service, perhaps because only a small

[2] Simultaneous coordinated multimodal interaction is the term used by W3C http://www.w3.org for the most complicated multimodal interaction, where all available input devices are active simultaneously, and their actions are interpreted in context.

proportion of the users is expected to be interested in this information (e.g., 'Who is the architect of this building?' and 'What other buildings has he designed in Paris?'). For the answers to these questions access will be provided to a multilingual Question/Answering (Q/A) system, developed by France Télécom R&D, that will try to find the answers on the Internet (Boualem and Filoche, n.y.).

2 The architecture of the demonstrator

The overall architecture of the MUST demonstrator is shown in Figure 1. The server side of the architecture combines a number of specialised modules, that exchange information among each other. The server is accessed by the user through a thin client that runs on the mobile terminal. The application server is based on the Portugal Telecom Inovação (Azevedo and Beires, 2001) and Telenor R&D (Knudsen et al., 2000) voice servers, which were originally designed for voice-only services, i.e. there are two versions of the demonstrator that only differ in the voice platforms used. The voice servers provide an interface to ISDN and PSTN telephony and advanced voice resources such as Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis (TTS). The ASR applied is Philips *SpeechPearl2000*, that supports all the languages in the project (English, French, Portuguese and Norwegian). ASR-features such as confidence scores and N-best lists are supported. The TTS engine is used to generate real-time speech output. Different TTS-engines are used for the different languages in MUST. Telenor and France Télécom use home-built TTS engines, while Portugal Telecom uses *RealSpeak* from L&H.

The multilingual question-answering (Q/A) system uses a combination of syntactic/semantic parsing and statistical natural Language Processing techniques to search the Web for potentially relevant documents. The search is based on a question expressed in natural language, and the system subsequently tries to extract a short answer from the documents. The size (in terms of number of characters) of the answer cannot be predicted in advance, but it is expected that most answers are short enough to fit into the text box that is used for presenting information that is already available in the database. If an answer is too long, it will be provided by Text to Speech.

The GALAXY Communicator Software Infrastructure, a public domain reference version of DARPA Communicator maintained by MITRE (http://fofoca.mitre.org), has been chosen as the underlying inter-module communication framework of the system. It also provides the HUB in Figure 1, through which nearly all the inter-module messages are passed. The main features of this framework are modularity, distributed nature, seamless integration of the modules, and flexibility in terms of inter-module data exchange (synchronous and asynchronous communication through HUB and directly between modules). GALAXY allows to 'glue' existing components (e.g., ASR, TTS, etc.) together in different ways by providing extensive facilities for passing messages between the components through the central HUB. A component can easily invoke a functionality that is being provided by other components without knowing which component provides it or where it is running.

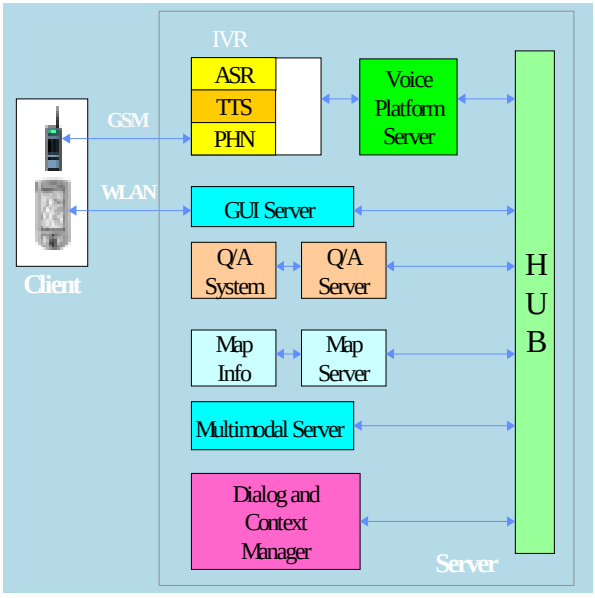


Figure 1. Schematic architecture of the MUST tourist guide to Paris

The processing in the HUB can be controlled using a script or it can act as a facilitator in an agent based system. In MUST the HUB messaging control is script based. The modules are written in Java and C/C++ under Linux and Windows NT.

In order to keep the format of the messages exchanged between the modules simple and flexible, it has been decided to use an XML based

mark-up language named MxML - MUST XML Mark up Language. MxML is used to represent most of the multimodal content that is exchanged between the modules. Parameters required for set-up, synchronization, and disconnection of modules use key pair (name - value) attributes in Galaxy messages.
The client part of the demonstrator is implemented on a COMPAQ iPAQ Pocket PC running Microsoft CE with WLAN connection. The speech part is handled by a mobile phone. The user will not notice this "two part" solution, since the phone will be hidden and the interface will be transparent. Only the headset (microphone and earphones) with a wireless connection will be visible for the user.
The spoken utterances are forwarded to the speech recogniser by the telephony module. The text and pen inputs are transferred from the GUI Client via the TCP/IP connection to the GUI Server. The inputs from the speech recogniser and the GUI Server are integrated in the Multimodal Server (late fusion) and passed to the Dialogue/Context Manager (DM). The DM interprets the result and acts accordingly, for example by contacting the Map Server and fetching the information to be presented for the user. The information is then sent to the GUI Server and Voice Server via the Multimodal Server that performs the fission. Fission consists of the extraction of data addressed to the output modalities (speech and graphics in this case).
MUST set out to investigate implementation issues related to coordinated simultaneous multimodal input, i.e. *all* parallel inputs must be interpreted in combination, depending on the fusion of the information from all channels. In our implementation we opted for the "late fusion" approach, where recogniser outputs are combined at a semantic interpretation level. The temporal relationship between different input channels is obtained by considering all input contents within a reasonable time window. The length of this time window has a default value of 1 second and is a variable parameter that can be adjusted dynamically according to the dialog context.

3 The user interface of the demonstrator

One important feature for the user interface is the "Tap While Talk" functionality. When the pen is used shortly before, during or shortly after speech, the two input actions are integrated into one combined action. An example is the utterance "Show hotels here", while tapping at Notre Dame. When the time between tapping and speech is longer than a pre-set threshold, the actions are considered as sequential and independent.

The overall interaction strategy is user controlled, in accordance with what is usual in graphical user interfaces. This implies that the speech recogniser must always be open to capture input. Obviously, this complicates signal processing and speech recognition. However, it is difficult to imagine an alternative for a continuously active ASR without changing the interaction strategy. Users can revert to sequential operation by leaving enough time between speech and pen actions.
The output information is mainly presented in the form of text (e.g. "the entrance fee is 3 euro") and graphics (maps and pictures of hotels and restaurants). The text output appears in a text box on the screen.

To help the user keep track of the system status, the system will always respond to an input. In most cases the response is graphical. For example, when a Point of Interest (POI) has been selected, the system will respond by showing the corresponding map. If the system detects an ambiguity (e.g. if audio input was detected, but ASR was not able to recognise the input with sufficiently high confidence), it provides a prompt saying that it did not understand the utterance.
The graphical part of the user interface consists of two types of maps: an overview map showing all POIs, and detailed maps with a POI in the centre. The Dialogue/Context Manager is designed such that the interaction starts without a focus for the dialogue. Thus, the first action that a user must take is to select a POI. The selected object automatically becomes the focus of the dialogue: all deictic pronouns, requests etc. now refer to the selected object. Selection can be accomplished in three ways: by speaking, by pointing, or by both simultaneously. Irrespective of the selection mode, the application responds by showing the section map that contains the POI. A selected object is marked by a red frame surrounding it, as a graphical response to the selection action. All additional selectable objects on a map are indicated by green frames. When

the user has selected a POI, several facilities such as hotels and restaurants can be shown as objects on the maps. This can be accomplished by means of speech (by asking a question such as 'What hotels are there in this neighbourhood?'), or by tapping on one of the 'facility' buttons that appear at the bottom of the screen, just below each section map.

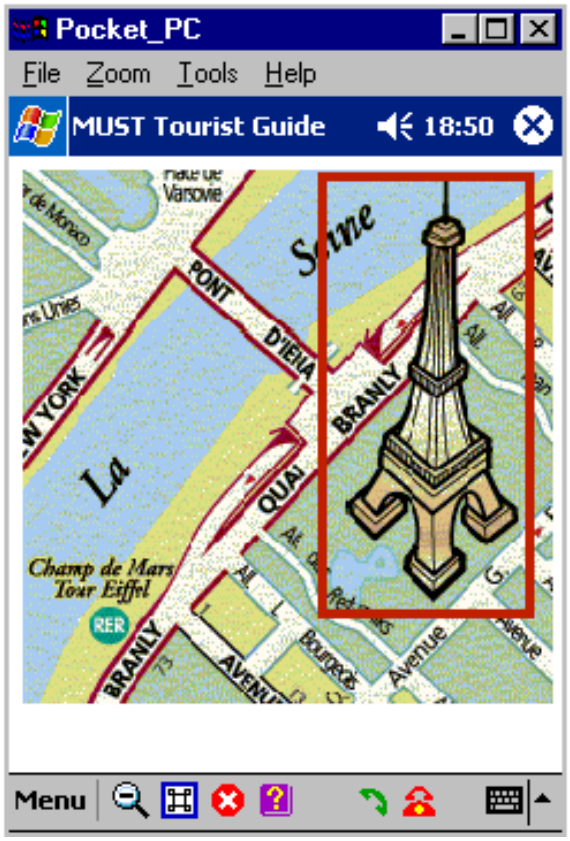


Figure 2. Screen Layout of the MUST tourist guide

Fig. 2 shows the buttons that were present in the toolbar of the first version of the GUI. Two buttons are related to the functionality of the service (hotels and restaurants), and three buttons are related to navigation: a help button, a home button, and a back button. The back button will make the application go back to the previous state of the dialogue as a kind of error recovery mechanism to deal with recognition failures. 'Help' was context independent in the first version of the demonstrator; the only help that was provided was a short statement saying that speech and pen can be one by one or combined to interact with the application.

Speech input allows what we call shortcuts. For example, at the top navigation level (where the overview map with POIs is on the screen) the user can ask questions such as 'What hotels are there near the Notre Dame?'. That request will result in the detailed map of the Notre Dame, with the locations of hotels indicated as selectable objects. However, until one of the hotels is selected, the Notre Dame will be considered as the topic of the dialogue.

4 Expert review

The MUST application was investigated by Norwegian and Portuguese experts in human-machine interaction. Since only twelve experts participated in this evaluation, results should be interpreted with due caution. There were great similarities between the remarks and observations of the Portuguese and Norwegian experts. The most noteworthy observations will be discussed here.

During the exploratory phase of the evaluation, most experts started to use the two input modalities one by one, and some of them never tried to use them simultaneously. After a while five of the twelve experts started to use pen and speech simultaneously.

Timing between speech and pointing has been studied in other experiments (Martin et al. 1998; Kehler et al., 1998). In the expert evaluation we observed that the experts typically tapped at the end or shortly after the utterance. This was especially the case when the utterances ended with deictic expressions like 'here' or 'there'. If no deictic expressions were present, tapping often occurred somewhat earlier. Timing relations between speech and pointing will be investigated in more detail in the user evaluation experiment that is now being designed.

The results from the exploratory phase indicate that frequent PC and PDA users are so accustomed to use a single modality (pen or mouse) to select objects or navigate through menus to narrow down the search space, that even if they are told that it is possible to use speech and pen simultaneously, they will have to go through a learning process to get accustomed to the new simultaneous coordinated multimodal interaction style. But once they have discovered and experienced it, the learning curve appears to be quite steep.

It was not intuitive and obvious that the interface was multimodal, and in particular that the two modalities could be used simultaneously. This indicates that for the naïve user evaluation we should pay much attention to the introduction phase where we explain the service and the interface to the user.

During the expert evaluation many usability issues were revealed. They can be divided into interaction style issues and issues that are specific for the MUST tourist guide. The MUST guide specific issues were mainly related to buttons, feedback, prompts, the way selected objects were highlighted, and the location of the POIs on the screen. Most of the problems can be solved rather easily. The comments from the experts gave helpful advice to improve the graphical interface and button-design for the second version of the demonstrator that will be used for the user evaluation experiments.
Almost all experts agreed that without some initial training and instruction, the users would probably not use a simultaneous multimodal interaction style. They also believed that the users will probably be able to use such an interaction style with small cognitive effort, once they are aware of the systems capabilities. This is also supported by our observations of the experts behaviour during the explorative phase
With the present lack of multimodal applications for the general public, there is a need to introduce the capabilities of simultaneous coordinated interaction explicitly before customers start using the new products. According to the experts a short video or animation would be suitable for this purpose. This issue will be studied during the user experiments that will be carried out in September. The introduction that is given to the users before they start to use the tourist guide will be the main parameter in this experiment. Then we will also gain more information on how naïve users benefit from adding the simultaneous coordinated actions in a multimodal tourist guide. In our demonstrator it is not necessary for the user to input several modalities simultaneously. The choice of sequential/simultaneous mode is controlled by the user.
Another issue pointed out by the experts is the importance of a well-designed help mechanism in speech-centric user initiative information services. In these services it is difficult for the system to convey information about its capabilities and limitations (Walker and Passonneau, 2001).

5 Conclusion and further work

The aim of MUST is to provide Telecom Operators with useful information on multimodal services. We have built a stable, real-time multimodal demonstrator using standard components without too much effort.
The first version was evaluated by human-factor experts. One of the main conclusions was that naïve users will need instructions before being able to benefit from a simultaneous coordinated multimodal interaction. Once aware of the systems capabilities they should be able to use the system with small cognitive effort. This will be studied more in forthcoming user experiments. Another issue we will study in this experiment is the timing of the input, especially when deictic expressions are used.

References

Azevedo, J., Beires N. (2001) *InoVox - MultiService Platform Datasheet*, Portugal Telecom Inovação.

Boualem, M. and Filoche, P. (n.y.) Question-Answering System in Natural Language on Internet and Intranets, *YET2 marketplace*, http://www.yet2.com/

Boves, L., and Den Os, E. (Eds.) (2002) *Multimodal services – a MUST for UMTS*. http://www.eurescom.de/public/projectresults/P1100-series/P1104-D1.asp

Cheyer, A. and Julia, L. (1998) Multimodal Maps: An agent-based approach. In: H. Bunt, Beun, Borghuis (Eds) *Multimodal Human-computer communication*, Springer Verlag, pp. 111-121.

EURESCOM (2002) *Multimodal and Multilingual Services for Small Mobile Terminals.* Heidelberg, EURESCOM Brochure Series.

Kehler, A., Martin, J.-C., Cheyer, A. Julia, L., Hobbs, J. and Bear, J. (1998) On representing salience and reference in multimodal human-computer interaction. *AAAI'98, Representations for multimodal human-computer interaction*, Madison, pp. 33-39.

Knudsen, J.E., Johansen, F.T. and Rugelbak, J. (2000) *Tabulib 1.4 Reference Manual*, Telenor R&D scientific document N-36/2000.

Martin, J.-C. Julia, L. and Cheyer, A. (1998) A theoretical framework for multimodal user studies, *CMC-'98*, pp. 104-110.

Nielsen, J. and Mack, R.L. (eds), (1994) *Usability Inspection Methods*, Jon Wiley & Sons, Inc

Oviatt, S. (1999) Ten Myths of Multimodal Interaction, *Communications of the ACM*. Vol. 42, No. 11, pp. 74-81.

Oviatt, S. et al. (2000) Designing the user interface for multimodal speech and gesture applications: state-of-the-art systems and research directions for 2000 and beyond. In: J. Carroll (ed) *Human-computer interaction in the new millennium.* Boston: Addison-Wesley Press.

Oviatt, S. & Cohen, P. (2000) Multimodal Interfaces That Process What Comes Naturally, *Communications of the ACM*, Vol. 43, No. 3, pp. 45-53.

Oviatt, S. L., DeAngeli, A. & Kuhn, K. (1997) Integration and synchronization of input modes during multimodal human-computer interaction, *Proc. Conf. on Human Factors in Computing Systems: CHI '97*, New York, ACM Press, 415-422.

Wahlster, W., Reithinger, N., and Blocher, A. (2001) SmartKom: Multimodal Communication with a Life-Like Character, *EUROSPEECH-2001*, Aalborg, Denmark, pp 1547–1550.

Walker, M. A., and Passonneau, R. (2001) DATE: A Dialog Act Tagging Scheme for Evaluation of Spoken Dialog Systems. *Human Language Technology Conference*. San Diego, March 2001.

Wyard, P. and Churcher, G. (1999) The MUeSLI multimodal 3D retail system, *Proc. ESCA Workshop on Interactive Dialogue in Multimodal Systems*, Kloster Irsee, pp. 17-20.

How to relate User Satisfaction and System Performance in Multimodal Dialogue Situations? - a Graphical Approach

Nicole Beringer, Sebastian Hans, Katerina Louka, Jie Tang
Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
Schellingstr. 3, 80799 München
{beringer, hanss, kalo, tang}@phonetik.uni-muenchen.de

Abstract

The goal of this paper is to present in detail a graphical evaluation tool for multimodal dialogue systems which is used to compare the users' satisfaction with the system's technical performance quasi objectively. It is also used to define weights for the calculation of overall system performance (user satisfaction & technical performance) of multimodal dialogue systems.

1 Keywords

Multimodal dialogue systems, end-to-end evaluation, graphical evaluation tool, evaluation framework, SmartKom

2 Introduction

When evaluating multimodal dialogue systems the evaluators have many problems to solve and only partly can transfer established methods from spoken dialogue system evaluations (see (Beringer et al., 2002b) for further details) such as the PARADISE framework (Walker et al., 1997). In the end-to-end evaluation of the multimodal dialogue system SmartKom the evaluators have to deal with the innovative character of multimodality. Therefore, we developed a new evaluation framework for multimodal dialogue systems, PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation) (Beringer et al., 2002a) since established methods cannot be transferred unambiguously from monomodal frameworks like PARADISE.

This new framework combines established methods from spoken dialogue evaluations and takes into account new methods to handle multimodal characteristics like gestural input combined with speech input, graphical vs. speech output or userstate information via facial expression of the user.

One of the features of spoken dialogue evaluation frameworks, namely the independence of systems and tasks by weighting objectively measured qualities and quantities of the system by subjective user satisfaction has been implemented in PROMISE as well.

To obtain weights, user satisfaction can be directly correlated with the objective quality and quantity measures (further referred to as costs) via Pearson correlation (see (Beringer et al., 2002a) for further details). But not all costs can be given corresponding questions in usability questionnaires. However, to obtain a normalization over differing systems, scenarios and tasks we have to weight them somehow.

To handle this problem, the evaluators have to objectively compare with costs the recorded passages of the dialogues in question.

This was the motivation to develop a graphical approach to check and relate user satisfaction and system performance objectively.

The paper is structured as follows: section 3 describes briefly the function of the multimodal SmartKom dialogue system which has to be evaluated. In section 4 we give a general outline of PROMISE. Section 5 describes the possibility to define weights to normalize over systems, scenarios and tasks. The requirements and characteristics of the graphical evaluation tool as well as some positive

side effects will be presented in section 5. The paper finishes with a short summary and outlines our future work.

3 The SmartKom project

In the SmartKom project, an intelligent computer-user interface is being developed which deals with various kinds of oral or physical input. Potential benefits of SmartKom include the ease of use and the naturalness of the man-machine interaction due to multimodal input and output. However, a very critical obstacle to progress in this area is the lack of a general methodology for evaluating and comparing the performance of the three possible scenarios provided by SmartKom:

- SmartKom Home/Office to communicate and operate machines at home (e.g. TV, workstation, radio),
- SmartKom Public to have a public access to public services, and
- SmartKom Mobile as a mobile assistant.

The system understands input in the form of natural speech as well as in the form of gestures. In order to "react" properly to the intentions of the user, the emotional status is analyzed via the facial expression and the prosody of speech. One of the requirements of the project is to develop new modalities and new techniques.

4 General Outline of the PROMISE Framework

PROMISE (**Pro**cedure for **M**ultimodal **I**nteractive **S**ystem **E**valuation) is an extended evaluation framework for multimodal dialogue systems (Beringer et al., 2002a), where we aimed to solve the problems of scoring multimodal inputs and outputs, weighting the different recognition modalities and how to deal with non-directed task definitions and the resulting, potentially uncompleted tasks by the users.

Advantages of established methods like abstracting from systems, scenarios and tasks are included in PROMISE as well as modality specific measures. The latter, of course, is the basis of a number of problems. The most challenging is described in the following subsection.

4.1 Scoring multimodal inputs and outputs

In contrast to interactive monomodal spoken dialogue systems, multimodal dialogue systems consist of several equivalent technologies which are functionally similar to each other. In other words, multimodal dialogue systems are based on many component technologies like speech recognition, gesture recognition, recognition of emotional states, text-to-speech, natural language understanding, natural language generation, generation of graphical presentation, synchronization of speech and graphics and database query languages. Taking the example of recognition, the different modalities can and will interfere with each other.

To evaluate interfering functionalities multimodal inputs have to be identified and further processed on according to the "total or nothing" principle[1].

Another problem in scoring multimodal input is how to estimate the accuracy of different recognizers. I.e., in talking about speech recognition, we have to deal with a very complicated pattern match, whereas gesture recognition has a limited set of recognizible gestures which can be found in a given coordinate plane.

4.2 Dealing with non-directed task definitions

Apart from the scoring problems PROMISE also offers a solution for handling non-directed task definitions. In contrast to task requirements, where the user has to check several functions of a system in a defined order, SmartKom offers a variety of different functions which can be combined in any order to get the wanted information. Therefore we had to define more dynamic "keys" (a PARADISE

[1] Only those combined inputs get access to the evaluation which are not ambiguous in their content.

term) to extract different superordinate concepts depending on the task at hand referred to below as "information chunks" or "bits" . These "information chunks" are carefully selected, categorized and weighted by hand before the tests start to compute, normalize and compare across different tasks and scenarios.

The number of information chunks can vary within one completed task, but it must define a task unambiguously in order to complete it successfully.

4.3 Abstraction from systems, scenarios and tasks

The requirement of abstracting from systems, scenarios and tasks is assumed to be necessary not only for spoken dialogue systems evaluation but also for the evaluation of multimodal dialogue systems. This can be done by weighting successfully completed tasks with correlation coefficient of the Pearson correlation between user satisfaction values and task completion. Due to the more dynamic task definition in SmartKom, PROMISE allows only two values for task success:
$\tau_j = +1$: task success;
$\tau_j = -1$: task failure;
where j is the index of the corresponding tests.

To abstract from dialogues, PROMISE uses the mean value $\bar{\tau}$.

To compute the system performance we have to normalize over the cost functions via a z-scored normalization function: $\mathcal{N}(c_i) = \frac{c_i - \bar{c_i}}{\sigma_{c_i}}$, where c_i i-th cost, σ_{c_i} variance of c_i, $\bar{c_i}$ the mean of c_i.

5 The SmartKom Graphical Evaluation Tool

5.1 Requirements

For the SmartKom evaluation we designed a graphical evaluation tool which gives the possibility to compare user satisfaction values (out of a usability questionnaire) about a given functionality with the corresponding quality and quantity measures (objectively measurable technical evaluation) of the respective dialogue. A human evaluator checks both parts and decides which of the two is more likely or if both are equally likely. To get a better idea of the dialogue (s)he has to evaluate, the video or parts of it can be displayed as well - either by choosing the start and endpoint by hand or by clicking offered timestamps (see section 5.4 below). For evaluators' comments we must also provide a text field. The quasi objective scoring is written in an SQL database described below in section 5.3.

Table 1 gives an overview of the cost and usability pairs we defined for SmartKom.

The tool offers a defined course of evaluation results. This balances individual differences both of the evaluators and of the users. Making the data available to a number of evaluators by offering a platform independent tool the evaluation can be done highly objectively.

5.2 Video display

For the linux version of our Graphical Evaluation Tool we had to provide a video player which allows starting, stopping and repeating parts of the filmed dialogues. Video playback is implemented using a specially hacked up copy of XAnim[2]. XAnim is embedded into the interface by means of two wrappers: embedded_xanim and XAnimRemoteControl.

5.2.1 Modifiactions to XAnim

XAnim had to be modified to allow viewing of single parts of a video and to allow rapid stopping of replay. It now sports four new command line options, to specify start and end times in frames or milliseconds.

5.2.2 embedded_xanim

embedded_xanim is an interactive interface for XAnim written in C.
It displays a window with the specified size at the specified position. It then goes into a loop reading lines from standard input which are interpreted as XAnim command line arguments and an optional window title. XAnim is run with these arguments, thus playing the desired video. Furthermore, the video can be stopped at any point during replay using the 'stop' command.

[2]XAnim 2.80.1, see
http://xanim.va.pubnix.com/home.html

Quality and quantity measures	usability question
Transaction success Task complexity	The task was easy to solve
Misunderstanding of input Offtalk	SmartKom has understood my input
Misunderstanding of output	SmartKom can easily be understood
Semantical, syntactical correctness Incremental compatibility	SmartKom has answered properly in most cases
Mean system response time Mean user response time Timeout	The speed of the system was acceptable for each situation I always knew what to say
Acc. gesture recognition	The gestural input was successful
Acc. ASR	The speech input was successful
Dialogue complexity	SmartKom worked as assumed SmartKom reacted quickly to my input SmartKom is easy to handle
Percentage of appropriate/inappropriate system directive diagnostic utterances Percentage of explicit recovery answers repetitions No. of ambiguities	SmartKom offered an adequate amount of high quality information SmartKom is easy to handle
Diagnostic error messages Rejections Timeout	SmartKom needs input only once to successfully complete a task
Help-analyzer	SmartKom offers adequate help
Output complexity (display)	The display is clearly designed
Mean elapsed time Task completion time Dialogue elapsed time	SmartKom reacted fast to my input
Duration of speech input Duration of ASR	SmartKom reacted fast to speech input
Duration of gestural input Duration of gesture recognition	SmartKom reacted fast to gestural input
BargeIn Cancel	SmartKom allows interrupts
Dialogue complexity	Was the task difficult?
Gesture turns Ways of interaction	input via graphical display
Display turns	output via graphical display
Speech input	speech input
Speech synthesis (synchronicity)	speech output
N-way communication Ways of interaction Error rate of questions Input complexity	Possibility to interact in a quasi-human way with SmartKom

Table continues next page

Quality and quantity measures	usability question
Recognition/duration of facial expression Prosodic features	SmartKom reacted towards my emotional state
Synchronicity Graphical output (turns) Cooperativity	How do you score the competence of the agent? Were actions of the persona natural?
Gestural input	Gestural input

Table 1: Quality and quantity measures for the SmartKom evaluation compared with user satisfaction values

5.2.3 XAnimRemoteControl

XAnimRemoteControl is a high-level Java wrapper-class for embedded_xanim.
During initialization, an instance of embedded_xanim is started. Then, arbitrary videofiles can be played with the play()-method, and replay can be stopped with the stop()-method. The quit()-method cleans up and kills embedded_xanim.

5.3 The Back-end - A MySQL Database

The data processed by the tool comes from and goes into a MySQL database.
The questions and answers of the technical evaluation (mainly extracted from the logfiles) are read in from the TE_Frage & TE_Antwort table pair[3]. The results of the ergonomic evaluation (from the questionnaire) are read from a slightly more complicated arrangement of tables, where the questions are split into one or more subquestions. The technical and ergonomic questions are connected by a link table. The result is written into a separate synthesis table.

5.4 The Evaluation GUI

The evaluation process is based on a GUI programm written in Java. It is deliberately selected for its character of platform independence, so the program can be easily executed on the different computer types in the institute.

[3]Name of the cost and the value

5.4.1 Functional description

The graphical interface is composed of 3 sections: a Java window for the evaluation, a XAnim window[4] for the video display and an emacs window for the label display. Figure 1 shows a screenshot of the GUI.

The Java window is divided into 3 boxes of the same size. The left box is displayed for the result of technical evaluation, and the data is imported from the related MySQL database, in which the results of different sessions are located. The right one shows a display for questions and answers of ergonomic evalution related to the technical part. The middle section offers pre-defined decision categories. Comments can be written in plain text in the comment field.
When the GUI is started with a session, an aligned XAnim window with the video file in questions as well as an aligned emacs window including the corresponding annotations is initialized.

5.4.2 XAnim control

During the evaluation process, human evaluators can navigate through the video file by clicking the timestamps in order to see a sepecified part of the video sequence. It is integrated into the Java program, so the evaluator can mark the start and end point in milliseconds to fire an event, which is handled correspondingly in the programm to run the embedded_xanim with the specified segment.

5.4.3 Database connection

Another main component of this tool, which runs in the background, is the database

[4]see the section 5.2 Video display

connection. The database used in the evaluation process is a MySQL database, which includes the essential tables for the whole evaluation(See the section 5.3 for details). The program communicates with the database in two ways:

- In the initialization process of the GUI, the program makes queries to the database to extract the related results of technical and ergonomic evaluation in order to graphically represent them in the provided boxes.
- During the evalution process the evaluator makes decisions and comments. The outgoing results are then immediately inserted into the corresponding table of the database.

5.5 Side Effects

Apart from allocating user satisfaction and system performance via the evaluation GUI the tool offers some positive side effects, namely

- the possibility of annotating userstate and gestural input (Steininger et al., 2001; Steininger et al., 2002)
- the possibility to score multimodal recognition facilities
- the repetition of video sequences

Using a graphical interface not only for evaluation purposes but also for presentation of detailled evaluation results we make it possible for every scenario that the developers get access to all content-related problems, results and protocols within the corresponding evaluation phase. While using the display, they can easily find out by checking the related objectively measured and user satisfaction values on which evaluation part to concentrate (user satisfaction scores or objective costs) to improve the system.

6 Conclusion

The Graphical Evaluation Tool allows a visual representation of dialogues that can be used both for evaluating the system and comparing user satisfaction with objectively measured costs. Human evaluators have to follow the same course of evaluation by the Graphical Evaluation Tool. This balances individual differences both between the evaluators as well as the users.

Apart from the primary functionality of evaluation, our tool offers the possibility of annotating user state and gestural input.

Via the GUI, it is possible to score the different recognition modalities as well.

Finally, the controlled playing of video sequences can be done platform independently due to the Java implementation.

7 Acknowledgements

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the SmartKom project (01IL905E/6).

References

Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, and Uli Tuerk. 2002a. Promise - a procedure for multimodal interactive system evaluation. *Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation (to appear).*

Nicole Beringer, Katerina Louka, Victoria Penide-Lopez, and Uli Tuerk. 2002b. End-to-end evaluation of multimodal dialogue systems - can we transfer established methods? *Proc. of the LREC 2002, Gran Canaria, Spain (to appear).*

S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in smartkom - the coding system. *Springer Gesture Workshop 2001, London (to appear)*, LNAI 2298.

S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. 2002. Development of the user-state conventions for the multimodal corpus in smartkom. *Proc. of the LREC Workshop on Multimodal Resources and Multimodal Systems Evaluation (to appear).*

M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluation spoken dialogue agents. *Annnual Meeting of the Association of Computational Linguistics. ACL.*

Figure 1: Graphical evaluation tool for the SmartKom evaluation

Phone vs. Face-to-Face with Virtual Persons

Timothy Bickmore
MIT Media Lab
20 Ames St., Room E15-320
Cambridge, MA 02139
+1 612 253 7368
bickmore@media.mit.edu

Justine Cassell
MIT Media Lab
20 Ames St., Room E15-315
Cambridge, MA 02139
+1 612 253 4899
justine@media.mit.edu

Abstract

This study compares people's interactions with Embodied Conversational Agents to similar interactions over the phone, and investigates the impact these media have on a wide range of behavioral, task and subjective measures. While the behavioral measures were consistent with previous studies, the subjective measures indicated that the fit of an ECA's persona to the task and style of interaction can overwhelm the effects of media on subjects' assessment of the ECA and the interaction.

Introduction

Social psychologists have compared face-to-face conversation with phone conversation, video-mediated communication and other mediated modalities, showing the effect various media have on psychosocial variables such as interpersonal vs. task-orientation, cooperation, trust, metacognition, person perception, veracity and task outcomes in negotiation and collaborative problem-solving (Rutter, 1987). Studies comparing human-human to human-computer interaction have demonstrated effects on speech disfluency, turn length and frequency, utterance length and interruptions (e.g., Oviatt, 1995). Few studies to date, however, have investigated how interaction with embodied conversational agents (ECAs) compares with these other well-understood modalities.

In this paper we present the results of a study comparing interaction with an embodied conversational agent to interaction with a phone-based dialogue system. This study extends previous work investigating the effects of social dialogue ("small talk") in a real estate sales domain, which demonstrated that social dialogue can have a significant impact on a user's trust of a computer agent (Bickmore and Cassell, 2001). In addition to varying medium (phone vs. embodied) and dialogue style (social dialogue vs. task-only) we also assessed the user's personality along the introversion/extroversion dimension, since extroversion is one indicator of a person's comfort level with face-to-face interaction.

1 Related Work

Work on the development of ECAs, as a distinct field of development, is best summarized in (Cassell, Sullivan et al., 2000). The current study is based on the REA ECA (see Figure 1), a simulated real-estate agent, who uses vision-based gesture recognition, speech recognition, discourse planning, sentence and gesture planning, speech synthesis and animation of a 3D body (Cassell, Bickmore et al., 1999). Some of the other major systems developed to date are Steve (Rickel and Johnson, 1998), the DFKI Persona (Andre, Muller et al., 1996), Olga (Beskow and McGlashan, 1997), and pedagogical agents developed by Lester, et al, (Lester, Stone et al., 1999). These systems vary in their linguistic generativity, input modalities, and task domains, but all aim to engage the user in natural, embodied conversation.

Figure 1. REA

1.2 User Studies on Embodied Conversational Agents

Koda and Maes (Koda and Maes, 1996) and Takeuchi and Naito (Takeuchi and Naito, 1995) studied interfaces with static or animated faces, and found that users rated them to be more engaging and entertaining than functionally equivalent interfaces without a face. Kiesler and Sproull (Kiesler and Sproull, 1997) found that users were more likely to be cooperative with an interface agent when it had a human face (vs. a dog or cartoon dog).

Andre, Rist and Muller found that users rated their animated presentation agent ("PPP Persona") as more entertaining and helpful than an equivalent interface without the agent (Andre, Rist et al., 1998). However, there was no difference in actual performance (comprehension and recall of presented material) in interfaces with the agent vs. interfaces without it.

In a user study of the Gandalf system (Cassell and Thorisson, 1999), users rated the smoothness of the interaction and the agent's language skills significantly higher under test conditions in which Gandalf utilized limited conversational behavior (gaze, turn-taking and beat gesture) than when these behaviors were disabled.

Sproull et al. (Sproull, Subramani et al., 1997) showed that subjects rated a female embodied interface significantly lower in sociability and gave it a significantly more negative social evaluation compared to a text-only interface. Subjects also reported themselves to be more aroused (less relaxed and assured) when interacting with the embodied interface than when interacting with the text interface. They also presented themselves in a more positive light (gave themselves significantly higher scores on social desirability scales) and disclosed less (wrote significantly less and skipped more questions in response to queries by the interface) when interacting with an embodied interface vs. a text-only interface. Men were found to disclose more in the embodied condition and women disclosed more in the text-only condition.

Most of these evaluations have tried to address whether embodiment of a system is useful at all, by including or not including an animated figure. In their survey of user studies on embodied agents, Dehn and van Mulken conclude that there is no "persona effect", that is a general advantage of an interface with an animated agent over one without an animated agent (Dehn and Mulken, 1999). However, they believe that lack of evidence and inconsistencies in the studies performed to date may be attributable to methodological shortcomings and variations in the kinds of animations used, the kinds of comparisons made (control conditions), the specific measures used for the dependent variables, and the task and context of the interaction.

1.3 User Studies on Human-Human vs. Human-Computer Communication

Several studies have shown that people speak differently to a computer than another person, even though there are typically no differences in task outcomes in these evaluations. Hauptmann and Rudnicky (Hauptmann and Rudnicky, 1988) performed one of the first studies in this area. They asked subjects to carry out a simple information-gathering task through a (simulated) natural language speech interface, and compared this with speech to a co-present human in the same task. They found that speech to the simulated computer system was telegraphic and formal, approximating a command language. In particular, when speaking to what they believed to be a computer, subject's utterances used a small vocabulary, often sounding like system commands, with very few task-unrelated utterances, and fewer filled pauses and other disfluencies.

These results were extended in research conducted by Oviatt (Oviatt, 1995; Oviatt, Levow et al., 1998; Oviatt and Cohen, 2000), in which she found that speech to a computer system was characterized by a low rate of disfluencies relative to speech to a co-present human. She also noted that visual feedback has an effect on disfluency: telephone calls have a higher rate of disfluency than co-present dialogue. From these results, it seems that people speak more carefully and less naturally when interacting with a computer.

Boyle and Anderson (Boyle, Anderson et al., 1994) compared pairs of subjects working on a map-based task who were visible to each other with pairs of subjects who were co-present but could not see each other. Although no performance difference was found between the two conditions, when subjects could not see one another, they compensated by giving more verbal feedback and using longer utterances. Their conversation was found to be less smooth than that between mutually visible partners, indicated by more interruptions, and less efficient, as more turns were required to complete the task. The researchers concluded that visual feedback improves the smoothness and efficiency of the interaction, but that we have devices to compensate for this when visibility is restricted.

Daly-Jones, et al. (Daly-Jones, Monk et al., 1998), also failed to find any difference in performance between video-mediated and audio-mediated conversations, although they did find differences in the quality of the interactions (e.g., more explicit questions in audio-only condition).

Whittaker and O'Conaill (Whittaker and O'Conaill, 1997) survey the results of several studies which compared video-mediated communication with audio-only communication and concluded that the visual channel does not significantly impact performance outcomes in task-oriented collaborations, although it does affect social and affective dimensions of communication. Comparing video-mediated communication to face-to-face and audio-only conversations, they also found that speakers used more formal turn-taking techniques in the video condition even though users reported that they perceived many benefits to video conferencing relative to the audio-only mode.

1.4 Trait-based Variation in User Responses

Several studies have shown that users react differently to social agents based on their own personality and other dispositional traits. For example, Reeves and Nass have shown that users like agents that match their own personality (on the introversion/ extraversion dimension) more than those which do not, regardless of whether the personality is portrayed through text or speech (Reeves and Nass, 1996; Nass and Lee, 2000). Resnick and Lammers showed that in order to change user behavior via corrective error messages, the messages should have different degrees of "humanness" depending on whether the user has high or low self-esteem ("computer-ese" messages should be used with low self-esteem users, while "human-like" messages should be used with high-esteem users) (Resnick and Lammers, 1985). Rickenberg and Reeves showed that different types of animated agents affected the anxiety level of users differentially as a function of whether users tended towards internal or external locus of control [20].

2. Experimental Methods

This was a multivariate, multiple-factor, between-subjects experimental design, involving 58 subjects (69% male and 31% female).

2.1 Apparatus

One wall of the experiment room was a rear-projection screen. In the EMBODIED condition Rea appeared life-sized on the screen, in front of the 3D virtual apartments she showed, and her synthetic voice was played through two speakers on the floor in front of the screen. In the PHONE condition only the 3D virtual apartments were displayed and subjects interacted with Rea over an ordinary telephone placed on a table in front of the screen.

For the purpose of this experiment, Rea was controlled via a wizard-of-oz setup on another computer positioned behind the projection screen. The interaction script included verbal and nonverbal behavior specifications for Rea (e.g., gesture and gaze commands as well as speech), and embedded commands describing when different rooms in the virtual apartments should be shown. Three pieces of information obtained from the user during the interview were entered into the control system by the wizard: the city the subject wanted to live in; the number of bedrooms s/he wanted; and how much s/he was willing to spend. The first apartment shown was in the specified city, but had twice as many bedrooms as the subject requested and cost twice as much as s/he could afford (they were also told the price was "firm"). The second apartment shown was in the specified city, had the exact number of bedrooms requested, but cost 50% more than the subject could afford (but this time, the subject was told that the price was "negotiable"). The scripts for the TASK and SOCIAL conditions were identical, except that the SOCIAL script had additional small talk utterances added to it, as described in (Bickmore and Cassell, 2001). The part of the script governing the dialogue from the showing of the second apartment through the end of the interaction was identical in both conditions.

Procedure. Subjects were told that they would be interacting with Rea, who played the role of a real

estate agent and could show them apartments she had for rent. They were told that they were to play the role of someone looking for an apartment in the Boston area. In both conditions subjects were told that they could talk to Rea "just like you would to another person".

2.2 Measures

Subjective evaluations of Rea -- including how friendly, credible, lifelike, warm, competent, reliable, efficient, informed, knowledgeable and intelligent she was -- were measured by single items on nine-point Likert scales. Evaluations of the interaction--including how tedious, involving, enjoyable, natural, satisfying, fun, engaging, comfortable and successful it was--were also measured on nine-point Likert scales. Evaluation of how well subjects felt they knew Rea, how well she knew and understood them and how close they felt to her were measured in the same manner.

Liking of REA was an index composed of three items--how likeable and pleasant Rea was and how much subjects liked her--measured items on nine-point Likert scales (Cronbach's alpha = .87).

Amount Willing to Pay was computed as follows. During the interview, Rea asked subjects how much they were able to pay for an apartment; subjects' responses were entered as $X per month. Rea then offered the second apartment for $Y (where Y = 1.5 X), and mentioned that the price was negotiable. On the questionnaire, subjects were asked how much they would be willing to pay for the second apartment, and this was encoded as Z. The task measure used was (Z - X) / (Y - X), which varies from 0% if the user did not budge from their original requested price, to 100% if they offered the full asking price.

Trust was measured by a standardized trust scale (Wheeless and Grotz, 1977) (alpha = .93).

Given literature on the relationship between user personality and preference for computer behavior, we were concerned that subjects might respond differentially based on predisposition. Thus, we also included composite measures for introversion and extroversion on the questionnaire.

Extrovertedness was an index composed of seven Wiggins (Wiggins, 1979) extrovert adjective items: Cheerful, Enthusiastic, Extroverted, Jovial, Outgoing, and Perky. It was used for assessment of the subject's personality (alpha = .87).

Introvertedness was an index composed of seven Wiggins (Wiggins, 1979) introvert adjective items: Bashful, Introverted, Inward, Shy, Undemonstrative, Unrevealing, and Unsparkling. It was used for assessment of the subject's personality (alpha = .84).

Behavioral Measures

Rates of speech disfluency (as defined in Oviatt, 1995) and utterance length were coded from the video data.

Observation of the videotaped data made it clear that some subjects took the initiative in the conversation, while others allowed Rea to lead. Unfortunately, Rea is not yet able to deal with user-initiated talk, and so user initiative often led to Rea interrupting the speaker. To assess the effect of this phenomenon, we therefore divided subjects into *PASSIVE* (below the mean on number of user-initiated utterances) and *ACTIVE* (above the mean on number of user-initiated utterances). To our surprise, these measures turned out to be independent of introversion/extroversion (Pearson r=0.042), and to not be predicted by these latter variables.

3. Results

Full factorial single measure ANOVAs were run, with SOCIALITY (Task vs. Social), PERSONALITY OF SUBJECT (Introvert vs. Extrovert), MEDIUM (Phone vs. Embodied) and INITIATION (Active vs. Passive) as independent variables.

3.1. Subjective Assessments of Rea

In looking at the questionnaire data we find that subjects seemed to feel more comfortable interacting with Rea over the phone than face-to-face. Thus, subjects in the phone condition felt that they knew Rea better ($F=5.02$; $p<.05$), liked her more ($F=4.70$; $p<.05$), felt closer to her ($F=13.37$; $p<.001$), felt more comfortable with the interaction ($F=3.59$; $p<.07$), and thought Rea was more friendly ($F=8.65$;$p<.005$), warm ($F=6.72$; $p<.05$), informed ($F=5.73$; $p<.05$), and knowledgeable ($F=3.86$; $p<.06$) than those in the embodied condition.

However, in the remainder of the results section, as we look more closely at different users, different kinds of dialogue styles, and users' actual behaviour, a more complicated picture emerges. Subjects felt that Rea knew them ($F=3.95$; $p<.06$) and understood them ($F=7.13$; $p<.05$) better when she used task-only dialogue face-to-face; these trends were reversed for phone-based interactions. Task-only dialogue was more fun ($F=3.36$; $p<.08$) and less tedious ($F=8.77$; $p<.005$; see Figure 2) when embodied, while social dialogue was more fun and less tedious on the phone.

That is, in the face-to-face condition, subjects preferred Rea to simply "get down to business."

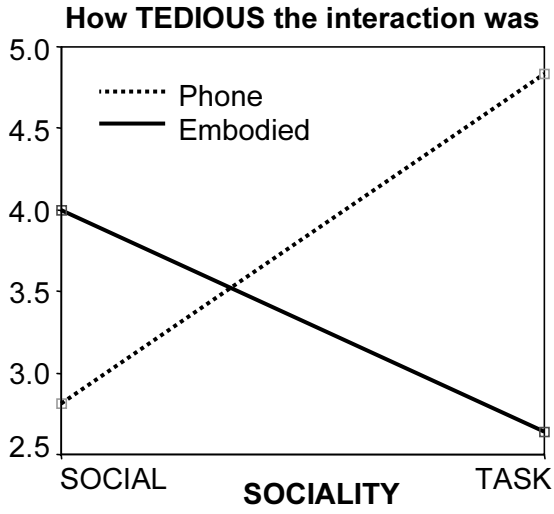


Figure 2. Ratings of TEDIOUS

These results may be telling us that Rea's nonverbal behavior inadvertently projected an unfriendly, introverted personality that was especially inappropriate for social dialogue. Rea's smiles are limited to those related to the ends of turns, and at the time of this experiment, she did not have a model of immediacy or other nonverbal cues for liking and warmth typical of social interaction (Argyle, 1988). According to Whittaker and O'Connail (Whittaker and O'Conaill, 1993), nonverbal information is especially crucial in interactions involving affective cues, such as negotiation or relational dialogue, and less important in purely problem-solving tasks. This interpretation of the results is backed up by comments such as this response from a subject in the face-to-face social condition:

> The only problem was how she would respond. She would pause then just say "OK", or "Yes". Also when she looked to the side and then back before saying something was a little bit unnatural.

This may explain why subjects preferred task interactions face-to-face, while on the phone Rea's social dialogue had its intended effect of making subjects feel that they knew REA better, that she understood them better, and that the experience was more fun and less tedious.

In our earlier study, looking only at an embodied interface, we reported that extroverts trusted the system more when it engaged in small talk, while introverts were not affected by the use of small talk (Bickmore and Cassell, 2001). In the current study, these results were re-confirmed, but only in the embodied interaction; that is, a three-way interaction between SOCIALITY, PERSONALITY and MEDIUM ($F=3.96$; $p<.06$) indicated that extroverts trusted Rea more when she used social dialogue in embodied interactions, but there was essentially no effect of user's personality and social dialogue on trust in phone interactions. Further analysis of the data indicated that this result derived from the substantial difference between introverts and extroverts in the face-to-face task-only condition. Introverts trusted her significantly more in the face-to-face task-only condition than in the other conditions ($p<.03$), while extroverts trusted her significantly less in this condition than in the other conditions ($p.<01$).

In light of these new observations, our earlier results indicating that social dialogue leads to increased trust (for extroverts at least) needs to be revised. This further analysis indicates that the effects we observed may be due to the attraction of a computer displaying similar personality characteristics, rather than the process of trust-building. In the face-to-face, task-only condition both verbal and nonverbal channels were clearly indicating that Rea was an introvert (also supported by the comments that REA's gaze-away behavior was too frequent, an indication of introversion (Wilson, 1977)), and in this condition we find the introverts trusting more, and extroverts trusting less. In all other conditions, the personality cues are either conflicting (a mismatch between verbal and nonverbal behavior has been demonstrated to be disconcerting to users (Nass, Isbister et al., 2000)) or only one channel of cues is available (i.e. on the phone), yielding trust ratings that are close to the overall mean.

There was, nevertheless, a preference by extroverts for social dialogue as demonstrated by the fact that, overall, extroverts liked Rea more when she used social dialogue, while introverts liked her more when she only talked about the task ($F=8.09$; $p<.01$).

Passive subjects felt more comfortable interacting with Rea than active subjects did, regardless of whether the interaction was face-to-face or on the phone, or whether Rea used social dialogue or not. Passive subjects said that they enjoyed the interaction more ($F=4.47$; $p<.05$), felt it was more successful ($F=6.04$; $p<.05$) and liked Rea more ($F=3.24$; $p<.08$), and that Rea was more intelligent ($F=3.40$; $p<.08$),

and knew them better (F=3.42; p<.08) than active subjects. These differences may be explained by the fixed-initiative dialogue model used in the WOZ script. Rea's interaction was designed for passive users--there was very little capability in the interaction script to respond to unanticipated user questions or statements--and user initiation attempts were typically met with uncooperative system responses or interruptions. But, given the choice between phone and face-to-face, passive users preferred to interact with Rea face-to-face: they rated her as more friendly (F=3.56; p<.07) and informed (F=6.30; p<.05) in this condition. Passive users also found the phone to be more tedious, while active users also found the phone to be less tedious (F=5.15; p<.05). Active users may have found the face-to-face condition particularly frustrating since processing delays may have led to the perception that the floor was open (inviting an initiation attempt), when in fact the wizard had already instructed Rea to produce her next utterance.

However, when interacting on the phone, active users differed from passive users in that active users felt she was more reliable when using social dialogue and passive users felt she was more reliable when using task-only dialogue. When interacting face-to-face with Rea, there was no such distinction between active and passive users (F=4.67; p<.05).

3.2. Effects on Task Measure

One of the most tantalizing results obtained is that extroverts were willing to pay more for the same apartment in the embodied condition, while introverts were willing to pay more over the phone (F=3.41; p<.08), as shown in Figure 3.

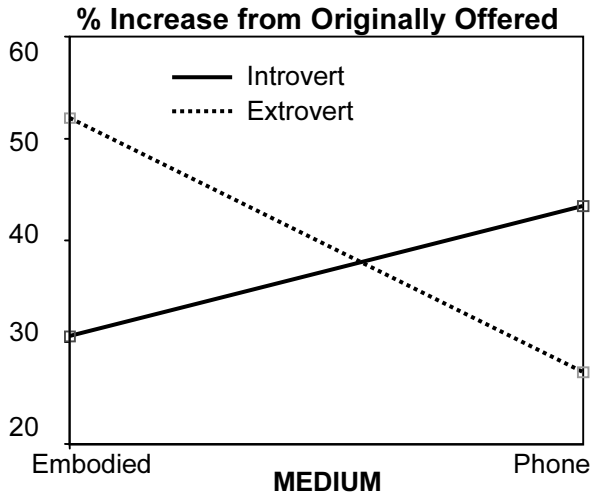


Figure 3. Amount Subjects Were Willing to Pay

While potentially very significant, this finding is a little difficult to explain, especially given that trust did not seem to play a role in the evaluation. Perhaps, since we asked our subjects to simply play the role of someone looking for an apartment, and given that the apartments displayed were cartoon renditions, the subjects may not have felt personally invested in the outcome, and thus may have been more likely to be persuaded by associative factors like the perceived liking and credibility of Rea. In fact, trust has been shown to not play a role in persuasion when "peripheral route" decisions are made, which is the case when the outcome is not of personal significance (Petty and Wegener, 1998). Further, extroverts are not only more sociable, but more impulsive than introverts (Wilson, 1977), and impulse buying is governed primarily by novelty (Onkvisit and Shaw, 1994). Extroverts did rate face-to-face interaction as more engaging than phone-based interaction (though not at a level of statistical significance), while introverts rated phone-based interactions as more engaging, providing some support for this explanation. It is also possible that this measure tells us more about subjects' assessment of the house than of the realtor. In future experiments we may ask more directly whether the subject felt that the realtor was asking a fair price.

3.3. Gender Effects

Women felt that Rea was more efficient (F=5.61; p<.05) and reliable (F=4.99; p<.05) in the embodied condition than when interacting with her over the phone, while men felt that she was more efficient and reliable by phone. Of course, Rea has a female body and a female voice and so in order to have a clearer picture of the meaning of these results, a similar study would need to be carried out with a male realtor.

3.4. Effects on Behavioral Measures

Although subjects' beliefs about Rea and about the interaction are important, it is at least equally important to look at how subjects *act*, independent of their conscious beliefs.

In this context we examined subjects' disfluencies when speaking with Rea. Remember that disfluency can be a measure of naturalness – human-human conversation demonstrates *more* disfluency than does human-computer communication. The rates of speech disfluencies (per 100 words) are shown in Table 1. Comparing these to results from previous studies (see

Table 2) indicates that interactions with REA were more similar to human-human conversation than to human-computer interaction. When asked if he was interacting with a computer or a person, one subject replied "A computer-person I guess. It was a lot like a human."

	Embodied	Phone	Overall
Disfluencies	4.83	6.73	5.83

Table 1. Speech Disfluencies per 100 Words

Human-human speech	
Two-person telephone call	8.83
Two-person face-to-face dialogue	5.5
Human-computer speech	
Unconstrained computer interaction	1.80
Structured computer interaction	0.83

Table 2. Speech Disfluencies per 100 Words for Different Types of Human-Human and Simulated Human-Computer Interactions (adapted from (Oviatt, 1995))

There were no significant differences in utterance length (MLU) across any of the conditions.
The behavioral measures indicate that, with respect to speech disfluency rates, talking to REA is more like talking to a person than talking to a computer.

Once again, there were significant effects of MEDIA, SOCIALITY and PERSONALITY on disfluency rate ($F=7.09$; $p<.05$), such that disfluency rates were higher in TASK than SOCIAL, higher overall for INTROVERTs than EXTROVERTs, higher for EXTROVERTs on the PHONE, and higher for INTROVERTs in EMBODIED condition. These effects on disfluency rates are consistent with the secondary hypothesis that the primary driver on disfluency is cognitive load, once the length of the utterance is controlled for (Oviatt, 1995). Given our results, this hypothesis would indicate that social dialogue requires lower cognitive load than task-oriented dialogue, that conversation requires a higher cognitive load on introverts than extraverts, that talking on the phone is more demanding than talking face-to-face for extraverts, and that talking face-to-face is more demanding than talking on the phone for introverts, all of which seem reasonable.

4. Conclusion

The complex results of this study give us hope for the future of embodied conversational agents, but also a clear roadmap for future research. In terms of their behaviour with Rea, users demonstrated that they treat conversation with her more like human-human conversation than like human-computer conversation. Their verbal disfluencies are the mark of unplanned speech, of a conversational style. However, in terms of their assessment of her abilities, this did not mean that users saw Rea through rose-colored glasses. They were clear about the necessity not only to embody the interaction, but to design every aspect of the embodiment in the service of the same interaction. That is, face-to-face conversations with ECAs must demonstrate the same quick timing of nonverbal behaviors as humans (not an easy task, given the state of the technologies we employ). In addition, the persona and nonverbal behavior of an ECA must be carefully designed to match the task, a conversational style, and user expectations. And finally, as computers begin to resemble humans, the bar of user expectations is raised: people expect that Rea will hold up her end of the conversation, including dealing with interruptions by active users.
We have begun to demonstrate the feasibility of embodied interfaces. Now it is time to design ECAs that people wish to spend time with, and that are able to use their bodies for conversational tasks for which human face-to-face interaction is unparalleled, such as social dialogue, initial business meetings, and negotiation.

Acknowledgements

Thanks to Ian Gouldstone, Jennifer Smith and Elisabeth Sylvan for help in conducting the experiment and analyzing data, and to the rest of the Gesture and Narrative Language Group for their help and support.

References

Andre, E., J. Muller, et al. (1996). The PPP Persona: A Multipurpose Animated Presentation Agent. Advanced Visual Interfaces, Palermo, Italy, ACM Press.

Andre, E., T. Rist, et al. (1998). Integrating Reactive and Scripted Behaviors in a Life-Like Presentation Agent. AGENTS '98, Minneapolis, MN.

Argyle, M. (1988). Bodily Communication. New York, Methuen & Co. Ltd.

Beskow, J. and S. McGlashan (1997). Olga: A Conversational Agent with Gestures. IJCAI'97 workshop on Animated Interface Agents - Making them Intelligent, Nagoya, Japan.

Bickmore, T. W. and J. Cassell (2001). Relational Agents: A Model and Implementation of Building User Trust. CHI 2001, Seattle, WA.

Boyle, E., A. Anderson, et al. (1994). "The Effects of Visibility in a Cooperative Problem Solving Task." Language and Speech **37**(1): 1-20.

Cassell, J., T. Bickmore, et al. (1999). Embodiment in Conversational Interfaces: Rea. CHI 99, Pittsburgh, PA, ACM.

Cassell, J., J. Sullivan, et al. (2000). Embodied Conversational Agents. Cambridge, MIT Press.

Cassell, J. and K. R. Thorisson (1999). "The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents." Applied Artificial Intelligence **13**: 519-538.

Daly-Jones, O., A. Monk, et al. (1998). "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus." International Journal of Human-Computer Studies **49**(1): 21-58.

Dehn, D. M. and S. v. Mulken (1999). The Impact of Animated Interface Agents: A Review of Empirical Research. Saarbrucken, Germany, University of Saarland.

Hauptmann, A. G. and A. I. Rudnicky (1988). "Talking to computers: an empirical investigation." International Journal of Man-Machine Studies, **8**(6): 583 - 604.

Kiesler, S. and L. Sproull (1997). 'Social' Human-Computer Interaction. Human Values and the Design of Computer Technology. B. Friedman. Stanford, CA, CSLI Publications**:** 191-199.

Koda, T. and P. Maes (1996). Agents with Faces: The Effects of Personification of Agents. IEEE Robot-Human Communication '96, Tsukuba, Japan.

Lester, J. C., B. Stone, et al. (1999). "Lifelike Pedagogical agents for Mixed-Initiative Problem Solving in Constuctivist Learning Environments." User Modeling and User-Adapted Interaction **9**(1-2): 1-44.

Nass, C., K. Isbister, et al. (2000). Truth is Beauty: Researching Embodied Conversational Agents. Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost and E. Churchill. Cambridge, MA, MIT Press**:** 374-402.

Nass, C. and K. Lee (2000). Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. CHI 2000, The Hague, Amsterdam, ACM Press.

Onkvisit, S. and J. J. Shaw (1994). Consumer Behavior: Strategy and Analysis. New York, Macmillan College Publishing Company.

Oviatt, S. (1995). "Predicting spoken disfluencies during human-computer interaction." Computer Speech and Language **9**: 19-35.

Oviatt, S. and P. Cohen (2000). "Multimodal Interfaces That Process What Comes Naturally." Communications of the ACM **43**(3): 45-53.

Oviatt, S., G.-A. Levow, et al. (1998). "Modeling global and focal hyperarticulation during human-computer error resolution." Journal of the Acoustical Society of America **104**(5): 3080-3091.

Petty, R. and D. Wegener (1998). Attitude Change: Multiple Roles for Persuasion Variables. The Handbook of Social Psychology. D. Gilbert, S. Fiske and G. Lindzey. New York, McGraw-Hill**:** 323-390.

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

Resnick, P. V. and H. B. Lammers (1985). "The Influence of Self-esteem on Cognitive Responses to Machine-Like Versus Human-Like Computer Feedback." The Journal of Social Psychology **125**(6): 761--769.

Rickel, J. and W. L. Johnson (1998). "Task-Oriented Dialogs with Animated Agents in Virtual Reality." Proceedings of the 1st Workshop on Embodied Conversational Characters: 39-46.

Rickenberg, R. and B. Reeves (2000). The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. CHI 2000, The Hague, Amsterdam.

Rutter, D. R. (1987). Communicating by Telephone. New York, Pergamon Press.

Sproull, L., M. Subramani, et al. (1997). When the Interface is a FAce. Human Values and the Design of Computer Technology. B. Friedman. Stanford, CA, CSLI Publications**:** 163-190.

Takeuchi, A. and T. Naito (1995). Situated Facial Displays: Towards Social Interaction. Human Factors in Computing Systems: CHI'95, Denver, CO, ACM Press.

Wheeless, L. and J. Grotz (1977). "The Measurement of Trust and Its Relationship to Self-Disclosure." Human Communication Research **3**(3): 250-257.

Whittaker, S. and B. O'Conaill (1993). An Evaluation of Video Mediated Communication. Conference on Human Factors and Computing Systems - INTERACT/CHI '93, Amsterdam, The Netherlands.

Whittaker, S. and B. O'Conaill (1997). The Role of Vision in Face-to-Face and Mediated Communication. Video-Mediated Communication. K. Finn, A. Sellen and S. Wilbur, Lawrence Erlbaum Associates, Inc.**:** 23-49.

Wiggins, J. (1979). "A psychological taxonomy of trait-descriptive terms." Journal of Personality and Social Psychology **37**(3): 395-412.

Wilson, G. (1977). Introversion/ Extraversion. Personality Variables in Social Behavior. T. Blass. New York, John Wiley & Sons**:** 179-218.

A Dialogue Architecture for Multimodal Control of Robots

Carl BURKE
The MITRE Corporation
7515 Colshire Drive
McLean VA 22102 USA
cburke@mitre.org

Lisa HARPER
The MITRE Corporation
7515 Colshire Drive
McLean VA 22102 USA
lisah@mitre.org

Dan LOEHR
The MITRE Corporation
7515 Colshire Drive
McLean VA 22102 USA
loehr@mitre.org

Abstract

Robots typically execute only pre-programmed, limited instructions. For humans to command teams of semi-autonomous robots in non-trivial, mobile, and dynamically changing tasks, the human-robot interface will need to include several aspects of human-human communication. These aspects include cooperatively detecting and resolving problems, making using of context, and maintaining contexts across multiple conversations. In this paper, we describe the architecture we are developing to support this dialogue system, based on the TRINDIKit framework.

Introduction

Robotics research has recently experienced a surge of interest due to a growing awareness that robots can work collaboratively with humans to perform tasks in situations unsafe for humans. The 1997 Mars Sojourner rover was tasked to act as a "mobile remote geologist" and conducted soil experiments in several different terrains (NASA 1997). Teleoperated robots assisted at the site of the World Trade Center in New York City after the September 11 attack. Robots were able to penetrate into areas of rubble debris in cavities too narrow and dangerous for humans and dogs (Kahney 2001). Finally, the US Government's Defense Advanced Research Projects Agency (DARPA) has invested substantial funding toward a vision in which robots will support future combat systems.

Despite this increased activity in robotics, relatively few advances have been made in the area of human-robot interaction. In a recent Robocup Rescue event, the best contenders in the competition relied upon teleoperation (joystick-style control) by human controllers (Eyler-Walker, p.c.). Though ultimately supervisory control of teams of semi-autonomous robots is a very promising avenue for future research in robot search and rescue, this technological approach does not yet reach the level of competence of teleoperation. Recently, NASA has been concerned with human-machine interaction are commanded by high-level commands rather than sequences of low level commands. A grapefruit-sized Personal Satellite Assistant (PSA) is being developed to operate aboard the Space Shuttle's flight deck. It will navigate using its own navigation sensors, wireless network connections, and propulsion components. Rayner et al. (2000a, 2000b) describe an architecture for a spoken interface with the PSA.

An alternative approach to human-robot interaction by Fong et al (2001) bridges teleoperation with "collaborative control". In this model, humans and robots act as peers exchanging information in dialogue to achieve goals. Instead of controlling the vehicular robot solely by direct (manual) control, the human specifies a set of waypoints that the robot can achieve on its own. One problem observed with waypoint driving is that robots may encounter obstacles for which its vision system is inadequate to assess. In such a circumstance, the robot can query the human about the nature of the obstacle and receive assistance.

In this paper we describe a dialogue architecture we are developing for a Personal Digital Assistant (PDA)-based dialogue interface to a robot, which we plan to extend toward a team-based search and rescue task. Currently, the PDA supports single user, single robot dialogue in a limited navigation and question-answer scenario for visitors to a technology trade show. Using touch gestures and speech, users may ask the robot to guide them to a particular booth, show images from remotely located robots, and answer questions about exhibits at the trade show.

Our primary research interest is the development of a dialogue system architecture robust enough to tolerate continuous operational use, flexible enough for porting to different domains and tasks, and able to support multiple, simultaneous conversations involving one or more humans and one or more cooperative robot entities. The dialogue management architecture we are developing is based on the TRINDIKit (Task oRiented Instructional Dialogue Toolkit) (TRINDI 2002) framework, although we have introduced a number of implementational changes in the reengineering of a TRINDIKit architecture.

1 Original Architecture

Our architecture was first assembled for development of a demonstration system called Mia, the MITRE Information Assistant. Mia is an information kiosk equipped with a touch screen and a microphone, and stocked with information about MITRE's internal research projects for use as a visitors' guide to a MITRE trade show.

Mia was built as a set of independent modules that communicated using SRI's Open Agent Architecture (OAA). The Graphical User Interface (GUI) was written in Tcl/Tk. The GUI handled push-to-talk for the speech recognizer, maintained a text menu of possible user utterances, showed a map of the overall trade show layout with the ability to zoom in on specific rooms, and displayed prerecorded output videos of the animated agent speaking and gesturing. Dialogue management was done with the TRINDIKit system.

TRINDIKit itself provides the basic infrastructure of a dialogue manager. It provides structured data types and the means to define an Information State (IS) from those types, a language for defining the modules of a Dialogue Move Engine (DME), and a language for controlling the application of individual modules to the job of dialogue management. With all TRINDIKit provides, it does not implement a theory of dialogue. For that we used the GoDiS (Gothenburg Dialogue System) (Larsson et al 2000) system, which implements the Questions Under Discussion model in TRINDIKit. We were able to adapt existing GoDiS dialogues to our kiosk domain in a very short time.

In order to integrate TRINDIKit into the kiosk using the OAA, we used TRINDIKit's concurrent mode, which incorporates support for use of the OAA. While this seemed to be a natural choice, and allowed more natural definition of module interactions, it also raised several problems, as discussed below.

1.1 Speed

TRINDIKit in concurrent mode ran very slowly, on a 750 MHz Pentium2 with 384 MB RAM running WindowsNT and no other processes. We believe using the OAA for data transport caused the delays, as a large number of messages were exchanged. Lewin et al (2000:45) report that running GoDiS with TRINDIKit on the OAA yielded a 2-second user utterance to system utterance time, compared to a 0.5 second time when using TRINDIKit's internal agent environment (which is not available for use with non-prolog components). Although modules run independently in concurrent mode, updates to IS were still transmitted to each module individually. Updates were sent whether they were used by that module or not, and all other processing waited until that module finished its work.

1.2 Data Consistency

TRINDIKit does not exercise good controls over asynchronous modifications to IS. At one point we had to build artificial delays into our system to work around these limitations. The dialogue manager we built for Mia was based on GoDiS, which requires very structured turn-taking. In several cases, however, the interactions with the user flowed better if these responses were automatic. Processing was sufficiently slow that our GUI's automatic acknowledgement often arrived and was processed before TRINDIKit was finished cleaning up from the previous utterance. As a result, it was possible to change the IS twice before the DME could respond to one change, and the system lost track of the dialogue state. Consistency of data needs to be assured throughout the design of the system.

1.3 Inconsistent Semantics

We encountered situations where constructs of the GoDiS plan language were interpreted differently depending on the depth of the plan. With the proliferation of small languages im

plemented by different sets of macros, it was difficult to track down bugs in the rules and conversation scripts. This was made more difficult by the nature of Prolog. Clauses that fail do not normally generate any error messages, because failure is a normal aspect of program execution. Unfortunately, database bugs and misspelled names often caused unexpected failures, causing the system to generate either no response or a response that looked reasonable but was in fact incorrect. We feel it's necessary to provide explicit notification of certain kinds of failure, such as failure to find a named variable, failure to find a matching value in a table, and so on.

1.4 Lack of Multimodal Support

Neither TRINDIKit nor GoDiS provides any direct support for multimodal processing. The primary interface driving the development of these systems was language; there is no separation of events by source, no temporal tagging of input events, and no provision for assessing temporal relationships between different inputs.

2 Revised Architecture

We are moving ahead with the design for a dialogue manager for robot control. From our experience with the dialogue manager in Mia, we're convinced of the advantages of a kit-based approach. We feel that TRINDIKit was a good first cut at it, and hope that our efforts will lead to a second, somewhat better iteration.

2.1 Distributed Information State

We've chosen to model all of our module interactions as if they were asynchronous. This provides the cleanest separation of modules, and the cleanest conceptual integration with the asynchronous requirements of robot control. Our approach to solving this problem is to define an explicit interface definition language, which will be used to define every module's interface with the outside world. We explicitly include the information state structure in this interface definition, perhaps as a module in itself. Since TRINDIKit does not include a separate language for specifying module interfaces, we are designing our own. This language is analogous to CORBA Interface Definition Language, but with less concern for the physical implementation.

There are a large number of protocols and infrastructures that have been developed to support communications between agents, each with characteristics optimized for particular tasks or emphasizing desired functionality. We intend to support small standard set of operations that have wide applicability across programming languages and communication protocols.

2.2 Controlled Extensibility

Our interface specifications will need to be translated into specific computer languages before they can be executed. The translation will vary depending on the underlying protocol used to communicate between modules. While we want to support the widest possible audience, we don't want to get bogged down in the construction of translators for every possible set of implementation language and protocol. Our approach is to exploit an existing standard set of translation software, namely XML and XSLT processors such as Xalan. We are specifying a dialect of XML for modules interface definitions, and a small set of templates for realizing interfaces with specific combinations of programming language and protocol. Additional templates can be written as necessary to extend the kit to other languages and protocols without requiring modification of the kit itself.

The same approach extends to the specifications of DME rules, module synchronization and control, and the definition of new "languages" for the kit. We have defined what well-formed formulas look like in our kit's scripting language: what names look like, the types of expressions that are possible, how expressions and statements are aggregated to form programs. What is left unspecified is the exact sequences of expressions that form statements in any particular script language. Those are specified using templates analogous to XML DTDs, which gives us the flexibility to define new constructs as needed.

2.3 Rule Engine

The DME rules in TRINDIKit have strong similarities to rules in expert systems. We plan to implement these rules in both a sequential form, equivalent to the current TRINDIKit, and in an expert system form which may be more efficient. We expect that there will be differences in operating characteristics between those

two styles, and we want to identify and quantify those differences.

One issue we must address in our design is unification. While logic variables are natural for modeling discourse given the history of the field, most languages typically used to implement robot software do not support it directly. Our kit must ensure that sound unification procedures are provided for every language it supports, so that semantics are common across all realizations of a script. We must also provide for backtracking or iteration through the set of alternatives in a straightforward way.

2.4 Control and Synchronization

Our primary focus is multimodal communication, potentially multiparty as well. We are extending TRINDIKit's triggers to include support for consideration of temporal relationships between events, both within and across modes.

2.5 Integrated Environment

An ideal toolkit would have an integrated set of tools for designing, testing, and debugging dialogues. We would like to support static and dynamic analysis of dialogues, recording and playback of dialogues, graphical dialogue design tools, a "validation suite" of tests to support extension of the toolkit to new programming languages and agent protocols, and above all, the ability to plug-in as-yet-undefined capabilities.

3 Future Work

Significant effort has been devoted to defining our mutable language capability. This capability provides both a reasonable transition path from TRINDIKit scripts and a means for specifying module interfaces and information state structure using a common XML representation.

Our intent is to provide support for several different transport mechanisms to explore the limitations of our approach. To date, we have completed an initial interface definition specification and have developed templates to realize those interfaces with the OAA. DARPA's Galaxy Communicator is the second transport mechanism we will be considering. Time and resources permitting, we will examine some additional transports with differing characteristics, such as CORBA, Java Remote Method Invocation (RMI), or Linda.

We have devoted considerable time to up-front consideration of scripting languages, portable code generation, and module communications, and are now beginning the task of implementing our versions of the TRINDIKit scripting languages. Our target realization for these scripts is a combination of Java code and expert systems that can be executed within a Java program.

We plan to port and formally evaluate our dialogue toolkit within three domains (question-answering, automated tutoring, and multimodal robot control). Our dialogue toolkit will be openly available, as well as sample implementations for each of these domains.

Conclusion

We have described our evolving architecture (based on the TRINDIKit framework) for a flexible dialogue system capable of robust, multimodal, multiparty control of robots.

References

Fong T., C. Thorpe, and C. Baur (2002), Robot as Partner: Vehicle Teleoperation with Collaborative Control, Workshop on Multi-Robot Systems NRL, Washington, D.C.

Kahney, Leander (2001) *Robots Scour WTC Wreckage.* Wired Magazine, http://www.wired.com/news/print/0,1294,46930,00.html

Larsson, Staffan, Robin Cooper, Stina Ericsson (2000) *System Description of GoDis*. Third Workshop in Human-Computer Conversation, Bellagio, Italy.

Lewin, I., Rupp, C., Hieronymus, J., Milward, D, Larsson S., and Berman, A. (2000) 'SIRIDUS Project Deliverable D6.1, System Architecture and Interface Report (Baseline), URL (May 2002): http://www.ling.gu.se/projekt/siridus/Publications/publications.html.

NASA (1997) *Past Missions – Mars Pathfinder*. NASA, http://www.jpl.nasa.gov/missions/past/marspathfinder.html

Rayner, M., B.A. Hockey, and F. James. (2000a) *Turning Speech into Scripts.* AAAI Spring Symposium on Natural Dialogues with Practical Robotic Devices.

Rayner, M., B.A. Hockey, and F. James (2000b) *A compact architecture for dialogue management based on scripts and meta-outputs.* ANLP.

TRINDI (2002) . http://www.ling.gu.se/projekt/trindi.

Do oral messages help visual exploration?

Noëlle CARBONELL
LORIA, CNRS & INRIA
Campus Scientifique, BP 239
F54506 Vandœuvre-lès-Nancy Cedex, France
Noelle.Carbonell@loria.fr

Suzanne KIEFFER
LORIA, CNRS & INRIA
Campus Scientifique, BP 239
F54506 Vandœuvre-lès-Nancy Cedex, France
Suzanne.Kieffer@loria.fr

Abstract

A preliminary experimental study is presented, which aims at eliciting the contribution of oral messages to facilitating visual search tasks in crowded visual displays.
Results of quantitative and qualitative analyses suggest that appropriate verbal messages can improve both target selection time and accuracy. In particular, multimodal messages including a visual presentation of the isolated target together with absolute spatial oral information on its location in the displayed scene are most effective.

1 Context and motivation

1.1 Multimodality: state of the art

Numerous forms of speech-based input multimodality have been proposed, implemented and tested. Combinations of speech with gestural modalities have been studied extensively, especially combinations of speech with modalities exploiting new input media, such as touch screens, pens, data gloves, haptic devices. Both usability and implementation issues have been considered; see, among others, Oviatt *et al.* (1997) [1], Robbe *et al.* (2000) [2], for the first category of issues; Nigay and Coutaz (1993), for the second category.

Contrastingly, speech combined with text and graphics has motivated few studies. As an output modality, speech is mostly used either as a substitute for standard visual presentation modes (*cf.* phone services) or for supplementing deficiencies in visual exchange channels. Recent research efforts have been focusing on two main application domains: providing blind or partially sighted users with easy computer access, and implementing appropriate interaction facilities in situations where only small displays are available (*e.g.*, PDAs and wearable computers), or where the gaze is involved in other activities (*e.g.*, while driving a car); see, for instance, Baber (2001) concerning the first application domain, and De Vries and Johnson (1997) concerning the second one.
However, there is not yet, at least to our knowledge, a substantial amount of scientific work on the integration of speech into the system output modalities, with a view to enriching standard graphical human-computer interaction [3]; that is interaction intended for standard categories of users using standard application software in standard environments and contexts of use.
Published research on output types of multimodality including speech amounts to usability studies of the role of speech in multimedia presentations, such as Faraday and Sutcliffe (1997), and contributions to the automatic generation of multimedia presentations (*cf.* André and Rist, 1993; Maybury, 2001).
In fact, multimedia and multimodality refer to two different concepts, although these terms are often used as synonyms, especially when applied to system outputs. Precise definitions are presented in the next paragraph.

1.2 Definitions: multimodality vs multimedia

Coutaz and Caelen (1991), Maybury (1993; 2001) and Bernsen (1994), among others, define 'media' and 'modalities' contrastingly.
They use the first term for referring to the hardware and software channels through which information is conveyed, and the second one for designating the coupling of a medium with the

[1] on speech and pen.

[2] on speech and finger gestures on a touch screen.

[3] *cf.* direct manipulation, the present prevailing interaction paradigm (Shneiderman, 1983).

interpretation processes required for transforming physical representations of information into meaningful symbols or messages. In other words, and focusing on output media and modalities:

> '… by *media* we mean the carrier of information such as text, graphics, audio, or video. Broadly, we include any necessary physical interactive device (*e.g.*, keyboard, mouse, microphone, speaker, screen). In contrast, by *mode* or *modality* we refer to the human senses (more generally agent senses) employed to process incoming information, *e.g.*, vision, audition, and haptics.'
> (Maybury, 2001)

To characterize the various possible combinations of modalities, taxonomies have been proposed. In Coutaz and Caelen (1991), multimodality is characterized in terms of temporal and semantic coordination strategies for controlling the use of modalities.
In addition, Coutaz *et al.* (1995) defines four properties which prove useful for comparing modalities in terms of their expressive power (*i.e.*, complementarity versus equivalence), and their usage in multimodal contexts (*i.e.*, redundancy versus complementarity).
As for Bernsen's taxonomy (*cf.* Bernsen, 1994), it is a thorough inventory of the output modalities available to user interface designers.

1.3 Motivation and objectives

The frequent confusion between multimedia and multimodality may explain why the generation of system multimodal responses has raised but little interest in the user interface research community, especially from an ergonomic angle, save for studies focused on specific categories of users or specific contexts of use.
However, if standard users are offered speech facilities together with other input modalities, it is mandatory that the system responses are not limited to visual messages. Communication situations where one interlocutor can speak and the other cannot, are rather unusual. Research is then needed on the usability and software issues concerning the generation of appropriate multimodal system responses in standard human-computer interaction environments, and for standard user categories, including the general public.

The main objective of the early experimental study presented here is to contribute to scientific advances in this research area, as it addresses one of the major usability issues relating to the generation of effective oral system messages, namely:

> How to design oral messages which facilitate the visual exploration of crowded displays?
> In particular, how to design messages which effectively help users to locate specific graphical objects in such displays?

Resorting to deictics and visual enhancements of graphical targets is a solution which seems "natural". However, it is no more effective than the sole visual enhancement of the target.
Another approach is to implement a human-like embodiment of the system and to endow it with a pointing device; see, for instance the PPP [4] persona with a pointing stick in André (1997). However, the contribution of personae to the usability and effectiveness of human-computer interaction seems questionable (*cf.* Mulken *et al.*, 1999). Further testing is required in order to determine the usefulness of animated human-like system embodiments in this context.

These reasons explain why we chose to focus first on assessing whether oral messages including spatial information actually facilitate the visual exploration of complex displays, especially the localization of graphical targets in the context of standard computer environments for standard categories of users.

We selected visual search as the experimental task for the following reasons. It is one of the few human visual activities, besides reading, which have motivated a significant amount of psychological research (*cf.* Henderson and Hollingworht, 1998; Findlay and Gilchrist, 1998). And mostly, the design of numerous computer applications may benefit from a better knowledge of this activity, especially applications for the general public such as:

- Online help to current interactive application software. For instance, novice users interacting with present graphical interfaces are often confused by the increasing number of toolbars and icons displayed concurrently.

[4] PPP means 'Personalized Plan Based Presenter'.

- Map reading environments (*cf.* geographical applications), and navigation systems in vehicles.
- Data mining through exploring visualizations of very large data sets (*cf.*, for instance, the hyperbolic graph visualizations proposed by Lamping *et al.*, 1995).

The methodology and experimental set-up are described in the next section, together with the underlying working hypotheses.
Quantitative and qualitative results are presented and discussed in section 3.
Future research directions stemming from these results are included in the conclusion.

2 Methodology and experimental set-up

2.1 Overall experimental protocol

To assess the potential contribution of oral spatial information to facilitating visual search, we designed a first experimental study with:

- target presentation mode as independent variable,
- target search+selection time, and accuracy of target selection, as dependent variables.

Eighteen subjects were to locate and select visual targets in thirty six complex displayed scenes, using the mouse. They were requested to carry out target localization and selection as fast as they could. Colour displays only were used.

Each scene display was preceded by one out of three available presentations of the target:

- Display of the isolated target at the centre of the screen during three seconds (VP);
- Oral characterization of the target (*i.e.*, name of the relevant graphical object), plus spatial information on its position in the scene (OP);
- Simultaneous display of the visual and oral presentations of the target (*i.e.*, multimodal presentation, MP).

These three sets of thirty six dual stimuli (*i.e.*, pairs of stimuli [5]) determined three experimental situations, namely the VP, OP and MP situations. The visual and oral presentations of the targets used in the MP situation were identical to those used in the VP and OP situations.

[5] each pair consisting of a scene and a presentation mode.

Subjects were randomly split up into three groups (of six), so that each subject processed twelve pairs of stimuli in each set, and each pair of stimuli was processed by six subjects.
In order to avoid task learning effects, the processing order was balanced inside each group: three subjects processed pairs in the VP set first, while the other three subjects processed pairs in the OP set first. All subjects processed pairs in the MP set last.

Experimental design choices were motivated by the intent to assess the soundness of the three following working hypotheses, using the VP situation as the reference task context:

A. Multimodal presentations of targets will reduce selection times and error rates in comparison with visual presentations.
B. Oral presentations of targets will also improve selection times and accuracy, compared to visual presentations.
C. The type of spatial information included in the oral presentations of targets will influence selection times and error rates. In particular:
   absolute or relative spatial information will prove more effective than references to *a priori* knowledge (*cf.* subsection 2.3).

In the remainder of the section, further information is given on:
- the criteria used for selecting visual scenes and targets (subsection 2.2);
- the structure and information content of oral messages (subsection 2.3);
- subjects' profiles (subsection 2.4);
- the experimental set-up (subsection 2.5);
- the methodology adopted for analysing and interpreting subjects' results (subsection 2.6).

2.2 Scene selection criteria

Most visual scenes were taken from currently available Web pages in order to provide subjects with realistic task environments.

They were classified according to criteria stemming from Bernsen's taxonomy of output modalities (*cf.* Bernsen, 1994), our aim being to investigate the possible influence of the type of visual information displayed, on target selection times and accuracy.
Actually, our classification was derived from the graphical categories in Bernsen's taxonomy as follows. We focused on static graphical displays

exclusively [6], on the ground that the localization and selection of moving targets in animated visual presentations is a much more complex activity than the selection of still targets in static visual presentations. Issues relating to the exploration of visual animated scenes will be addressed at a later stage in our research.

We established two main classes of static presentations:

- Class 1 comprises representations of structured or unstructured collections of symbolic or arbitrary graphical objects, such as maps, flags, graphs, geometrical forms (*cf.* classes 9, 11, 21, 25 in Bernsen's taxonomy);
- Class 2 includes representations of realistic objects or scenes, namely photographs or naturalistic drawings figuring complex real objects (*e.g.*, monuments) or everyday life environments, such as views of rooms, town or country landscapes, … (*cf.* class 10 in Bernsen's taxonomy).

Half of the thirty six visual scenes belonged to class 1, and the other half to class 2. Class 1 and class 2 scenes in each of the three scene subsets described in subsection 2.1 were randomly ordered. See the overall task set-up in table 1.

Group	VP	OP	MP	Group	OP	VP	MP
G1	P1	P2	P3	G4	P2	P1	P3
G2	P3	P1	P2	G5	P1	P3	P2
G3	P2	P3	P1	G6	P3	P2	P1

Table 1: Overall task set-up.

G_i: group of 3 subjects (3 x 6 = 18 subjects)
P_i: set of 12 visual scenes (3 x 12 = 36 scenes)

Targets were objects or component parts of complex objects (*cf.* the complex real objects in class 2).

They were chosen according to the following criteria. An acceptable target was a unique autonomous [7] graphical object which could be designated unequivocally by a short simple verbal phrase. Although all targets were unique, some of them could be easily confused with other objects in the scene, such confusions being impossible for the other ones.

In order to avoid task learning effects, target size [8] and position were varied from one scene to another.

2.3 Message structures and contents

All messages included a noun phrase meant to designate the target unequivocally. For instance, "the pear" refers to the target unequivocally in the realistic scene reproduced in figure 1.

Figure 1: Basket with fruit (class 2)
"On the left of the apple, the pear."

Three types of spatial information on target locations were experimented:

- Absolute spatial information (ASI), such as "on the left/right", "at the top/bottom";
- Relative spatial information (RSI), for instance "on the left of the apple" (*cf.* figure 1);
- Implicit spatial information (ISI), that is spatial information that can be easily inferred from common *a priori* knowledge and the visual context; for instance, it is easy to locate the Mexican flag among twenty other national flags, from the simple message "The Mexican flag.", if each flag representation is placed on a planisphere inside the matching country (*cf.* figure 2).

Messages included one or two spatial phrases of the same type (ASI, RSI or ISI), or two spatial phrases of different types (namely, ASI + RSI or ASI + ISI), according to the scene complexity.

In order to make the assessment of hypothesis C possible, all messages had the same syntactical structure so that information content was the only variable in messages which might influence localization times and selection errors. The following structure, which emphasizes spatial

[6] *cf.* the five types of static graphical presentations in Bernsen's taxonomy, namely classes 9, 10, 11, 21, 25.

[7] *cf.* constituent parts of complex real objects.

[8] within the limits of the fixed size presentation box.

information, was adopted for most messages, some ISI messages including no spatial phrase:

[Spatial phrase] + Noun phrase (designation)

N.B. Target choices were also influenced by message design requirements: some targets were rejected because they could not be designated verbally in a simple unequivocal way.

Figure 2.: National flags (class 1)
"The Mexican flag."

2.4 Subjects' profiles

As this study involved a restricted number of subjects (18) and was a first attempt at validating hypotheses A, B and C, we defined constraints on subjects' profiles in order to reduce inter-individual diversity so that the selected group would be likely to carry out the prescribed tasks successfully.

To achieve homogeneity, we selected 18 undergraduate or graduate students in computer science with normal eyesight [9], and ages ranging from 22 to 29.
Then, all participants were expert mouse users with alike quick motor reactions, familiar with visual search tasks, and capable of performing the experimental tasks accurately and rapidly.

2.5 Experimental set-up

First, the experimenter presented the overall experimental set-up. Then, after a short training (6 target selections in the VP situation), each subject processed 12 scenes per situation, in the order VP+OP+MP or OP+VP+MP. Before each change in target presentation, the experimenter explained the new specific set-up to the subject.

For each visual scene:

- The target was presented first, during three seconds:
  - either visually in a fixed-size box in the centre of the screen,
  - or orally (together with a blank screen),
  - or orally and visually, simultaneously.
- Then, a button appeared in the centre of the screen together with a written message requesting the subject to click on the button to launch the display of the scene. Therefore, at the beginning of each target selection step, the mouse was positioned in the centre of the screen, making it possible to compare subjects' selection times.
- The next target was presented as soon as the subject had clicked on any object in the displayed scene.

At the end of the session, subjects had to fill in a questionnaire requesting them to rate the difficulty of each task using a six degree scale (ranging from "very easy" to "very difficult"). The session ended up in a debriefing interview.

2.6 Analysis methodology

Quantitative results comprise:
- average {localization + selection} times,
- and error (*i.e.,* wrong target selections) counts or percentages,

computed over all subjects and scenes, as well as per scene category and per type of oral message. We made no attempt at determining the statistical significance of these results, by reason of the small amount of available data.

Qualitative analyses of subjects' results, especially comparisons between target selection errors in the VP, OP and MP situations, provided useful information for defining further research directions.
In order to elicit the possible factors at the origin of selection errors, scenes and targets were characterized using the following features:

- Scene characterisation:
  - complexity (according to the number of displayed objects);
  - and, for class 1 scenes only, visual structure (*e.g.*, random layout of objects; tree, crown or matrix structures; …).
- Target characterization:
  - position on the screen (centre, left, …);
  - visual salience;

[9] save for one subject who was slightly colour-blind.

- familiarity versus unfamiliarity/oddness;
- unicity versus ambiguity;
- in the case of unicity, possible confusions.

Quantitative and qualitative results are presented and discussed in the next section.

3. Results: presentation and discussion

3.1 Experimental data checking

In the first place, we checked whether experimental design choices might have biased subjects' error rates and selection times.

No effect of task order (VP–OP versus OP–VP) could be detected on the basis of comparisons between average selection times, respectively error rates, in the following group pairs: G1 versus G4, G2 versus G5, and G3 versus G6 (*cf.* the groups in table 1).

In addition, subjects achieved similar results in each of the three situations (VP, OP and MP), whatever subset of scenes (P1, P2 or P3) they processed. Comparisons between the G1+G4, G2+G5 and G3+G6 groups showed no evidence of scene group influence on subjects' results.

Finally, two scenes (both in class 1) had to be excluded from the analysis of subjects' results, due to technical incidents.

3.2 Quantitative results

3.2.1 Global analysis

As regards selection accuracy, oral messages proved much more effective than visual target presentations, as shown by comparisons between the VP and OP situations. However, selection was slower in the absence of prior visualizations of isolated targets (*cf.* table 2). The total number of errors in the OP situation decreased by 55%, while average selection time increased by 28%.
Slower average selection time in the OP situation, together with a much higher standard deviation, may be explained by the fact that subjects were unfamiliar with the visual search tasks in the OP situation; this situation being rather unusual compared to the VP and MP situations which occur frequently in everyday life. Therefore, the higher variability of selection times in the OP situation may be assumed to reflect the high inter-individual diversity of cognitive abilities and processes.

In keeping with these results, table 2 also shows that multimodal presentations of targets reduced both selection times and error rates, in comparison with visual presentations.

Target presentation mode	Number of errors	Average selection time (sec.)	Standard deviation (sec.)
VP	31	2.83	1.70
OP	14	3.92	3.50
MP	8	2.70	1.93

Table 2: Results per target presentation mode.

This finding is compatible with perception models which postulate the existence of higher level multimodal processes resulting from interferences or collaborations between lower level visual and auditory unimodal processes (*cf.*, for instance, Engelkamp, 1992).

To conclude, as no task learning effect was observed (*cf.* subsection 3.1), these results contribute to validating hypothesis A, while they confirm hypothesis B partly. However, if our interpretation of the longer selection times in the OP situation is correct, hypothesis B may be held to be true for users familiar with visual target selection tasks in OP contexts.

These quantitative global results also suggest useful recommendations for improving user interface design.
In order to facilitate and improve the effectiveness of visual search tasks in crowded displays, two forms of user support may prove useful:

a. if *accuracy* only is sought for, an oral message comprising an unambiguous verbal designation of the graphical target and spatial information on its location in the display will prove sufficient.
b. if both *accuracy* and *rapidity* are sought for, a multimodal message will be more appropriate, that is a message comprising a context-free visual presentation of the target together with an oral message including the same information as the message in a.

However, further experimental research is needed to confirm these recommendations beyond doubt, in-as-much as they have been inferred from a relatively small sample of experimental data and measurements.

In addition, oral and multimodal messages should be compared, in terms of effectiveness and comfort, with other forms of user support, such as target visual enhancement through colour, animation, zooming, etc. Until a sufficient amount of experimental data has been collected, recommendations a. and b. should be considered as tentative.

Analyses of results per class of scenes and type of message (i.e., type of spatial oral information) are presented next. These analyses make it possible to refine the above recommendations.

3.2.2 Detailed analysis

Results per class of scenes

Subjects' results, grouped per scene class and target presentation mode, are presented in table 3. Error percentages have been computed over 96 samples for class 1 (*cf.*, in subsection 3.1, the exclusion of two class 1 scenes), and 108 samples for class 2.

Multimodal messages proved most effective, especially for scenes representing symbolic or arbitrary objects, in comparison with the VP and OP situations. For scenes in class 1, errors were reduced by 86% and 73%, respectively, while average selection times were decreased by 7% and 30%.

As for realistic scenes, the average selection time in the MP situation is similar to the VP one and markedly inferior (by 33%) to the OP one, while the number of errors is similar to the OP one, and inferior (by 35%) to the VP one.

These results suggest that subjects in the MP situation took advantage of all the information available in the VP and OP situations. They used successfully visual information to solve ambiguous verbal designations of targets, and oral information or both types of information, to solve visual ambiguities between the target and other graphical objects in the scene. In addition, verbal spatial information helped them to locate targets more rapidly, or so it seems.

Finally, the fact that average selection times were consistently longer for class 1 scenes than for class 2 scenes can be explained as follows.

If the target is a familiar object (such as a pan) in a familiar realistic scene (a kitchen, for instance), visual exploration of the scene is facilitated by *a priori* knowledge of the standard structure of the scene and the likely locations of the target therein. Such knowledge is not available in the case of unrealistic scenes such as class 1 scenes; the structure of the scene and the possible locations of the target cannot be foreseen using *a priori* knowledge, so that a more careful search, or even an exhaustive exploration, of the scene is necessary for locating the target object.

Target presentation mode	Percentage of errors	Average selection time (sec.)	Standard deviation (sec.)
VP-C1	14.6	3.27	1.94
VP-C2	15.7	2.43	1.39
OP-C1	7.3	4.30	4.09
OP-C2	6.5	3.58	2.87
MP-C1	2	3.03	2.36
MP-C2	5.5	2.40	1.36

Table 3: Results per target presentation mode and class of scene

This hypothesis may also explains why multimodal target presentations proved most effective for scenes belonging to class 1: both oral and visual information contributed to compensate for the lack of *a priori* knowledge.

Results per message type

Five categories of verbal messages were experimented (*cf.* subsection 2.3). Messages were classified according to the type of spatial information they comprised: absolute (ASI), relative (RSI), implicit (ISI), plus absolute-relative (ASI+RSI) and absolute-implicit (ASI+ISI). Subjects' results, grouped according to these messages categories, are presented in table 4. As the number of scenes varied from one message class to the other, error percentages have been computed over 48 samples (ASI), 72 samples (RSI), 24 samples (ISI), 36 samples (ASI+RSI) and 12 samples (ASI+ISI) [10].

For each category of messages, comparisons between results achieved by subjects in the VP, OP and MP situations suggest that absolute and/or relative spatial information improved selection accuracy markedly (*cf.* the ASI, RSI and ASI+RSI types of messages).

[10] that is 192 samples instead of 6 x 36 = 204 samples (*cf.* the two scenes which were excluded).

However, the usefulness of ISI messages seems questionable, at least in the OP situation. Their effectiveness in the MP situation denotes the complexity of the interpretation processes at work in the interpretation of multimodal stimuli.

Average RSI and ISI selection times were much longer in the OP situation (4.03, 6.12) than in the other situations (*i.e.*, 2.86 and 1.84 for VP, 3.12 and 2.1 for MP).

For RSI messages, this effect may be due to the complexity of the visual search strategy induced by relative spatial information. This strategy probably includes two steps: first, localization of the reference graphical object, then exploration of its vicinity in search of the target.

As for ISI messages, their interpretation involves cognitive processes which may slow down selection.

	VP	Situation	
Scenes grouped per message type	**Percentage of errors**	**Average selection time (sec.)**	**Standard deviation (sec.)**
ASI	10.4	2.87	1.19
RSI	26.4	2.86	2.14
ISI	4.17	1.84	0.57
ASI+RSI	13.89	3.57	1.99
ASI+ISI	8.33	3.54	0.98
	0P	**Situation**	
ASI	0	2.91	3.41
RSI	8.33	4.03	5.94
ISI	16.67	6.12	3.78
ASI+RSI	5.56	3.82	3.78
ASI+ISI	16.67	5.19	3.37
	MP	**Situation**	
ASI	4.17	2.42	1.41
RSI	8.33	3.12	2.43
ISI	0	2.1	1.06
ASI+RSI	0	2.82	1.84
ASI+ISI	0	2.98	2.53

Table 4: Results per target presentation mode and type of verbal message

This interpretation may also explain why RSI messages did not affect selection times in the MP situation noticeably. The target being in the vicinity of the reference object and having been viewed previously, it can be recognized through peripheral vision, so that one eye fixation only is required for locating both the reference object and the target (*cf.* van Diepen *et al.*, 1998).

However, it is also possible that, in the MP situation, subjects tended to adopt a simpler search strategy based exclusively on the available visual information, hence comprising a single visual search step, whenever the oral message induced a complex slow selection strategy. This second interpretation has the advantage to explain why both RSI and ISI messages exerted no perceptible influence, in the MP situation, on selection times.

To sum up, while the inclusion of any category of verbal spatial information in multimodal target presentations seems worthwhile, absolute spatial information should be preferred over other information types in the design of oral target presentations, in order to improve both selection times and accuracy. This conclusion partly confirms hypothesis C.

3.3 Qualitative analyses

Qualitative analyses were focused on the subjects' errors exclusively, with a view to:

- getting a better understanding of the contribution of verbal messages to assisting users in visual search tasks,
- and obtaining useful knowledge for improving message design.

These analyses involve the detailed characterisations of scenes and messages listed in subsection 2.6, as well as the subjects' subjective ratings of the difficulty of the 36 experimental visual tasks (*cf.* the post-session questionnaires mentioned in subsection 2.5).

Scenes were filtered so that, in each situation, only the scenes which had occasioned more than one error were considered, on the basis of the following assumption:

> for a given scene in a given situation, the reasons for the failure of one single subject are more likely to be related to the subject' capabilities than to the scene characteristics or the message information content.

3.3.1 Visual situation

The main plausible factors at the origin of the selection errors observed in the visual situation are presented next. Percentages represent:

– the number of errors which a given characteristic of the scene may explain, by itself or in conjunction with other factors;
– computed over the total number of filtered errors (*i.e.*, 26).

Factors are listed in decreasing order of the percentages of errors they contribute to explain:

- Concerning targets:
  lack of salience (85%), eccentric position in the scene (69%), possible confusions with other objects (69%), unfamiliarity (50%).
- Concerning scenes:
  crowded (69%), unstructured (46%), representing geometric forms (42%).

This analysis of subjects' errors in the VP situation will be used as a reference in the next subsection which is focused on errors in the OP and MP situations.

3.3.2 Oral and multimodal situations

Five scenes in the OP situation and only two in the MP situation occasioned more than one error, against eight in the VP situation.
In addition, 23 errors in the VP situation were "corrected" in the OP situation, so that six out of the eight scenes occasioning more than one error in the VP situation yielded error-free results in the OP situation.
These comparisons bring out the usefulness of oral messages for improving target selection accuracy.

However, four scenes yielding error-free results in the VP and MP situations occasioned ten out of the twelve filtered errors observed in the OP situation [11]. Therefore, it is likely that the main factor at the origin of these errors is the poor quality of the information content of the corresponding oral messages.
The analysis of the four corresponding messages, together with the information provided by questionnaires and debriefings, support this conclusion. Four errors were motivated by an ISI message which referred to knowledge unfamiliar to the majority of subjects. A too complex ASI+ISI message (structure and length) referring to knowledge some subjects were unfamiliar with may account for two other errors. As for the two other pairs of errors, they may be reliably ascribed to the use, in the two verbal target designations, of technical substantives the exact meanings of which were unfamiliar to some subjects.

The fact that none of these errors occurred in the MP situation, together with the fact that two scenes only occasioned the four filtered errors observed in this situation, illustrates the advantages of combining visual and verbal information in target presentations.
Two errors occurred, in this situation, on a "difficult" scene which occasioned six errors in the VP situation (crowded scene, and non salient, unfamiliar target, easy to confuse with other objects), and two errors in the OP situation (use of technical vocabulary). The other two errors were occasioned by a scene which was processed successfully by all subjects in the OP situation, but occasioned three errors in the VP situation. This may hint that the processing of multimodal incoming information is guided or controlled by visual perception strategies rather than by cognitive processes.

The qualitative analysis of errors confirms the usefulness of oral messages for improving the accuracy of visual target identification, provided that:
– messages are short, their syntactical structure straightforward, the vocabulary used familiar to users,
– and mostly, provided that their information content is appropriate.

4. Conclusion

A preliminary experimental study has been presented, which aims at eliciting the contribution of oral messages to facilitating visual search tasks in crowded visual displays.
Results of quantitative and qualitative analyses suggest that appropriate verbal messages can improve both target selection time and accuracy. In particular, multimodal messages including a visual presentation of the isolated target, and absolute spatial oral information on its location in the visual scene, are most effective.
However, these results are tentative by reason of the small number of subjects involved in the experiment (18), the limited number of scenes they had to process, and the coarseness of measurements based on target selection using the mouse. Nevertheless, these results are

[11] Error patterns for these images were as follows: 4, 2, 2, 2.

encouraging, especially the qualitative ones, in-as-much as they suggest that the research directions we have initiated are worth pursuing.

In the short term, we intend to conduct an experimental study along the same lines as the study presented here. The protocol and set-up will be roughly similar, save for the following features:

- it will involve a greater number of subjects so that the statistical significance of quantitative results can be assessed;
- an eye-tracker will be used for measuring target localization time and accuracy more precisely; the analysis of ocular movements will also enable us to gain an insight into subjects' visual search strategies according to the scene characteristics; the present classification of scenes will be refined accordingly.

In the medium term, we shall attempt to compare oral assistance to visual search with target visual enhancement techniques, in terms of effectiveness and user comfort.

References

André, E. (1997), WIP and PPP: A Comparison of two Multimedia Presentation Systems in Terms of the Standard Reference Model. *Computer Standards and Interfaces*, Vol. 18, 6-7, 555-564.

André, E., Rist, T. (1993). The Design of Illustrated Documents as a Planning Task. In M.T. Maybury (Ed.), *Intelligent Multimedia Interfaces*, Memlo Park (CA): AAAI/MIT Press, pp. 94-116.

Baber, C. (2001). Computing in a multimodal world. In *Proceedings UAHCI'01 (HCI International 2001)*, Mahwah (NJ): Lawrence Erlbaum, pp. 232-236.

Bernsen, N.-O. (1994). Foundations of multimodal representations, a taxonomy of representational modalities. *Interacting with computers*, 6, 347-371.

Coutaz, J., Caelen, J. (1991). A taxonomy for multimedia and multimodal user interfaces. In *Proceedings of the 1st ERCIM Workshop on Multimodal HCI*, Lisbon: INESC, pp. 143-148.

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., Young, R. (1995). Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. In *Proceedings INTERACT'95* (Lillehammer, Norway), pp. 115-120.

De Vries, G., Johnson, G.I. (1997). Spoken help for a car stereo: an exploratory study. *Behaviour and Information Technology*, 16(2), 79-87.

Engelkamp, J. (1992). Modality and modularity of the mind. In *Actes du 5ème Colloque de l'ARC 'Percevoir, raisonner, agir – Articulation de modèles cognitifs'*, 24-26 mars 1992, Nancy, pp. 321-343.

Faraday, P., Sutcliffe, A. (1997). Designing Effective Multimedia Presentations. In *Proceedings CHI'97*, NewYork: ACM Press & Addison Wesley, pp. 272-278.

Findlay, J.M., Gilchrist, D. (1998). Eye Guidance and Visual Search. In (Underwood, 1998), chapter 13, pp. 295-312.

Henderson, J.M., Hollingworth, A. (1998). Eye Movements during Scene Viewing: an Overview. In (Underwood, 1998), chapter 12, pp. 269-293.

Lamping, J., Rao, R., Pirolli, P. (1995). A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In *Proceedings CHI'95*, New York: ACM Press & Addison Wesley, pp. 401-408.

Maybury M.T. (Ed.) (1993). *Intelligent Multimedia Interfaces*. Memlo Park, (CA): AAAI/MIT Press.

Maybury, M.T. (2001). Universal multimedia information access. In *Proceedings UAHCI'01 (HCI International 2001)*, Mahwah (NJ): Lawrence Erlbaum, pp. 382-386.

Mulken, S., André, E., Müller, J. (1999). An Empirical Study on the Trustworthiness of Life-Like Interface Agents. In *Proceedings HCI International 1999*, Mahwah (NJ): Lawrence Erlbaum, pp. 152-156.

Nigay, L., Coutaz, J. (1993). A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In *Proceedings INTERCHI'93*, New York: ACM Press & Addison Wesley, pp. 172-178.

Oviatt, S., DeAngeli, A., Kuhn, K. (1997). Integration and Synchronisation of Input Modes during Multimodal Human-Computer Interaction. In *Proceedings CHI'97*, New York: ACM Press & Addison Wesley, pp. 415-422.

Robbe, S., Carbonell, N., Dauchy, P. (2000). Expression constraints in multimodal human-computer interaction. In *Proceedings IUI'2000*, New York: ACM Press, pp. 225-229.

Shneiderman, B. (1983). Direct manipulation: a step beyond programming languages. *IEEE Computer*, 16, 57-69.

Underwood, G. , Ed. (1998). *Eye Guidance in Reading and Scene Perception*. Amsterdam: Elsevier.

Van Diepen, M.J., Wampers, M., d'Ydewall, G. (1998). Functional Division of the Visual Field: Moving Masks and Moving Windows. In (Underwood, 1998), chapter 15, pp. 337-355.

MIND: A Semantics-based Multimodal Interpretation Framework for Conversational Systems

Joyce Chai, Shimei Pan and Michelle X. Zhou

IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532, USA
{jchai, shimei, mzhou}@us.ibm.com

Abstract

To facilitate a full-fledged multimodal human-machine conversation, we are developing an intelligent infrastructure called Responsive Information Architect (RIA). As a part of this effort, we are building a semantics-based multimodal interpretation framework called MIND (**M**ultimodal **I**nterpretation for **N**atural **D**ialog). MIND addresses both multimodal input understanding and discourse interpretation in a conversation setting. In particular, MIND has two unique features. First, MIND characterizes intention and attention of user inputs and the entire conversation from multiple dimensions. This fine grained semantic model provides a computational basis for multimodal interpretation. Second, MIND uses rich contexts such as conversation discourse, domain knowledge, visual context, user and environment models to facilitate multimodal understanding. This approach allows MIND to improve understanding of user inputs, including those abbreviated or imprecise ones.

Keywords: multimodal user interfaces, multimodal input interpretation, multimodal conversation

1. Introduction

To aid users in their information-seeking process, we are building an intelligent infrastructure, called Responsive Information Architect (RIA), which can engage users in a full-fledged multimodal conversation. Currently, RIA is embodied in a testbed, called Real Hunter™, a real-estate application for helping users find residential properties.

Figure 1 shows RIA's main components. A user can interact with RIA using multiple input channels, such as speech and gesture. To understand a user input, the multimodal interpreter exploits various contexts (e.g., conversation history) to produce an interpretation frame that captures the meanings of the input. Based on the interpretation frame, the conversation facilitator decides how RIA should act by generating a set of conversation acts (e.g., Describe information to the user). Upon receiving the conversation acts, the presentation broker sketches a presentation draft that expresses the outline of a multimedia presentation. Based on this draft, the language and visual designers work together to author a multimedia blueprint that contains fully coordinated and detailed multimedia presentation. The blueprint is then sent to the producer to be realized. To support all components described above, an information server supplies various contextual information, including domain data (e.g., houses and cities for a real-estate application), a conversation history (e.g., detailed conversation exchanges between RIA and a user), a user model (e.g., user profiles), and an environment model (e.g., device capabilities).

Our focus in this paper is on the interpretation of multimodal user inputs. Specifically, we are developing a semantics-based multimodal interpretation framework called MIND (**M**ultimodal **I**nterpretation for **N**atural **D**ialog). MIND is inspired by the earlier works on input interpretation from both multimodal interaction systems (e.g.,[Bolt 1980, Burger and Marshall 1993, Cohen et al 1996, Zancanaro et al 1997, Wahlster 1998, Johnston and Bangalore 2000]) and spoken dialog systems [Allen et al 2001, Wahlster 2000]. In particular, MIND provides a unique framework that addresses both multimodal input interpretation (capture the meanings of a multimodal input at a particular conversation turn) and discourse interpretation (identify the impact of indi-

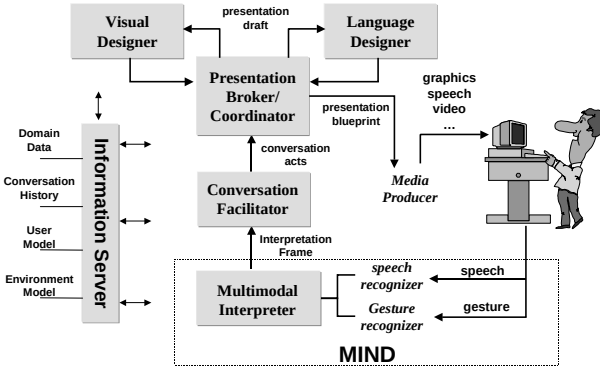


Figure 1. RIA Infrastructure

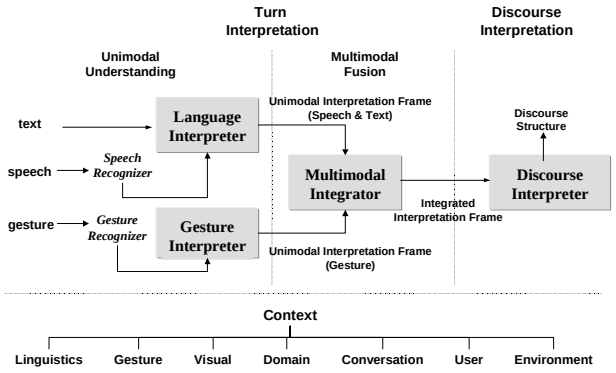


Figure 2. MIND components

vidual inputs on the overall progress of a conversation). Specifically, MIND presents two features. The first is a fine-grained semantic model that characterizes the meanings of user inputs and the overall conversation. The second is an integrated interpretation approach that identifies the semantics of user inputs and the overall conversation using a wide variety of contexts (e.g., conversation history and domain knowledge). These two features enable MIND to improve understanding of user inputs, including those ambiguous and incomplete inputs.

Before presenting MIND in details, we begin with an overview and an example scenario to help explain our work.

2. MIND Overview

MIND supports two major processes (Figure 2): turn interpretation and discourse interpretation. *Turn interpretation* identifies the semantics of user multimodal inputs at a particular turn of a conversation. Turn interpretation is a two-step process. First, an array of recognizers (e.g., a speech recognizer) convert input signals (e.g., speech signals) to modality-specific outputs (e.g., text). These outputs are then processed by modality-specific interpreters (e.g., a natural language interpreter). As a result, the meanings of each unimodal input are captured by a unimodal interpretation frame. Based on these meanings, a multimodal integrator uses proper contextual information to infer and create an integrated interpretation frame. This frame captures the overall meanings of the multimodal inputs at this conversation turn.

In addition to understanding user inputs at each conversation turn, MIND also captures the overall progress of a conversation. In particular, MIND utilizes a discourse interpreter to capture how a particular user input is related to the conversation as a whole. Based on a *discourse structure* that models the conversation history, MIND decides whether the current input contributes to an existing conversation topic or it initiates a new one.

3. Example Scenario

Table 1 logs a conversation fragment between a user Joe and RIA. Joe initiates the conversation by asking for houses in Irvington (U1), and RIA replies by showing a group of desired houses (R1). Based on the generated visual display, Joe points to the screen (a position between two houses) and asks for the price (U2). In this case, it is not clear which object Joe is pointing at. There are three candidates: two houses nearby and the town of Irvington†. Using our domain knowledge, MIND can rule out the town of Irvington, since Joe is asking for a price. At this point, MIND still can not determine which of the two house candidates is the desired one. To clarify this ambiguity, RIA highlights both houses and asks Joe to pinpoint the house of interest (R2).

Again, Joe's reply (U3) alone would be ambiguous, since there are multiple red objects on the screen. However, using the conversation history (R2) and the visual properties (Figure 3), MIND is able to infer that Joe is referring to the highlighted red house. Joe continues on to ask for the size (U4). This request by itself is incomplete, since Joe did not explicitly specify the object of interest (house). Nevertheless, MIND understands that Joe is asking for the size of the same red house based on the conversation history (U2–3).

Joe moves on to inquire about another house (U5). This input by itself does not indicate exactly what Joe wants. Again, using the conversation history (U4), MIND recognizes that Joe is most likely asking for *the size* of another house. Next Joe switches to asking

Joe:	*Speech:* Show me houses in Irvington. (**U1**)
RIA:	*Speech:* Here are the houses you requested. *Graphics:* Show a collection of houses on the map (**R1**)
Joe:	*Speech:* What's the cost? *Gesture:* Point to the screen (**U2**)
RIA:	*Speech:* Which house are you interested in? *Graphics*: Highlight two candidate houses (**R2**)
Joe:	*Speech:* The red one (**U3**)
RIA:	*Speech:* The asking price of this red house is 350,000 dollars. *Graphics:* Highlight the red house and show the price (**R3**)
Joe:	*Speech:* And the size? (**U4**)
RIA:	*Speech:* The size of this house is 2000 square feet (**R4**)
Joe:	*Speech:* This one? *Gesture:* Put a question mark on top of a house icon (**U5**)
RIA:	*Speech:* The size of this house is 2200 square feet. (**R5**) *Graphics:* Highlight the house icon
Joe:	*Speech:* By the way, where is the train station? (**U6**)
RIA:	*Speech:* Here is the train station in Irvington. *Graphics:* Indicate the train station on the map (**R6**)
Joe:	*Speech:* OK.Then...how many bedrooms does it have? (**U7**)
RIA:	*Speech:* This house has four bedrooms.(**R7**)

Table 1. A conversation fragment.

† The generated display has multiple layers, where the house icons are on top of the Irvington town map. Thus this deictic gesture could either refer to the town of Irvington or houses.

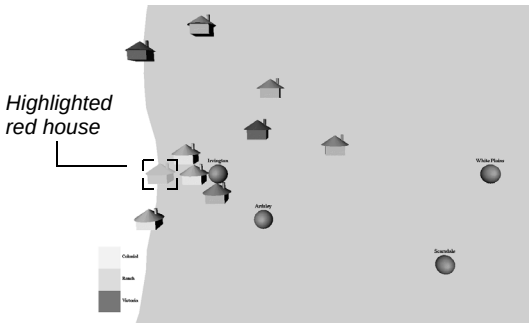


Figure 3. A snapshot of graphics output

for the location of a train station (U6). According to our domain knowledge, train stations are always related to towns. Although Joe did not specify the town at this turn, MIND is able to conclude that the relevant town is Irvington using the conversation history (U1). Finally Joe asks about the number of bedrooms (U7). Based on the current visual context (one house still being highlighted from U5), MIND infers that Joe now returns to the previously explored house.

4. Semantics-based User Input Modeling

To enable a thorough understanding of user multimodal inputs, we use a set of semantic features to model user-computer conversations. As described below, our model not only captures fine-grained semantic aspects of user inputs at each turn of a conversation, but also characterize the overall progress of the conversation.

4.1 Turn-level Modeling

In support of a full-fledged multimodal conversation, MIND has two goals. First, MIND must understand the meanings of user inputs precisely so that the conversation facilitator (Figure 1) can decide how the system should act. Second, MIND attempts to capture the user input styles (e.g., using a particular verbal expression or gesturing in a particular way) or user communicative preferences (e.g., preferring a verbal vs. a visual presentation). The captured information helps the multimedia generation components (visual or language designers in Figure 1) create more effective and tailored system responses. To accomplish both goals, MIND characterizes five aspects of a user input: intention, attention, interpretation status, presentation preference, and modality decomposition.

Intention. *Intention* describes the purpose of a user input [Grosz and Sidner 1986]. We characterize three aspects of intention: Motivator, Act, and Type. Motivator captures the purpose of an interaction. Since we focus on information-seeking applications, MIND currently distinguishes three top-level purposes: DataPresentation, DataAnalysis (e.g., comparison), and ExceptionHandling (e.g., disambiguation).

Act indicates one of the three user actions: request, reply, and inform. Request specifies that a user is making an information request (e.g., asking for a collection of houses in U1 Table 1). Reply indicates that the user is responding to a previous RIA request (e.g., confirming the house of interest in U3). Unlike Request or Reply, Inform states that a user is simply providing RIA with specific information, such as personal profiles or interests. For example, during a house exploration, a user may tell RIA that she has school-age children.

Type further refines an user action. For example, MIND may distinguish two different types of Request. One user may request RIA to Describe the desired information, such as the price of a house, while the other may request RIA simply to Identify the desired information (e.g., show a train station on the screen).

In some cases, Motivator, Act and Type can be directly captured from individual inputs (e.g. U1). However, in other situations, they can only be inferred from the conversation discourse. For example, from U3 itself, MIND only understands that the user is referring to a red house (i.e., Type: Refer). It is not clear whether this is a reply or inform. Moreover, the overall purpose of this input is also unknown. Nevertheless, based on the conversation context, MIND understands that this input is a reply to a previous RIA question (Act: Reply), and contributes to the overall purpose of ExceptionHandling (Motivator: ExceptionHandling).

In addition to the purpose of each user input, Motivator also captures discourse purposes (described later). Therefore, Motivator can be also viewed as characterizing subdialogues discussed in previous literatures [Lambert and Carberry 1992, Litman and Allen 1987]. For example, ExceptionHandling (with Type: correct) corresponds to Correction subdialogue. The difference from earlier works is that our Motivator is used to model intentions at both input and discourse levels. Finally, we model intention not only to support conversation, but also to facilitate multimedia generation. Specifically, Motivator and Type together direct RIA in its response generation. For example, RIA would consider Describe and Identify two different data presentation directives [Zhou and Pan 2001].

Figure 4(a) shows the Intention that MIND has identified from the user input U2 (Table 1). It says that the user is asking RIA to present him with some information. The information to be presented is captured in Attention (described next).

Attention. While Intention indicates the purpose of a user input, Attention captures the content of a user input with six parameters. Base specifies the semantic category of the content (e.g., all houses in our appli-

(a) Intention **Act**: Request **Motivator**: DataPresentation **Type**: Describe	**(d) Modality Decomposition** **Modality**: ^SpeechInput **Modality**: ^GestureInput
(b) Attention **Base**: House **Topic**: Instance **Focus**: SpecificAspect (^Topic) **Aspect**: Price **Constraint**: < > **Content**: [MLS0187652 | MLS0889234]	**(e) Presentation Preference** **Directive**: <Summary> **Media**: <Multimedia> **Device**: <Desktop> **Style**: < >
(c) Interpretation Status **SyntacticComplete**: Attentional-ContentAmbiguity **SemanticComplete**: TRUE	

Figure 4. The interpretation of a multimodal input U2[1].

1. Symbol ^ indicates a pointer and < > labels a default value. A defau value indicates that a pre-defined vlaue is given to a parameter since information concerning this parameter has been identified from the u input. A default value can be overwritten when information is identifi from other sources (e.g., context).

cation belong to the House category). Topic indicates whether the user is concerned with a concept, a relation, an instance, or a collection of instances. For example, in U1 (Table 1) the user is interested in a collection of House, while in U2 he is interested in a specific instance.

Focus further narrows down the scope of the content to distinguish whether the user is interested in a topic as a whole or just specific aspects of the topic. For example, in U2 the user focuses only on *one* specific aspect (price) of a house instance. Aspect enumerates the actual topical features that the user is interested in (e.g., the price in U2). Constraint holds the user constraints or preferences placed on the topic. For example, in U1 the user is only interested in the houses (Topic) located in Irvington (Constraint). The last parameter Content points to the actual data in our database.

Figure 4(b) records the Attention identified by MIND for the user input U2. It states that the user is interested in the price of a house instance, MLS0187652 or MLS0889234 (house ids from the Multiple Listing Service). As discussed later, our fine-grained modeling of Attention provides MIND the ability to discern subtle changes in user interaction (e.g., a user may focus on one topic but explore different aspects of the topic). This in turn helps MIND assess the overall progress of a conversation.

Interpretation Status. InterpretationStatus provides an overall assessment on how well MIND understands an input. This information is particularly helpful in guiding RIA's next move. Currently, it includes two features. SyntacticCompleteness assesses whether there is any unknown or ambiguous information in the interpretation result. SemanticCompleteness indicates whether the interpretation result makes sense. Using the status, MIND can inform other RIA components whether a certain exception has risen. For example, the value AttentionalContentAmbiguity in SyntacticCompleteness (Figure 4c) indicates that there is an ambiguity concerning Content in Attention, since MIND cannot determine whether the user is interested in MLS0187652 or MLS0889234. Based on this status, RIA would ask a clarification question to disambiguate the two houses (e.g., R2 in Table 1).

Presentation Preference. During a human-computer interaction, a user may indicate what type of responses she prefers. Currently, MIND captures user preferences along four dimensions. Directive specifies the high-level presentation goal (e.g., preferring a summary to details). Media indicates the preferred presentation medium (e.g., verbal vs. visual). Style describes what general formats should be used (e.g., using a chart vs. a diagram to illustrate information). Device states what devices would be used in the presentation (e.g., phone or PDA). Using the captured presentation preferences, RIA can generate multimedia presentations that are tailored to individual users and their goals. For example, Figure 4(e) records the user preferences from U2. Since the user did not explicitly specify any preferences, MIND uses the default values to represent those preferences. Presentation preferences can either directly derived from user inputs or inferred based on user and environment contexts.

Modality Decomposition. ModalityDecomposition (Figure 4d) maintains a reference to the interpretation result for each unimodal input, such as the gesture input in Figure 5(a–d) and the speech input in Figure 5(e–f). In addition to the meanings of each unimodal input (Intention and Attention), MIND also captures modality-specific characteristics from the

Gesture Input **(a)**	**(b)**	**(c)**	**(d)**	Speech Input **(e)**	**(f)**
Intention	**Attention (A1)**	**(A2)**	**(A3)**	**Intention**	**Attention**
Act: < > **Motivator**: < > **Type**: Refer **SurfaceAct**: Point	**Base**: House **Topic**: Instance **Focus**: < > **Aspect**: < > **Constraint**: < > **Content**: [MLS0187652]	**Base**: House **Topic**: Instance **Focus**:< > **Aspect**: < > **Constraint**: < > **Content**: [MLS0889234]	**Base**: City **Topic**: Instance **Focus**: < > **Aspect**: < > **Constraint**: < > **Content**: [Irvington]	**Act**: Request **Motivator**: DataPresentation **Type**: Describe **SurfaceAct**: Inquire	**Base**: < > **Topic**: Instance **Focus**: SpecificAspect(^Topic) **Aspect**: Price { <**SynCat**: Noun> <**Realization**: "cost">} **Constraint**: [ReferredBy THIS] **Content**: < >

Figure 5. Separate interpretation of two unimodal inputs in U2.

inputs. In particular, MIND uses SurfaceAct to distinguish different types of gesture/speech acts. For example, there is an Inquire speech act (Figure 5e) and a Point gesture act (Figure 5a). Furthermore, MIND captures the syntactic form of a speech input, including the syntactic category (SynCat) and the actual language realization (Realization) of important concepts (e.g., Topic and Aspect). For example, Aspect price is realized using a noun cost (Figure 5f). Using such information, RIA can learn to adapt itself to user input styles (e.g., using similar vocabulary).

4.2 Discourse-level Modeling.

In addition to modeling the meanings of user inputs at each conversation turn, we also model the entire progress of a conversation. Based on Grosz and Synder's conversation theory [Grosz and Sidner 1986], we establish a refined *discourse structure* that characterizes the conversation history for supporting a full-fledge multimodal conversation. This is different from other multimodal systems that maintain the conversation history by using a global focus space [Neal et al 1998], segmenting focus space based on intention [Burger and Marshall 1993], or establishing a single dialogue stack to keep track of open discourse segments [Stent et al 1999].

Conversation Unit and Segment. Our discourse structure has two main elements: conversation units and conversation segments. A *conversation unit* records user or RIA actions at a single turn of a conversation. These units can be grouped together to form a *segment* (e.g., based on their intentional similarities). Moreover, different segments can be organized into a hierarchy (e.g., based on intentions and sub-intentions). Figure 6 depicts the discourse structure that outlines the first eight turns of the conversation in Table 1. This structure contains eight units (rectangles U1–4 for the user, R1–4 for RIA) and three segments (ovals DS1–3).

Specifically, a user conversation unit contains the

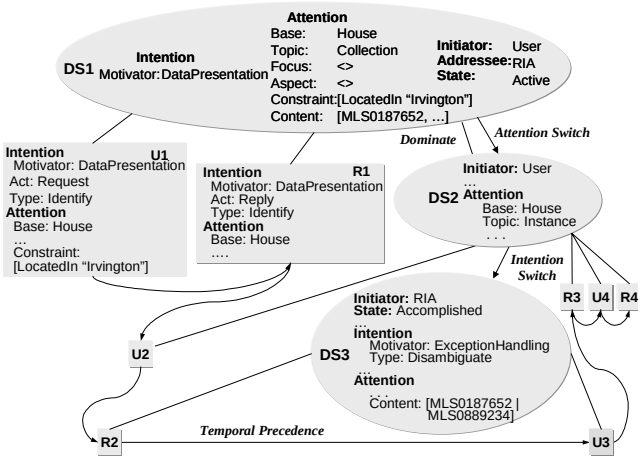


Figure 6. Fragment of a discourse structure

interpretation result of a user input discussed in the last section. A RIA unit contains the automatically generated multimedia response, including the semantic and syntactic structures of a multimedia presentation [Zhou and Pan 2001]. A segment has five features: Intention, Attention, Initiator, Addressee, and State. The Intention and Attention are similar to those modeled in the turns (see DS1, U1 and R1 in Figure 6). Our uniform modeling of intention and attention for both units and segments allows MIND to derive the content of a segment from multiple units (see Section 5.2) during discourse interpretation. In addition, Initiator indicates the conversation initiating participant (e.g., Initiator is User in DS1). Addressee indicates the recipient of the conversation (e.g., Addressee is RIA in DS1). Currently, we are focused on one-to-one conversation. However, MIND can be extended to multi-party conversations where the Addressee could be a group of agents. Finally, State reflects the current state of a segment: active, accomplished or suspended. For example, after U3 DS1 is still active, but DS3 is already accomplished since its purpose of disambiguating the content has been fulfilled.

Discourse Relations. To model the progress in a conversation, MIND captures three types of relations in the discourse: conversation structural relations, conversation transitional relations and data transitional relations.

Conversation structural relations reveal the intentional structure between the purposes of conversation segments. Following Grosz and Sidner's early work, there are currently two types: dominance and satisfaction-precedence. For example, in Figure 6, DS1 *dominates* DS2, since exploring all available houses in Irvington (DS1) comprises the exploration of a specific house in this collection (DS2).

Conversation transitional relations specify transitions between conversation segments and between conversation units as the conversation unfolds. Currently, two types of relations are identified between segments: intention switch and attention switch. The intention switch relates a segment which has a different intention from the current segment. Interruption is a subtype of an intention switch. The attention switch relates a segment that has the same intention but different attention from the current segment. For instance, in Figure 6, there is an *intention switch* from DS2 to DS3, since DS3 starts a new intention (ExceptionHandling). Furthermore, U5 starts a new segment which is related to DS2 through attention switch. In addition to segment relations, there is also temporal-precedence relation between conversation units that preserves the sequence of conversation.

Data transitional relations further categorize different types of attention switches. In particular, we distinguish eight types of attention switch including

Collection-to-Instance and Instance-to-Aspect. For example, the attention is switched from a collection of houses in DS1 to a specific house in DS2 (Figure 6). Data transitional relations allow MIND to capture user data exploration patterns. Such patterns in turn can help RIA decide potential data navigation paths and provide users with an efficient information-seeking environment.

Our studies showed that, in an information-seeking environment, the conversation flow usually centers around the data transitional relations. This is different from task oriented applications where dominance and satisfaction precedence are greatly observed. In an information seeking application, the communication is more focused on the type and the actual content of information which by itself does not impose any dominance or precedence relations.

5. Context-based Multimodal Interpretation

Based on the semantic model described above, MIND uses a wide variety of contexts to interpret the rich semantics of user inputs and conversation discourse.

5.1 Turn Interpretation

To capture the overall meaning of a multimodal input at a particular turn, MIND first interprets the meanings of individual unimodal inputs (e.g., understanding a speech utterance). It then combines all different inputs using contextual information to obtain a cohesive interpretation. The first step is known as *unimodal understanding*, and the latter, *multimodal understanding*.

Unimodal Understanding. Currently, we support three input modalities: speech, text, and gesture. Specifically, we use IBM ViaVoice to perform speech recognition, and a statistics-based natural language understanding component [Jelinek et al 1994] to process the natural language sentences. For gestures, we have developed a simple geometry-based gesture recognition and understanding component. Since understanding unimodal inputs is out of the scope of this paper, next we explain how to achieve an overall understanding of multimodal inputs.

Multimodal Understanding. Traditional multimodal understanding that focuses on multimodal integration is often inadequate to achieve a full understanding of user inputs in a conversation, since users often give partial information at a particular turn. For example, in U5 (Table 1) it is not clear what exactly the user wants by just merging the two inputs together. To address these inadequacies, MIND adds context-based inference. Our approach allows MIND to use rich contextual information to infer the unspecified information (e.g., the exact intention in U5) and resolve ambiguities rising in the user input (e.g., the gestural ambiguities in U2). In particular, MIND applies two operations: fusion and inference to achieve multimodal understanding.

Fusion. *Fusion* creates an integrated representation by combining multiple unimodal inputs. In this process, MIND first merges intention structures using a set of rules. Here is one of our rules for merging intentions from two unimodal inputs:

IF I1 is the intention from unimodal input 1 &
I2 is the intention from unimodal input 2 &
(I1 has non-default values) &
(I2.Type == Refer) & (I2.Motivator == DEFAULT) &
(I2.Act == DEFAULT)
THEN Select I1 as the fused intention

It asserts that when combining two intentions together, if one is only for referral purpose (e.g., the gesture of U2 in Figure 5a, where the Act and Motivator carry the default values), then the other (e.g., the speech of U2 in Figure 5e) serves as the combined intention (e.g., the integrated Intention of U2 in Figure 4a). The rational behind this rule is that a referral action without any overall purpose most likely complements another action that carries a main communicative intention. Thus, this communicative intention is the intention after fusion. Once intentions are merged, MIND unifies the corresponding attention structures. Two attentions can be unified if and only if parameter values in one structure subsume or are subsumed by the corresponding parameter values in the other structure†. The unified value is the subsumed value (e.g., the more specific or the shared value). For example, in U2 MIND produces two combined attention structures by unifying the Attention from the speech (Figure 5f) with each Attention from the gesture (Figure 5b-d). The result of fusion is shown in Figure 7. In this combined representation, there is an ambiguity about which of the two atten-

† Value V1 subsumes value V2 if V1 is more general than V2 or is the same as V2. A special case is that a default value subsumes any other non-default values.

(a)	(b)	(c)
Intention	**Attention**	**Attention**
Motivator: DataPresentation **Act**: Request **Type**: Describe	**Base**: House **Topic**: Instance **Focus**: SpecificAspect(^Topic) **Aspect**: Price **Constraint**: [ReferredBy THIS] **Content**: [MLS0187652 | MLS0889234]	**Base**: City **Topic**: Instance **Focus**: SpecificAspect(^Topic) **Aspect**: Price **Constraint**: [ReferredBy THIS] **Content**: [Irvington]

Figure 7. Combined interpretation as a result of multimodal fusion in U2.

tion structures is the true interpretation (Figure 7b, c). Furthermore, within the attention structure for House, there is an additional ambiguity on the exact object (Content in Figure 7b). This example shows that integration resulting from unification based multimodal fusion is not adequate to resolve ambiguities. We will show later that some ambiguities can be resolved based on rich contexts.

For simple user inputs, attention fusion is straightforward. However, it may become complicated when multiple attentions from one input need to be unified with multiple attentions from another input. Suppose that the user says "tell me more about the red house, this house, the blue house," and at the same time she points to two positions on the screen sequentially. To fuse these inputs, MIND first applies temporal constraints to align the attentions identified from each modality. This alignment can be easily performed when there is an overlapping or a clear temporal binding between a gesture and a particular phrase in the speech. However, in a situation where a gesture is followed (preceded) by a phrase without an obvious temporal association as in "tell me more about the red house (deictic gesture 1) this house (deictic gesture 2) the blue house," MIND uses contexts to determine which two of the three objects (the red house, this house, and the blue house) mentioned in the speech should be unified with the attentions from the gesture.

Modality integration in most existing multimodal systems is speech driven and relies on the assumption that speech always carries the main act, and others are complementary [Bolt 1980, Burger and Marshall 1993, Zancanaro et al 1997]. Our modality integration is based on the semantic contents of inputs rather than their forms of modalities. Thus MIND supports all modalities equally as in Quickset [Johnston 1998]. For example, the gesture input in U5 is the main act, while the speech input is the complementary act for reference.

Inference. *Inference* identifies user unspecified information and resolves input ambiguities using contexts. In a conversation, users often supply abbreviated or imprecise inputs at a particular turn, e.g., abbreviated inputs given in U3, U4, U5, and the imprecise gesture input in U2 (Table 1). Moreover, the abbreviated inputs often foster ambiguities in interpretation. To derive a thorough understanding from the partial user inputs and resolve ambiguities, MIND exploits various contexts.

The domain context is particularly useful in resolving input ambiguities, since it provides semantic and meta information about the data content. For example, fusion inputs in U2 which has imprecise gesture results in ambiguities (Figure 7). To resolve the ambiguity whether the attention is a city object or a house object, MIND uses the domain context. In this case, MIND eliminates the city candidate, since cities cannot have an attribute of price. As a result, MIND understands that the user is asking about the House.

In addition to the domain context, the conversation context also provides MIND with a useful context to derive the information not specified in the user inputs. In an information seeking environment, users tend to only explicitly or implicitly specify the new or changed aspects of their information of interest without repeating those that have been mentioned earlier in the conversation. Therefore, some required but unspecified information in a particular user input can be inferred from the conversation context. For example, the user did not explicitly specify the object of interest in U4 since he has provided such information in U3. However, MIND uses the conversation context and infers that the missing object in U4 is the house mentioned in U3. In another example U5, the user specified another house but did not mention the interested aspect of this new house. Again, based on the conversation context, MIND recognizes that the user is interested in the size aspect of the new house.

RIA's conversation history is inherently a complex structure with fine-grained information (e.g., Figure 6). However, with our hierarchical structure of conversation units and segments, MIND is able to traverse the conversation history efficiently. In our example scenario, the conversation between U1 and R5 contributes to one segment (DS1 in Figure 6), whose purpose is to explore houses in Irvington. U6 starts a new segment, in which the user asked for the location of a train station, but did not specify the relevant town name. However, MIND is able to infer that the relevant town is Irvington directly from DS1, since DS1 captures the town name Irvington. Without the segment structure, MIND would have to traverse all previous 10 turns to reach U1 to resolve the town reference.

As RIA provides a rich visual environment for users to interact with, users may refer to objects on the screen by their spatial (e.g., the house at the left corner) or perceptual attributes (e.g., the red house). To resolve these spatial/perceptual references, MIND exploits the visual context, which logs the detailed semantic and syntactic structures of visual objects and their relations. More specifically, visual encoding automatically generated for each object is maintained as a part of the system conversation unit in the conversation history. During reference resolution, MIND would identify potential candidates by mapping the referring expressions with the internal visual representation. For example, the object which is highlighted on the screen (R5) has an internal representation that associates the visual property Highlight with the object identifier. This allows MIND to correctly resolve referents for *it* in U7. In this reference resolution process, based on the Centering Theory

[Grosz et al 1995], MIND first identifies the referent most likely to be the train station since it is the preferred center in the previous utterance. However, according to the domain knowledge, such a referent is ruled out since the train station does not have the attribute of bedrooms. Nevertheless, based on the visual context, MIND recognizes a highlighted house on the screen. An earlier study indicates that objects in the visual focus are often referred by pronouns, rather than by full noun phrases or deictic gestures [Kehler 2000]. Therefore, MIND considers the object in the visual focus (i.e., the highlighted house) as a potential referent. In this case, since the highlighted house is the only candidate that satisfies the domain constraint, MIND resolves the pronoun *it* in U7 to be that house. Without the visual context, the referent in U7 would not be resolved.

Furthermore, the user context provides MIND with user profiles. A user profile is established through two means: explicit specification and automated learning. Using a registration process, information about user preferences can be gathered such as whether the school district is important. In addition, MIND can also learn user vocabularies and preferences based on real sessions between a user and RIA. Currently, we are investigating the use of user context for interpretation. One attempt is to use this context to map fuzzy terms in an input to precise query constraints. For example, the interpretation of the term *expensive* or *big* varies from one user to another. Based on different user profiles, MIND can interpret these fuzzy terms as different query constraints. Finally, the environment context provides device profiles that facilitate response generation. For example, if a user uses a PDA to interact with RIA, MIND would present information in a summary rather than an elaborated textual format because of the limited display capability.

5.2 Discourse Interpretation

While turn interpretation derives the meanings of user inputs at a particular turn, *discourse interpretation* identifies the contribution of user inputs toward the overall goal of a conversation. In particular, during discourse interpretation MIND decides whether the input at the current turn contributes to an existing segment or starts a new one. In the latter case, MIND also decides where to add the new segment and how this segment relates to existing segments in a conversation history. To make these decisions, MIND first calculates the semantic distances between the current turn and existing segments. Based on the distances, MIND then interprets how the turn is related to the overall conversation.

Some previous works on discourse interpretation are based on the shared plan model [Lochbaum 1998, Rich and Sidner 1998] where specific plans and recipes are defined for the applications. In an information seeking application, since users can freely browse or navigate information space, it would be difficult, if not impossible, to come up with a generic navigation plan. Therefore, our approach is centered around user information needs such as the desired operations on information, the type of information and the finer aspects of information. Specifically, our discourse interpretation is based on intention and attention that captures user information needs, and the discourse structure reflects the overall exchanged information at each point in the conversation. This discourse structure provides MIND an overall picture about what information has been conveyed, and thus guide MIND in more efficient information navigation (e.g., decide on what information needs to be delivered). At the core of this approach, is the semantic distance measurement.

Measuring Semantic Distance. Semantic distance measures the closeness of user information needs captured in a pair of intention/attention. For example, a user first requests for the information about the size of a house, and after a few interactions, he asks about the price of the same house. In this case, although there are a few interactions between these two requests, the second request is closely related to the first request since they both asking specific aspects about the same house object. Therefore, the semantic distance between intention/attention representing those two requests is small. For another example, suppose the user asks about the price of a house, and then in the next turn, he asks RIA to compare this house with a different house. Although these two requests are adjacent in the conversation, they are quite different since the first requests for data presentation and the second asks for data comparison. So the semantic distance between those two requests is larger than that in the first example.

Furthermore, since MIND consistently represents intention and attention in both conversation units and conversation segments, the semantic distance can be extended to measure the information needs represented in a new conversation unit and those represented in existing conversation segments. This measurement allows MIND to identify the closeness between a new information need (from an incoming user input) with other information exchanges in the prior conversation. By relating similar information needs together using the semantic distance measurement, MIND is able to construct a space of communicated information and its inter-relations.

Specifically, to measure the semantic distance between a user conversation unit and a segment, MIND compares their corresponding Intention and Attention. As in the following formula, the distance between two intentions (I_u and I_s) or attentions (A_u and A_s) is a weighted sum of distances between their corresponding parameters as the following, where w_i

(w_j) is the weight and d_i (d_j) is the parametric distance for each parameter i in Intention (or j in Attention).

$$\textit{Intention}: \ \text{Distance}(I_u, I_s) = \sum_i w_i \cdot d_i$$

$$\textit{Attention}: \ \text{Distance}(A_u, A_s) = \sum_j w_j \cdot d_j$$

Different weights help promote/demote the significance of different parameters in the distance measuring. For example, MIND assigns the highest weight to Motivator in Intention, since it manifests the main purpose of an input. Likewise, Aspect in Attention is given a least weight since it captures a very specific dimension of the content. To compare two parameters, MIND currently performs a binary comparison. That is, if two parameter values are equal or one value subsumes the other, the parametric distance is 0, otherwise 1. Once the semantic distance between a conversation unit and a conversation segment is computed, MIND determines the relationship between them using interpretation rules. Note that currently, our weights are manually assigned. In the future, those weights can be trained over a labeled corpus.

Applying Interpretation Rules. To determine how the current user input is related to the existing conversation, MIND first calculates the semantic distance between the conversation unit representing the current user input and every existing segment. Based on these distances, MIND will then choose the segment that is the closest and apply a set of rules to decide how the current unit relates to this segment. Specifically, these rules use a set of thresholds to help determine whether this unit belongs to the existing segment or starts a new segment. For example, when U2 is encountered, MIND first calculates the semantic distance between U2 and DS1 (the only existing segment at this point). Since the distance measurement satisfies the conditions in Ruleset 2(a) (Figure 8), a new segment DS2 is generated. Furthermore, Ruleset 2(b) helps MIND identify that DS2 is dominated by DS1, since the content of DS2 (MLS0187652 or MLS0889234, which is copied from the current turn) is a part of DS1 (a collection that includes both MLS0187652, MLS0889234). This structure indicates that, up to this point in the conversation, the overall purpose is presenting a collection of houses in Irvington, and this overall purpose contains a sub-purpose which is presenting a particular house in this collection. Similarly, for U4 MIND calculates the distance between U4 and three existing segments (DS1, DS2 and DS3). In this case, since DS2 is the closest, MIND attaches U4 to DS2 (Figure 6) according to Ruleset 1(a).

Ruleset 1:

(a) **IF** Unit U with Intention I_u and Attention A_u
& Segment S with Intention I_s and Attention A_s
& $Distance(I_u, I_s) < t_1$ & $Distance(A_u, A_s) < t_2$
THEN Add Unit U to Segment S & Update S

Ruleset 2:

(a) **IF** Unit U with Intention I_u and Attention A_u
& Segment S_1 with Intention I_{s1} and Attention A_{s1}
& $Distance(I_u, I_{s1}) < t_1$ & $t_2 < Distance(A_u, A_{s1}) < t_3$
THEN Create a new segment S_2
& Copy I_u and A_u to I_{s2} and A_{s2} & Add unit U to S_2

(b) **IF** $I_{S2}.content$ is a part of $I_{S1}.content$
THEN S_1 dominates S_2

Figure 8. Examples of interpretation rules

Our current approach to discourse interpretation relies on our fine-grained model of Intention and Attention. Different applications may require understanding the conversation at different levels of granularity (the granularity of segments). To accommodate different interpretation needs, MIND can vary the weights in the distance measurement and adjust the thresholds in the interpretation rules.

6. Evaluation

We have developed MIND as a research prototype. The modeling scheme and interpretation approach are implemented in Java. The prototype is currently running on Linux.

Our initial semantic models and interpretation algorithms were driven by a user study we conducted. In this study, one of our colleagues acted as RIA and interacted with users to help them find real estate in Westchester county. The analysis of the content and the flow of the interaction indicates that our semantic models and interpretation approaches are adequate to support these interactions.

After MIND was implemented, we conducted a series of testing on multimodal fusion and context-based inference (focusing on domain and conversation contexts). The testing consisted of a number of trials, where each trial was made up by a sequence of user inputs. Half of these inputs were specifically designed to be ambiguous and abbreviated. Since the focus of the testing was not on our language model, we designed the speech inputs so that they could be parsed successfully by our language understanding components. The testing showed that once the user speech input was correctly recognized and parsed, in about 90% of those trials, the overall meanings of user inputs were correctly identified. However, speech recognition is a bottleneck in MIND. To improve the robustness of MIND, we need to enhance the accuracy of speech recognition and improve the coverage of the language model. We plan to do more vigorous evaluations in the future.

7. Conclusions and Future Work

To support a full-fledged multimodal conversation, we have built MIND, which unifies multimodal input understanding and discourse interpretation. In particular, MIND has two unique features. The first is a

fine-grained semantic model that characterizes the meanings of user inputs and the overall conversation from multiple dimensions. The second is an integrated interpretation approach that identifies the semantics of user inputs and the overall conversation using a wide variety of contexts. These features enable MIND to achieve a deep understanding of user inputs.

Currently, multimodal fusion (for intention) and discourse interpretation rules are constructed based on typical interactions observed from our user study. These rules are modality independent. They can be applied to different information seeking applications such as searching for computers or cars. Our future work includes exploring learning techniques to automatically construct interpretation rules and incorporating confidence factors to further enhance input interpretation.

8. Acknowledgements

We would like to thank Keith Houck for his contributions on training models for speech/gesture recognition and natural language parsing, and Rosario Uceda-Sosa for her work on RIA information server.

References

Allen, J., D. Byron, M. Dzikovska, G. Ferguson, G. L., and A. Stent (2001) Toward conversational human computer interaction. *AI Magazine*, 22(4):27–37.

Bolt, R. A. (1980) Voice and gesture at the graphics interface. *Computer Graphics*, pages 262–270.

Burger, J. and R. Marshall. (1993) The application of natural language models to intelligent multimedia. In M. Maybury, editor, *Intelligent Multimedia Interfaces*, pages 429–440. MIT Press.

Cohen, P., M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow (1996) Quickset: Multimodal interaction for distributed applications. *Proc. ACM MM'96*, pages 31–40.

Grosz, B. J., A. K. Joshi, and S. Weinstein (1995) Towards a computational theory of discourse interpretation. *Computational Linguistics*, 21(2):203–225.

Grosz, B. J. and C. Sidner (1986) Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jelinek, F., J. Lafferty, D. M. Magerman, R. Mercer, and S. Roukos (1994) Decision tree parsing using a hidden derivation model. *Proc. Darpa Speech and Natural Language Workshop*, March.

Johnston, M. (1998) Unification-based multimodal parsing. *Proc. COLING-ACL'98.*

Johnston, M. and S. Bangalore (2000) Finite-state multimodal parsing and understanding. *Proc. COLING'00*.

Kehler, A. (2000) Cognitive status and form of reference in multimodal human-computer interaction. *Proc. AAAI'01*, pages 685–689.

Lambert, L. and S. Carberry (1992) Modeling negotiation subdialogues. *Proc. ACL'92*, pages 193–200.

Litman, D. J. and J. F. Allen. (1987) A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11:163–200.

Lochbaum, K. (1998) A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

Neal, J. G., C. Y. Thielman, Z. Dobes, S. M. Haller, and S. C. Shapiro (1998) Natural language with integrated deictic and graphic gestures. In M. Maybury and W. Wahlster, editors, *Intelligent User Interfaces*, pages 38–52.

Rich, C. and C. Sidner (1998) Collagen: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction*.

Stent, A., J. Dowding, J. M. Gawron, E. O. Bratt, and R. Moore (1999) The commandtalk spoken dialog system. *Proc. ACL'99*, pages 183–190.

Wahlster, W. (1998) User and discourse models for multimodal communication. In M. Maybury and W. Wahlster, editors, *Intelligent User Interfaces*, pages 359–370.

Wahlster, W. (2000). Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final Verbmobil system. *Verbmobile*, pages 3–21.

Zancanaro, M., O. Stock, and C. Strapparava (1997) Multimodal interaction for information access: Exploiting cohesion. *Computational Intelligence*, 13(4):439–464.

Zhou, M. X. and S. Pan (2001) Automated authoring of coherent multimedia discourse for conversation systems. *Proc. ACM MM'01*, pages 555–559.

A General Purpose Architecture for Intelligent Tutoring Systems

Brady Clark, Elizabeth Owen Bratt, Oliver Lemon, Stanley Peters, Heather Pon-Barry, Zack Thomsen-Gray, & Pucktada Treeratpituk
Stanford University
Center for the Study of Language Information
Stanford CA 94305-4115 USA
{bzack,ebratt,lemon,peters,ponbarry,ztgray,pucktada}@csli.stanford.edu

Abstract

The goal of the Conversational Interfaces project at CSLI is to develop a general purpose architecture which supports multi-modal dialogues with devices. Our systems use a common software base consisting of the Open Agent Architecture, Nuance speech recogniser, Gemini (SRI's parser and generator), Festival speech synthesis, and CSLI's "Conversational Intelligence Architecture" (CIA). This paper focuses on one application of this architecture — an automated tutor for shipboard damage control. We discuss the benefits of adopting this architecture for intelligent tutoring.

Keywords

dialogue system architecture, tutorial dialogue, intelligent tutoring system, multimodal

1 Introduction

Multi-modal, activity-oriented dialogues with devices present a challenge for dialogue system developers. Conversational interaction in these contexts is mixed-initiative and open-ended. Consider dialogue with an intelligent tutoring system (ITS). Dialogue can be unpredictable in tutorial interactions. The user may need to query the system; e.g., ask a definitional question. Further, the tutor must have a way of reacting to various types of user input; e.g., by adjusting the tutorial agenda when the student asks for clarification about past topics of discussion.

In this paper we discuss a new general purpose architecture for intelligent dialogue systems which addresses these issues: the Conversational Intelligence Architecture (CIA) developed at CSLI.

The CIA has previously been used in a dialogue system for multi-modal conversations with a robot helicopter (the WITAS system; Lemon et al. 2001, 2002). We focus on a parallel deployment of this architecture in the domain of automated tutoring. We will first discuss the ITS we are developing for shipboard damage control. Next, we discuss the CIA for dialogue systems and what benefits it has for intelligent tutoring.

2 An Intelligent Tutoring System for Damage Control

Shipboard damage control refers to the task of containing the effects of fire, explosions, and other critical events that can occur aboard Naval vessels. The high-stakes, high-stress nature of this task, together with limited opportunities for real-life training, make damage control an ideal target for AI-enabled educational technologies like intelligent tutoring systems.

We are developing an intelligent tutoring system for automated critiquing of student performance on a damage control simulator (Clark et al. 2001). The simulator is DC-Train (Bulitko and Wilkins 1999), an immersive, multimedia training environment for damage control. DC-Train's training scenarios simulate a mixture of physical phenomena (e.g., fire) and personnel issues (e.g., ca-

sualties). Figure 1 provides a sample of the type of tutorial interaction our system aims to support.

Conversation with automated tutors places the following requirements on dialogue management (see Lemon et al. 2001, Clark 1996):

1. *Flexibility*: user and system input should be interpreted as dialogue moves, even when they are not predictable in advance

2. *Open-ended*: there are not rigid predetermined goals for the dialogue

3. *Mixed-initiative*: in general, both the user and the system should be able to introduce topics

In the next section, we discuss a general purpose architecture for dialogue systems which meets these two demands.

3 An Architecture for Multi-modal Dialogue Systems

To facilitate the implementation of multimodal, mixed-initiative interactions we use the Open Agent Architecture (OAA) (Martin et al. 1999). OAA is a framework for coordinating multiple asynchronous communicating processes. The core of OAA is a 'facilitator' which manages message passing between a number of encapsulated software agents that specialize in certain tasks (e.g., speech recognition).

Our system uses OAA to coordinate the following agents:

1. The **Gemini** NLP system (Dowding et al. 1993). Gemini uses a single unification grammar both for *parsing* strings of words into logical forms (LFs) and for *generating* sentences from LF inputs. This agent enables us to give precise and reliable meaning representations which allow us to identify dialogue moves (e.g., *question*) given a linguistic input; e.g., the question "What happened next?" has the LF: `(ask(wh([past,happen])))`.

2. A **Nuance** speech recognition server, which converts spoken utterances to strings of words. The Nuance server relies on a language model, which is compiled directly from the Gemini grammar, ensuring that every recognized utterance is assigned a LF.

3. The **Festival** text-to-speech system, which 'speaks' word strings generated by Gemini.

4. The **Conversational Intelligence Architecture**, which coordinates inputs from the user, interprets the user's dialogue moves, updates the dialogue context, and delivers speech and graphical outputs to the user. This agent is discussed in Section 4.

Agents 1-3 are 'off-the-shelf' dialogue system components (apart from the Gemini grammar, which must be modified for each application). The CIA agent (4) was written in Java for dialogue management applications in general. This OAA/Gemini/Nuance/Festival/CIA architecture has also been deployed successfully in Lemon et al. 2001, 2002.

4 Dialogue Managment

Dialogue Management with the CIA makes use of several recent ideas in dialogue modeling, described in detail in Lemon et al. (2001, 2002). It creates and updates an *Information State*, corresponding to a notion of dialogue context. Dialogue moves (e.g., *assertion, wh-question*) update information states. A user's dialogue move might send a response to the system, elicit an assertion by the system, or prompt a follow-up question. The system itself generates dialogue moves which are treated just like the user's conversational contributions.

The CIA includes the following dynamically updated components (see Lemon et al. 2002 for full details):

1. The *Dialogue Move Tree:* a structured history of dialogue moves and 'threads', plus a list of 'active nodes'

T: There were two fires, and you put out both of them. You also handled one flood. The first fire was in the officer's baggage room. For a fire, what's the first thing you should do?

S: Send a repair team to set fire boundaries on the compartment.

T: No, you missed a step. What comes before setting fire boundaries?

S: I'm not sure.

T: Okay. First send a team to investigate. For the second fire, you also forgot to send a team to investigate.

T: Okay. After setting fire boundaries, what should you do?

S: Order a repair team to fight the fire.

T: No, you missed a step. What should you do before sending a repair team to fight the fire?

S: Isolate the space.

T: Yes. Always electrically and mechanically isolate a space before fighting a fire. For the second fire, you also forgot to isolate the space.

Figure 1: A Sample dialogue: T=tutor/system, S=student

2. The *Activity Tree*: a temporal and hierarchical structure of activities initiated by the system or the user, plus their execution status
3. The *System Agenda*: the issues to be raised by the system
4. *Salience List*: the objects referenced in the dialogue thus far, ordered by recency
5. The *Pending List*: the questions asked but not yet answered
6. The *Modality Buffer*: stores gestures for later resolution

The Dialogue Move Tree (DMT) provides a representation of dialogue context in terms of a structured history of dialogue moves. Further, the DMT determines whether or not user input can be interpreted in the current dialogue context, and how to interpret it. Recall the three requirements placed on automated tutors discussed in Section 2. The DMT structure is able to interpret user and system input as dialogue moves, even when they are not predictable in advance (*flexibility*). Further, the DMT can handle dialogues with no clear endpoint (*open-ended*). The CIA supports the third requirement of *mixed-initiative* by way of the System Agenda and generation component. In the next section, we discuss further benefits of the CIA for intelligent tutoring systems, both in the domain of shipboard damage control and in general.

5 Benefits of the Conversational Intelligence Architecture

The CIA has the following useful properties:

1. It embodies Clark's (1996) joint activity theory of dialogue, in which dialogue serves the activity the conversational participants are engaged in. By utilizing the Dialogue Move Tree/Activity Tree distinction, we are able to provide a model of the joint activities (the Activity Tree), and follow the structure of dialogue deployed in service of those activities (the Dialogue Move Tree).
2. While other intelligent tutoring systems employ finite-state automata which constrain the dialogue move option space for any input (e.g., AutoTutor; Graesser et

al. 2000), our CIA is not a finite-state machine, and is dynamically updated. The latter property is useful for handling unpredictable input and for shifting the agenda in response to user input.

3. The Dialogue Move Tree provides us with a rich representation of dialogue structure, which allows us to return to past topics of discussion in a principled, orderly way. For example, in the domain of shipboard damage control, the automated tutor might compare the handling of a later crisis to the handling of earlier crises. Further, the student might ask for clarification about the reasons for earlier actions, so we would like to be able to return to the earlier topic, and pick up the context at that point, as well as simply referring to the earlier crisis.

4. Dialogue moves used in the different implementations of the CIA are domain-general, and thus reusable across different domains. We are building a library of dialogue moves for use by any type of dialogue system. For example, tutorial dialogue will share with other systems dialogue moves such as *questions* and *answers*, but not others (e.g., *hints*).

5. The architecture separates dialogue management from "back-end" activities, such as robot control or tutorial strategies. In the tutorial case, it provides a high-level representation of the tutorial strategies (in the form of the Activity Tree) accessible by the Dialogue Move Tree.

6. The architecture supports multimodality by way of the Modality Buffer. For example, we are able to coordinate speech input and output with gestural input and output (e.g., the user can indicate a point on a map with a mouse click or the system can highlight a map region).

Acknowledgments

This work is supported by the Department of the Navy under research grant N000140010660, a multidisciplinary university research initiative on natural language interaction with intelligent tutoring systems.

References

Bulitko, V.V. and D.C. Wilkins. 1999. Automated instructor assistant for ship damage control. *Proceedings of AAAI-99*.

Clark, B., J. Fry, M. Ginzton, S. Peters, H. Pon-Barry, and Z. Thomsen-Gray. A Multi-Modal Intelligent Tutoring System for Shipboard Damage Control. *Proceedings of IPNMD-2001*: 121-125.

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

Dowding, J., J. Gawron, D. Appelt, J. Bear, L. Cherny, R.C. Moore and D. Moran. 1993. Gemini: A natural language system for spoken-language understanding. *Proceedings of the ARPA Workshop on Human Language Technology*.

Graesser, A., K. Wiemer-Hastings, P. Wiemer-Hastings, R. Kreuz and the Tutoring Research Group. 2000. AutoTutor: a simulation of a human tutor. *Journal of Cognitive Systems Research*. 1: 35-51.

Lemon, O., A. Bracy, A. Gruenstein and S. Peters. 2001. Information States in a Multimodal Dialogue System for Human-Robot Conversation. *Proceedings Bi-Dialog, 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, pages 57 - 67, 2001.

Lemon, O., A. Gruenstein and S. Peters. 2002. Collaborative Activities and Multitasking in Dialogue Systems *Traitement Automatique des Langues (TAL, special issue on dialogue)*, ed. Claire Gardent, (to appear).

Martin, D., A. Cheyer and D. Moran. 1999. The Open Agent Architecture: a framework for building distributed software systems. *Applied Artificial Intelligence* 13, 1-2.

Perceptive Animated Interfaces:
The Next Generation of Interactive Learning Tools

Ron Cole
Center for Spoken Language Research
University of Colorado, Boulder

We envision a new generation of human computer interfaces that engage users in natural face-to-face conversational interaction with intelligent animated characters. In specific learning domains, these perceptive animated interfaces will process auditory and visual signals presented by the user (e.g., speech sounds, lip movements, facial expressions, hand and body gestures) to interpret the user's spoken utterances and infer the user's intentions and cognitive state (e.g., focused, excited, frustrated). In addition, the system will use this information to build user models relevant to the learning task— e.g., if the system is helping the student learn about space science, it will analyze the student's responses to discover the student's preconceptions within that domain. Based on this information, the animated agent will orient to the user, provide real time feedback when the user speaks, and respond to the user much like a sensitive and effective teacher—through speech, facial expressions and hand and body gestures.

Research on perceptive animated interfaces at CSLR builds on advanced spoken dialogue systems research supported by DARPA and NSF. This research uses CU Communicator, an environment for researching and developing spoken dialogue systems that enable completely natural, unconstrained, mixed-initiative spoken dialogues in specific task domains. Communicator uses the public domain Galaxy hub-server architecture developed by the MIT speech group and maintained by MITRE under DARPA support. Spoken dialogue interaction in Communicator occurs via communication between users and various technology servers (all developed at CSLR) that pass messages through the Galaxy hub—audio server, speech recognizer, semantic parser, language generator, speech synthesizer, dialogue manager, and back-end servers that communicate with Web sites. By adding computer vision and computer animation servers, we have transformed Communicator into a platform for research and development of perceptive animated interfaces.

Our research on perceptive animated interfaces occurs in the context of Interactive Books: powerful learning tools that reside on client machines, and communicate with servers running Communicator. Interactive Books employ full-bodied 3D animated characters that integrate auditory and visual processing so the animated character can orient to the user, interpret the user's auditory and visual behaviors, and respond to these behaviors using speech, facial expressions and gestures. By studying the behaviors of master teachers working with individual students, and by working with these teachers to incorporate their best practices into our learning tools, we hope to invent animated agents that interact with students much like sensitive and effective teachers.

Our presentation will demonstrate the capabilities of Interactive Books, discuss the major research challenges involved, and describe ongoing work applying these learning tools to a number of learning domains, including foundational speech and reading skills, comprehension training and science education.

On The Relationships Among Speech, Gestures, And Object Manipulation In Virtual Environments: Initial Evidence

Andrea Corradini, Philip R. Cohen
Center for Human-Computer Communication
Department of Computer Science and Engineering
Oregon Health & Science University, Portland, OR, USA
{andrea,pcohen}@cse.ogi.edu

Abstract

This paper reports on a study whose goal was to investigate how people make use of gestures and spoken utterances while playing a videogame without the support of standard input devices.

We deploy a Wizard of Oz technique to collect audio-video- and body movement-related data on people's free use of gesture and speech input. Data was collected from eleven subjects for up to 60 minutes of game interaction each. We provide information on preferential mode use, as well as the predictability of gesture based on the objects in the scene.

The long-term goal of this on-going study is to collect natural and reliable data from different input modalities, which could provide training data for the design and development of a robust multimodal recognizer.

1. Introduction

Human-computer interaction in virtual environments has long been based on gesture, with the user's hand(s) being tracked acoustically, magnetically, or via computer vision. In order to execute operations in virtual environments, users often are equipped with datagloves, whose handshapes are captured digitally, or a tracked device that is equipped with multiple buttons. For a number of reasons, these systems have frequently been difficult to use. First, although the user's hand/arm motions are commonly called "gesture," the movements to be recognized are typically chosen by the developer. Thus, rather than recognize people's naturally occurring movements, such systems require users to learn how to move "properly." Secondly, gestural devices have many buttons and modes, making it difficult for a naïve subject to remember precisely which button in a given mode accomplishes which function. Third, the 3D interaction paradigm usually derives from the 2D-based direct manipulation style, in which one selects an object and then operates upon it. Some systems have modeled the virtual environment interface even more strongly upon the WIMP (windows, icons, menus, pointing device) graphical user interface, providing users with menus that need to be manipulated in the 3D world. Unfortunately, it turns out to be very difficult to select objects and menu entries in 3D environments.

Various researchers have attempted to overcome these awkward interfaces in different ways. For example, Hinckley [1] gave users real-world models to manipulate, causing analogous actions to take place in the virtual environment. Stoakley et al [2] provided a miniature copy of the virtual world in the user's virtual hand, thereby allowing smaller movements in hand to have analogous results on the world itself. Fisher et al. [3] developed an early multimodal 3D interface for simulated Space Station operations, incorporating limited speech recognition, as well as hand gestures using a VPL dataglove. Weimer and Ganapathy developed a prototype virtual environment interface [4] that incorporated a VPL dataglove, and a simple speech recognizer. Although only three gestures, all by the user's thumb, were recognized, and the speech system offered just a 40 word vocabulary, the authors remarked upon the apparent improvement in the interface once voice was added.

Based in part on this prior research, we hypothesize that multimodal interaction in virtual environments can ease the users' burden by distributing the communicative tasks to the most appropriate modalities. By employing speech for its strengths, such as asking questions, invoking actions, etc., while using gesture to point at locations and objects, trace paths, and manipulate objects, users can more easily engage in virtual environment interaction. In order to build such multimodal systems, we need to understand how, if at all, people would speak and/or gesture in virtual environments if given the choice. What would users do on their own, without being limited to the researchers' preferred gestures and language? Would gestures and language be predictable, and if so from what? Can the recognition of gesture and/or speech in virtual environments be improved by recognizing or

understanding the input in another modality, as we find in 2D map-systems [5, 6].

Regarding predictability of speech and gesture, we hypothesize that without instruction, people will manipulate manufactured objects in VE in the ways they were designed to be manipulated – using their affordances [7], [8]. Given data indicating a user's viewpoint on the object, and the degrees of freedom afforded by the object, a system should be able to predict how the user's hand/arm will move. If the user can also speak, will they employ the same gestures during multimodal interaction as they employ using gesture alone?
In order to answer these questions, and to provide a first set of data for training recognizers and statistically-basd multimodal integration systems, we conducted a Wizard-of-Oz study of multimodal interaction with a simple, though realistically rendered, computer game.

2. Study

Eleven volunteer subjects (ten adults, one 12 year-old child), interacted with the Myst™ III game played on a 2GHz Dell computer with Nvidia GeForce 3 graphics card. Myst is a semi-immersive 2.5-dimensional game in which a user moves around a complex world, containing both indoor and outdoor scenes. The user views the world through a (moderate) fish-eye viewport, which s/he can rotate 360 degrees, as well as tilt to see above and below. In Myst III, the user's task is (partially) to travel around an island, rotating a series of beacons so that they shine on one another in a certain sequence, etc. Thus, the game involves navigation, manipulation of objects (doors, a knife switch, beacons, push-buttons, etc.) and search.

The subjects wore a set of four 3D trackers attached to their head and dominant arm. They were told that they could interact with the game as they wished, and that the system could understand their speech and gestures.

2.1 Wizard of Oz study

The classic method for studying recognition-based systems before the appropriate recognizers have been trained is to employ a Wizard of Oz paradigm [9]. In this kind of study, an unseen assistant plays the role of the computer, processing the user's input and responding, as the system should. Importantly, the response time should be rapid enough to support satisfactory interactive performance. In the present study, subjects were told that they would be playing the Myst III computer game, to which they could speak and gesture freely. The user chose where s/he wanted to stand.

Subjects "played" the game standing in front of a 50" diagonal flat-panel plasma display in wide-view mode. They could and did speak and/or gesture without constraint. Unbeknownst to the subject, a researcher observed the subjects' inputs, and controlled the game on a local computer, whose audio and video output was sent to the subject. This "wizard's" response time averaged less than 0.5 seconds. Since Myst III (and its predecessors) assumes the user is employing a mouse, it

Figure 2: Example of experimental setup, utterance, gesture, and humanoid reproducing the subject's motions

is designed to minimize actual gesturing, allowing only mouse-selection. Although occasionally the Wizard made errors, subjects received no explicitly marked recognition errors. A research assistant was present in the room with the subject, and would upon request give the subject hints about how to play the game, though not about what to say or how to gesture.

2.2 Equipment

To acquire the gesture data, the six-degree-of-freedom Flock of Bird (FOB) magnetic tracking device from Ascension Technology Corporation [10] was used. We attached four sensors to the subject; one on the top head, one on the upper arm to register the position and orientation of the humerus, one on the lower arm for the wrist position and lower arm orientation, and finally one on the top of the hand. The last three sensors are aligned with each other anytime the subject stretched his or her arm to the side of the body, keeping the palm of the hand facing and parallel to the ground (see Figure 3).

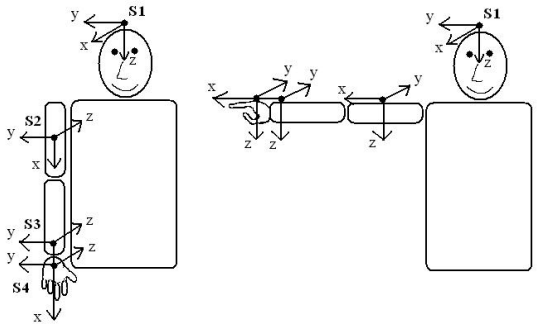


Figure 3: Arrangement of trackers on subject's body

The data from the FOB are delivered via serial lines, one for each sensor. The four data streams can be processed in real time by a single SGI Octane machine employing the Virtual Reality Peripheral Network (VRPN) package [11], which provides time-stamps of the data and distributes it to customer processes. Because the FOB devices uses a magnetic field that is affected by metallic objects, and the laboratory is constructed of steel-reinforced concrete, the data from the sensors is often distorted. As a result, the data is processed with a median filter to partially eliminate noise. A digital "humanoid" plays back the sensor data, providing both a check on accuracy and distortion.

3. Data Analysis

The subject's body motions were captured by the FOB, while the video recorded the subject's view, and vocal interaction (see Figure 2). The speech and gesture on the video were transcribed, an example of which follows:

TRANSCRIPT FROM GAME

Bold = speech
= location of gesture when not overlapping speech
(…) = hesitation/pause
XXX = undecipherable
[ ] = speech-gestural stroke overlapped event
Indications such as "08-01-42-25/44-11" = VCR time-code

08-01-42-25/44-11:**# go across the bridge**
[hand held palm open to point toward the bridge then hand used as cursor along bridge]

08-01-47-00: **keep going**
no gesture

08-01-50-03/57-27:**[grab] this thing (…) just [grab it] and pull it down and see what happen #**
reach for rim of telescope with hand, [close fist and pull hand from up to down], [one more time], [one more time]

08-01-59-29/07-04:**# can I pull this thing ? (…) ah ahaa #**
reach for rim of telescope, [close fist, pull hand from left to right circularly], [one more time]

08-02-10-00/15-16:**ok [look at] look at this purple and see if there is anything to see**
[move hand toward the purple ball as to push at it]

08-02-15-16/18-28:**# no (...) [back]**
move hand toward lens of telescope, [close fist to grab at rim of telescope and pull hand back toward the body], [move open hand again back toward the body]

08-02-19-04/26-25:**ok # [turn it] again #**
reach for rim of telescope, [close fist and pull hand from left to right circularly as to rotate rim of telescope], [one more time], [one more time]

The transcript includes both speech to the system, as well as self-talk, but not requests for hints asked of the research assistant.

3.1 Coding

The following categories were coded: For events that required explicit interaction with the system beyond causing the scene to rotate around, subjects were coded as using gesture-only, speech-only, or multimodal interaction. Numerous subcategories were coded, but this paper only reports on the subset of gesture and multimodal interaction for which the user employed gesture "manipulatively," when interacting with an

object. Interrater reliability for second-scoring of 18% of the multimodal data was 98%.

– *Speech manipulative + gesture NOT manipulative:* like above but the gesture does not match the way the object functions.

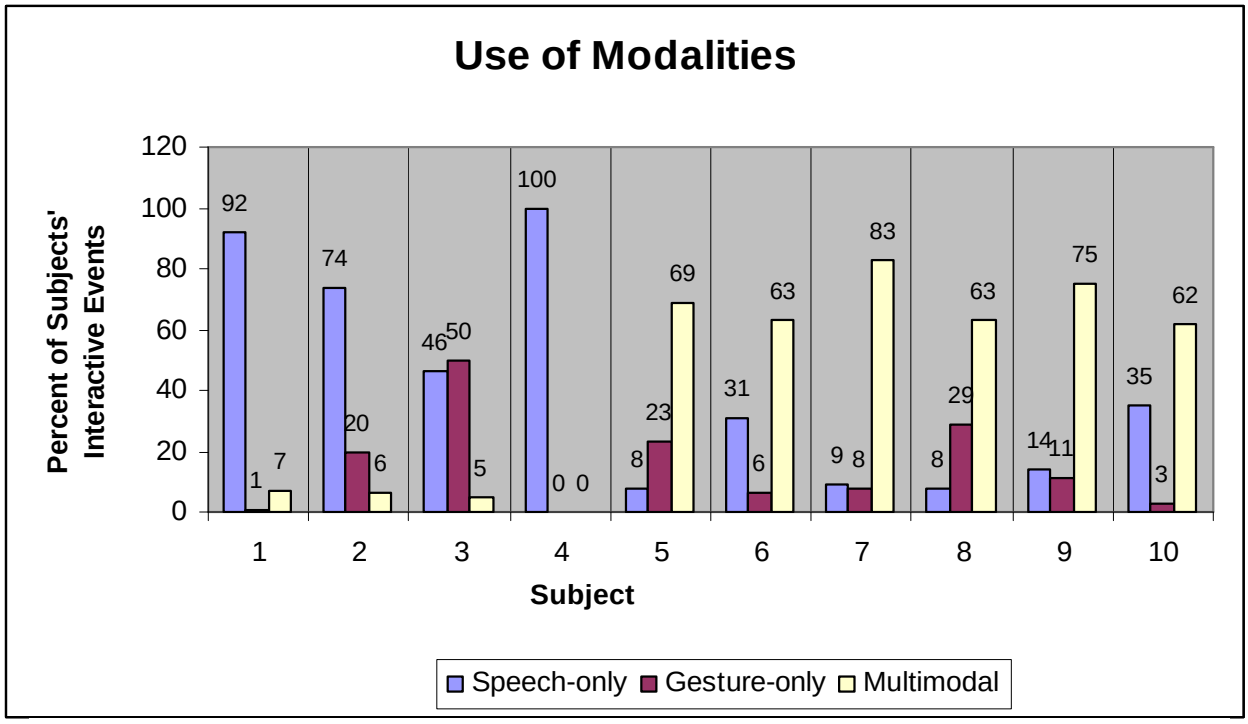


Figure 4: Subjects' use of modalities

For ONLY GESTURE:

– *Consistent manipulative gesture*: gesture used with the objective of changing the state of an object in the game --- e.g., turning a wheel or pressing a button. Such gestures are consistent if the movement matches the way the object operates.

– *Manipulative gesture, NOT consistent:* see above but the movement does not match the way the operate operates.

For ONLY SPEECH:

– *Speech manipulative*: involving any change of state of an object in the scene --- e.g., standing in front of the wheel and saying "turn the wheel", or "press the button" when in the elevator.

For SPEECH AND GESTURE TOGETHER

– *Speech manipulative + consistent manipulative gesture*: e.g. saying "turn the wheel" AND mimicking the gesture of turning a wheel.

4. Results

Of the 3956 "interactive events," we totaled the use of gesture-only, speech-only, or multimodal interaction (see Figure 4). Subjects were classified as "habitual users" of a mode of communication if they employed that mode during at least 60% of the available times. As can be seen, 90% of the subjects were habitual users of speech (including multimodal interaction), and 60% of the subjects were habitual users of multimodal interaction.

Subjects were classified as "habitual users" of the *consistent manipulation of objects strategy* if they gesturally manipulated the object at least 10 times in a given communication mode (using gesture-only or multimodally) and 60% of those times, their body motion was consistent with the affordances of that object (i.e., the ways humans would normally employ it). Figure 5 provides Myst III images that show some of those objects, whose affordances the reader can readily determine.

We examined the two cases separately: use of gesture-only and use of gesture within a multimodal event. Six subjects used gesture within their multimodal interaction manipulatively, rather than deictically, in order to change the state of an object. Of those six, five gestured in a fashion consistent with the object's affordances, indicating that subjects using multimodal interaction manipulated digital "artifacts" according to the actions afforded by their design (Wilcoxon, $p<0.03$, $Z= -1.89$, one-tailed). The one subject who did not do so used a speak-and-point strategy ("turn the wheel" <point>).

As for gesture alone, four subjects changed the state of an object with manipulative gesture used in a fashion consistent with the object affordances, whereas no subjects using manipulative gestures inconsistent with the object's affordances (Wilcoxon, $p=0.023$, $Z= -2.000$ one-tailed).

Figure 5 Other objects that can be manipulated in Myst III

4.1 Results from Questionnaire

The subjects filled out a post-test questionnaire (See Appendix) indicating, on scales of 1(very low) -10 (very high), their:

1. Experience with adventure games ($\bar{x} = 2.3$, $\sigma = 1.8$)
2. Immersive experience ($\bar{x} = 4.2$, $\sigma = 2.5$)
3. Involvement in the game ($\bar{x} = 4.1$, $\sigma = 2.6$)
4. Ranking of the interface ($\bar{x} = 6.1$, $\sigma = 1.7$)
5. System response latency ($\bar{x} = 5.3$, $\sigma = 2.6$)

In other words, the game itself was not particularly exciting, and subjects were only moderately involved in it. However, the speech/gesture interface received better than middling ratings, while the latency inherent in a Wizard-of-Oz did not make the interaction annoyingly slow.

4.2 Comments from subjects

Subjects were asked, "Do you think there was a 'preferred' channel between speech or gesture for the interaction? Or did you feel both channels were equally effective? Please explain". In response, we received the following comments, which often did not match the subjects' performance:

– "I feel that speech was easier for commands, since it is more economical in energy than gesture, and more precise. I grew tired of trying to motion precisely with my arms, and the time it took to turn. I would prefer to use speech over gesture, at least with the way gesture seemed to be implemented here." *Used speech-only 49%, gesture-only 45% of the time.*

– "I am confident that anything I was doing, I could do with either channel. Speech seemed more natural for some situations, and gestures for others, though." *A habitual user of speech-only interaction, using gesture 20% of the time, speech-only 74%.*

– "Preferred (channel) seemed to be speech. The movement felt like I was doing exercises." *A habitual user of speech (100%).*

– "I found voice commands worked better for me than the hand gestures." *A habitual user of multimodal interaction (62%), with no gesture-only interaction.*

– "I think the preferred channel is a gesture, with speech used to make clarifications." *A habitual user of multimodal interaction (83%).*

– "In the first few minutes, I realized that gesturing worked best for me. I always talk to my computer or my car while operating them, so that was more unconscious....I find it very involving to talk to the game...". *A habitual user of multimodal interaction (75%).*

– "I was under the distinct impression that gesture did nothing and the system was a speech-recognition one". A *habitual user of multimodal interaction (69%).*

– "I liked hands but used speech when confusing". *A habitual user of multimodal interaction (63%).*

5. Discussion

Results show that if given the opportunity, most subjects (60%) would use multimodal interaction more than 60% of the time for interacting with the game; an additional 30% would use speech-only interaction 60% of the time or more, but only 10% would even use gesture half the time. Overall, subjects were found to employ gesture alone 14% of the time, speech alone 41% of the time, and used speech and gesture for 45% of their interactions. These latter results are somewhat skewed by subjects who employed only speech, since in order to navigate in the scene, they issued many more navigation commands (e.g., "go left") than would be necessary if multimodal interaction were employed.

Subjects' opinions about which modalities were important were often belied by their actions. Some thought that gesture was the key modality, but used multimodal interaction habitually, while others thought speech was the essential modality, but also used multimodal interaction frequently. It would appear that having both available would suit just about all the subjects.

When given the opportunity to use whatever gestures they wanted, most subjects who attempted to manipulate an artifact generally did so according to the affordances of those objects – subjects turned wheels, pulled down knife switches, pushed in doorbells, etc. Thus, a future virtual environment control program that detects that a manufactured object is in the scene should be able to predict how a person's arm would be shaped and would move in order to manipulate that object properly. The gesture recognizer could then adapt to the scene itself, giving such gestures higher weight.

As novices to these kinds of games, the subjects found Myst III to be a modestly immersive experience with a moderate degree of involvement. Unlike true 3D games, there is no avatar in the scene that represents the user, and thus one would expect a lesser degree of immersion than a full 3D virtual reality environment. Also, given that the subjects were not supplied with explicit objectives in playing the game, it is not surprising that their degree of involvement with the game was moderate. A number of subjects very much liked the speech-gesture interface, while most gave it mixed reviews.

Finally, two female subjects who stood close to the screen experienced some degree of nausea or instability. Virtual environment illness is usually associated with 3D head-mounted displays or wrap-around environments [12]. Various factors affect VE illness, but it is not known precisely which are the most important. Factors that have been identified as possible contributors, and which may have been at work here include: subject gender (females are more prone), angle of view, distortion, response latency, refresh rate, and phosphor lag [13].

6. Related work

Most of the existing research on gestures has been performed by cognitive scientists who are interested in how people gesture, and the reasons why people gesture ([14-17]). Various taxonomies of gesture have been offered (e.g, McNeill's description of iconic, deictic, emblematic, and beat gestures [14], which usefully inform scientists who build gesture-based systems and avatars [18, 19]. Specific claims can also be useful to technologists. McNeill [14] argues that speech and gesture derive from an internal knowledge representation (termed a "catchment") that encodes both semantic and visual information. Our results tend to confirm this claim, in that the visual representation depicts the object's affordances, which then determines the manipulative gesture. The present corpus could also be used to confirm Kendon's [17] claim that the stroke phase of gestures tends to co-occur with phonologically prominent words.

Quek et al. [20] have provided a case-study of cross-modal cues for discourse segmentation during natural conversation by observing a subject describing her living space to an interlocutor. A comparative analysis of both video and audio data was performed to extract the segmentation cues while expert transcribed the gestures. Then both the gestural and spoken data was correlated for 32 seconds of video. A strong correlation between handedness and high-level semantic content of the discourse was found, and baseline data on the kinds of gestures used was provided.

Whereas this style of observational research is needed as a foundation, it needs to be combined with quantitative observational and experimental work in order to be useful to building computer systems. Wilson et al. [19] employed the McNeill theory to motivate their research to distinguish bi-phasic gestures (e.g, beats) from more meaningful tri-phasic gestures that have preparatory, stroke, and retraction phases.

Early work by Hauptmann [21] employed a Wizard of Oz paradigm and investigated the use of multimodal interaction for simple 3D tasks. It was found that people

prefer to use combined speech and gesture interaction over either modality alone, and given the opportunity to do so in a factorially designed experiment, chose to use both 70% of the time (vs. 13% gesture only, and 16% speech only). Their factorial study has the advantage that all subjects were exposed to all ways of communicating, whereas our more ecologically realistic study allowed users to develop their own ways of interacting. On the other hand, it also allowed users to become functionally fixed into their first successful way of operating.

Perhaps the most relevant work to ours is that of Sowa and Wachsmuth [22] who employed a WOZ paradigm to collect subjects' gestures as they attempt to describe a limited set of objects to a listener within a virtual construction domain. It was found that subjects' hand shapes corresponded to features of the objects themselves. Based on this data, a prototype system was built that decomposes each gesture into spatial primitives and identifies their interrelationships. The object recognition engine then employs a graph-matching technique to compare the structure of the objects and that of the gesture.

This latter work differs from the results presented here in that the game we studied included the task of manipulating the object rather than describing it. Thus, we find people attempt to manipulate artifacts in the ways they were built to be manipulated, whereas Sowa and Wachsmuth found people used their hands in describing objects to indicate their shape. Clearly, the subjects' goals and intentions make a difference to the kinds of gestures they use, whether in a multimodal or unimodal context.

7. Future work

The next steps in this research are to analyze the syntax and semantics of the utterances in conjunction with the form and meaning of the gesture. To date, we have employed unification as the primary information fusion operation for multimodal integration [5, 23]. We hypothesize that by using a feature-structured action representation that offers a "manner" attribute [24], whose value can itself be an action, and by representing the meanings of gestures as such actions, unification can again serve as the mechanism for information fusion across modalities. The analyses of utterance and gesture will be used to test this hypothesis. Given the identification of the same kind of gesture by the subjects to manipulate the same object, a gesture recognizer can then be trained [25] and used in a multimodal architecture. We also will investigate to what extent mutual disambiguation of modalities [6] can be used to overcome recognition errors in a complete multimodal virtual environment system [26].

8. Conclusions

This paper has provided initial empirical results based on a Wizard of Oz paradigm that indicate people will prefer to use multimodal interaction in virtual environments than speech or gesture alone. Furthermore, when they attempt to manipulate virtual representations of artifacts, they do so according to the objects' affordances – the ways the objects were designed to be manipulated. We therefore conclude that virtual environment systems should be capable of multimodal interaction that would employ subjects' natural gestures, and that such systems should in fact employ information about the objects currently in the user's view to predict the kinds of gestures that will ensue. In the future, information about the users' gaze could be employed to restrict still further the kinds of gestures that might be employed.

9. Acknowledgments

This research has been supported by the Office of Naval Research, under Grants N00014-99-1-0377, N00014-99-1-0380 and N00014-02-1-0038. We are thankful to Rachel F. Coulston for help in running the study, and to our volunteer subjects.

10. References

1. Hinckley, K., Pausch, R., Goble, J.C., and Kassell, N.F. Passive real-world interface props for neurosurgical visualization. In *Conference on Human Factors in Computing Systems (CHI '94)*. 1994.
2. Stoakley, R., Conway, M.J., and Pausch, R. Virtual reality on a WIM: interactive worlds in miniature. In *ACM Conference on Human Factors in Computing Systems (CHI '95)*. 1995: ACM Press.
3. Fisher, S.S., McGreevy, M., Humphries, J., and Robinett, W. Virtual environment display system. In *ACM Workshop on Interactive 3D Graphics*. 1986, ACM Press: Chapel Hill, N.C. p. 77-87.
4. Weimer, D., and Ganapathy, S. K. A synthetic visual environment with hand gesturing and voice input. In *Proceedings of Human Factors in Computing Systems (CHI'89)*, K. Bice, and Lewis,

C., Editor. 1989, ACM Press: New York. p. 235-240.
5. Cohen, P.R., Johnston, M., McGee, D.R., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. QuickSet: multimodal interaction for distributed applications. In *Fifth Annual ACM International Multimedia Conference (Multimedia '97)*. 1997. Seattle, WA: ACM Press.
6. Oviatt, S.L. Mutual disambiguation of recognition errors in a multimodal architecture. In *Human Factors in Computing Systems (CHI'99)*. 1999. New York: ACM Press.
7. Gibson, J.J. The theory of affordances. In *Perceiving, acting, and knowing*, R. Shaw, and J.Bransford, Eds. 1977, Erlbaum Associates: Hillsdale, N. J.
8. Norman, D.A. The Design of Everyday Things. 1988, New York, NY: Currency/Doubleday. 257.
9. Oviatt, S.L., Cohen, P.R., and Wang, M.Q. Toward Interface Design for Human Language Technology: Modality and Structure as Determinants of Linguistic Complexity. Speech Communication, 1994. **15**(3-4).
10. Ascension Technology, *Flock of Birds*, 2002. http://www.ascensiontech.com/
11. Taylor, R.M. VRPN: A Device-Independent, Network-Transparent VR Peripheral System. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. 2001, ACM Press: New York.
12. Kolasinski, E.M., Goldberg, S. L., and Hiller, J. H. Simulator sickness in virtual environments. 1995, US Army Research Institute for the Behavioral and Social Sciences: Alexandria, VA.
13. Pausch, R., Crea, T., and Conway, M. A literature survey for virtual environments: Military flight simulator visual systems and simulator sickness. *Presence*, 1992. **1**(3): p. 344-363.
14. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*. 1993, Chicago: University of Chicago Press.
15. Krauss, R.M. Why do we gesture when we speak? *Current directions in Psychological Science*, 1998. **7**: p. 54-59.
16. Goldin-Meadow, S.The role of gesture in communication and thinking. *Trends in Cognitive Science*, 1999. **3**: p. 419-429.
17. Kendon, A. Language and Gesture: Unity or Duality? In *Language and Gesture*, D. McNeill, Editor. 2000, Cambridge University Press: Cambridge, UK. p. 47-63.
18. Cassell, J., and Stone, M. Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. In *Proceedings of the AAAI 1999 Fall Symposium on Psychological Models of Communication in Collaborative Systems*. 1999, AAAI Press: North Falmouth, MA. p. 34-42.
19. Wilson, A.D., Bobick, A.F., and Cassell, J. Temporal classification of natural gesture and application to video coding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1997, IEEE Press: New York. p. 948-954.
20. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., and Ansari, R. Gesture and Speech Multimodal Conversational Interaction. 2001, Electrical Engineering and Computer Science Department, University of Illinois: Chicago.
21. Hauptmann, A.G. Speech and gestures for graphic image manipulation. In *Proceedings of Human Factors in Computing Systems (CHI'89)*. 1989, ACM Press: New York. p. 241-245.
22. Sowa, T., Wachsmuth, I. Interpretation of Shape-Related Iconic Gestures in Virtual Environments. In *Gesture and Sign Language in Human-Computer Interaction*. 2002, Springer Verlag: Berlin. p. 21-33.
23. Johnston, M., Cohen, P.R., McGee, D.R., Oviatt, S.L., Pittman, J.A., and Smith, I. Unification-based multimodal integration. In *35th Annual Meeting of the Association for Computational Linguistics (ACL '97)*. 1997. Madrid, Spain: ACL Press.
24. Badler, N., Bindinganavale, R., Allbeck, J., Schuler, W., Zao, L., and Palmer, M. Paramaterized action representatin for virtual human agents. In *Embodied conversational agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. 2000, The MIT Press: Cambridge, Massachusetts. p. 256-286.
25. Corradini, A. Real-time gesture recognition by means of hybrid recognizers. In *Gesture and Sign Language in Human-Computer Interaction*. 2002, Springer Verlag: Berlin. p. 34-46.
26. Cohen, P. R., McGee, D., Oviatt, S. L., Wu, L., Clow., J., King, R., Julier, S., Rosenblum, L., Multimodal interaction for 2D and 3D environments. *IEEE Computer Graphics and Applications*, July/August 1999, pp.10-13

11. Appendix: Questionnaire MYST III - EXILE

PRE-SESSION:

1) Have you ever played MYST III: Exile - the sequel to the MYST and RIVEN series ?

2) Have you ever played adventure or puzzle-like games?

3) How would you rank your experience in playing the game on a scale from 1 to 10 (1 = not experience at all - 10 = very experienced player)

4) Would you say that your current state of fitness is: as usual, you are sick, or both ? Please explain.

POST-SESSION:

5) Did you like the game ? Please explain.

6) Would you like to play again ? Please explain.

7) What did you find difficult when playing ? Please explain.

8) How did you feel physically during the game ? If you did not feel well, please report your symptoms.

9) If you did not feel well, would you say you had

    a. nausea,
    b. vomiting,
    c. eyestrain,
    d. disorientation,
    e. ataxia, (a kind of inability to coordinate voluntary muscular movements)
    f. vertigo
    g. disturbed locomotion,change in postural control,
    h. perceptual-motor disturbances,
    i. past pointing,
    j. flashbacks,
    k. drowsiness,
    l. fatigue,
    m. lowered arousal or mood
    n. pallor
    o. cold sweating
    p. increased salivation
    q. headache
    r. flushing
    s. dizziness
    t. other (describe)

10) How would you rank the immersive effect of the game on a scale from 1 to 10 (1 = not immersive at all – 10 = amazingly immersive) ?

11) How would you rank your involvement in playing the game on a scale from 1 to 10 (1 = I played because I was requested to – 10 = I fully wanted to get to the end of the game) ?
12) How would you rank the interface on a scale from 1 to 10 (1 = extremely bad, it never did what I wanted to – 10 = amazingly good) ?

13) How would you rank the latency response of the interface on a scale from 1 to 10 (1 = when I entered a command the interface reacted too slowly – 10 = I did not even realize there was a latency between my commands and the response of the interface)

14) Would you have preferred to play using a mouse/joystick/trackball/keyboard instead the way you played ? Please explain.

15) Did you consciously decide when to use speech or gesture for playing ? Please explain.

16) Did you have some hints concerning if and how to use gesture or speech or did you just do what it come naturally to you ? Please explain.

17) Do you think there was a “preferred” channel between speech or gesture for the interaction ? Or did you feel both channels were equally effective ? Please explain.

18) Were the sensors attached to your body cumbersome ? Did you feel restricted in your movements? Please explain.

19) Do you have any suggestions/criticism concerning the experiment ?
20)

Geometric and statistical approaches to audiovisual segmentation for untethered interaction

T. Darrell, J. Fisher, and K. Wilson
Artificial Intelligence Laboratory
M.I.T., Cambridge, MA 02139 USA
trevor,fisher,kwilson@ai.mit.edu

Keywords: source separation, vision tracking, microphone array, mutual information.

Abstract

Multimodal approaches are proposed for segmenting multiple speakers using geometric or statistical techniques. When multiple microphones and cameras are available, 3-D audiovisual tracking is used for source segmentation and array processing. With just a single camera and microphone, an information theoretic criteria separates speakers in a video sequence and associates relevant portions of the audio signal. Results are shown for each approach, and their integration discussed as future work.

1 Introduction

Conversational dialog systems have become practically useful in many application domains, including travel reservations, traffic information, and database access. However most existing conversational speech systems require *tethered* interaction, and work primarily for a single user. Users must wear an attached microphone or speak into a telephone handset, and do so one at a time. This limits the range of use of dialog systems, since in many applications users might expect to freely approach and interact with a device. Worse, they may wish to arrive as a group, and talk among themselves while interacting with the system. To date it has been difficult for speech recognition systems to handle such conditions, and correctly recognize the utterances intended for the device.

Given only a single sensing modality, and perhaps only a single sensor, disambiguating the audio from multiple speakers can be a challenge. But with multiple modalities, and possibly multiple sets of sensors, segmentation can become feasible. In this paper we present two methods for audiovisual segmentation of multiple speakers, based on geometric and statistical source separation approaches.

We have explored two configurations which are of practical interest. The first is based on a "smart environment" or "smart room" enabled with multiple stereo cameras and a ceiling mounted large-aperture microphone array grid. In this configuration users can move arbitrarily through the room or environment while focused audiovisual streams are generated from their appearance and utterance. In the second configuration we presume a single omnidirectional microphone and single camera is available. This is akin to what one might find on a PDA or cellphone, or low-cost PC videoconferencing installation.

In a multi-sensor environment we use a geometric approach, and use multi-view image correspondence and tracking methods combined with acoustic beamforming techniques. A multimodal approach can track sources even in acoustically reverberant environments with dynamic illumination, conditions that are tough for audio or video processing alone.

When only a single multimodal sensor pair (audio and video) is available we use a statistical approach, jointly modelling audio and video variation to identify cross-modal correspondences. We show how this approach can detect which user is speaking when several are facing a device. This allows the segregation of users' utterances from each other's speech, and from background noise events.

We first review related work, and then present our method for geometric source separation and vision-guided microphone array processing. We then describe our single camera/microphone method for audiovisual correspondence using joint statistical processing. We show results with each of these techniques, and describe how the two approaches may be integrated in future work.

2 Related Work

Humans routinely perform tasks in which ambiguous auditory and visual data are combined in order to support accurate perception. In contrast,

automated approaches for statistical processing of multi-modal data sources lag far behind. This is primarily due to the fact that few methods adequately model the complexity of the audio/visual relationship. Classical approaches to multi-modal fusion at a signal processing level often either assume a statistical relationship which is too simple (e.g. jointly Gaussian) or defer fusion to the decision level when many of the joint (and useful) properties have been lost. While such pragmatic choices may lead to simple statistical measures, they do so at the cost of modelling capacity.

An information theoretic approach motivates fusion at the measurement level without regard to specific parametric densities. The idea of using information-theoretic principles in an adaptive framework is not new (e.g. see [8] for an overview) with many approaches suggested over the last 30 years. A critical distinction in most information theoretic approaches lies in how densities are modelled (either explicitly or implicitly), how entropy (and by extension mutual information) is approximated or estimated, and the types of mappings which are used (e.g. linear vs. nonlinear). Approaches which use a Gaussian assumption include Plumbley [19, 18] and Becker[1]. Additionally, [1] applies the method to fusion of artificially generated random dot stereograms.

There has been substantial progress on feature-level integration of speech and vision. For example, Meier *et al* [17], Stork [25] and others have built visual speech reading systems that can improve speech recognition results dramatically. Our system, described below, is designed to be able to detect and disambiguate cases where audio and video signals are coming from different sources. Other audio/visual work which is closely related to ours is that of Hershey and Movellan [13] which examined the per-pixel correlation relative to an audio track, detecting which pixels have related variation. Again, an inherent assumption of this method was that the joint statistics were Gaussian. Slaney and Covell [21] looked at optimizing temporal alignment between audio and video tracks using canonical correlations (equivalent to mutual information in the jointly Gaussian case), but did not address the problem of detecting whether two signals came from the same person.

Several authors have explored geometric approaches to audiovisual segmentation using array processing techniques. Microphone arrays are a special case of the more general problem of sensor arrays, which have been studied extensively in the context of applications such as radar and sonar [22]. The Huge Microphone Array project[20] is investigating the use of very large arrays containing hundreds of microphones. Their work concentrates on audio-only solutions to array processing. Another related project is Wang and Brandstein's audio-guided active camera[24], which uses audio localization to steer a camera on a pan/tilt base.

A number of projects [2, 3, 4] have used vision to steer a microphone array, but because they use a single camera to steer a far-field array, they cannot obtain or make use of full 3-D position information; they can only select sound coming from a certain direction.

We are exploring both a geometric and statistical approach to audiovisual segmentation. In the next section we describe our geometric approach, based on microphone and camera arrays. Following that we present our statistical approach, using an information theoretic measure to relate single channel audio and video signals.

3 Multi-modal multi-sensor domain

The association between sound and location makes a microphone array a powerful tool for audiovisual segmentation. In combination with additional sensors and contextual information from the environment, a microphone array can effectively amplify and separate sounds of interest from complex background noise.

To focus a microphone array, the location of the speaker(s) of interest must be known. A number of techniques exist for localizing sound sources using only acoustic cues [23], but the performance of these localization techniques tends to degrade significantly in the presence of reverberation and/or multiple sound sources. Unfortunately, most common office and meeting room environments are highly reverberant, with reflective wall and table surfaces, and will normally contain multiple speakers.

However, in a multimodal setting we can take advantage of other sensors in the environment to perform localization of multiple speakers despite reverberation. We use a set of cameras to track the position of speakers in the environment, and report the relative geometry of speakers, cameras, and microphones.

The vision modality is not effected by acoustic reverberation, but its accuracy will depend on the the calibration and segmentation procedures. In practice we use video information to restrict the range of possible acoustic source locations to a region small enough to allow for acoustic localization techniques to operate without severe problems with reverberation and multiple speakers.

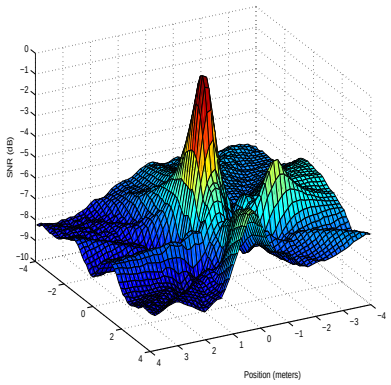


Figure 1: Array power response as a function of position (two speakers). This plot shows the array output power as the array's focus is scanned through a plane centered on one speaker while another speaker is nearby. The central speaker is easily discernible in the plot, but the peak corresponding to the weaker speaker is difficult to distinguish among the sidelobe peaks. Using vision-based person tracking cues can disambiguate this case.

3.1 Microphone array processing overview

Many problems can be addressed through array processing. The two array processing problems that are relevant to our system are beamforming and source localization.

Beamforming is a type of spatial filtering in which the signals from individual array elements are filtered and added together to produce an output that amplifies signals coming from selected regions of space and attenuates sounds from other regions of space. In the simplest form of beamforming, delay-and-sum beamforming, each channel's filter is a pure delay. The delay for each channel is chosen such that signals from a chosen "target location" are aligned in the array output. Signals from other locations will tend to be combined incoherently.

Source localization is a complementary problem to beamforming whose goal is to estimate the location of a signal source. One way to do this is to beamform to all candidate locations and to pick the location that yields the strongest response. This method works well, but the amount of computation required to do a full search of a room is prohibitively large. Another method for source localization consists of estimating relative delays among channels and using these delays to calculate the location of the source. Delay-estimation techniques are computationally efficient but tend to perform poorly in the presence of multiple sources and/or reverberation.

For microphone arrays that are small in size compared to the distance to the sources of interest, incoming wavefronts are approximately planar. Because of this, only source direction can be determined; source distance remains ambiguous. When the array is large compared to the source distance, the sphericity of the incoming wavefronts is detectable, and both direction and distance can be determined. These effects of array size apply both to localization and to beamforming, so if sources at different distances in the same direction must be separated, a large array must be used. As a result, with large arrays the signal-to-noise ratio (for a given source) at different sensors will vary with source location. Because of this, signals with better signal-to-noise ratios should be weighted more heavily in the output of the array. Our formulation of the steering algorithm presented below takes this into account.

3.2 Person tracking overview

Tracking people in known environments has recently become an active area of research in computer vision. Several person-tracking systems have been developed to detect the number of people present as well as their 3D position over time. These systems use a combination of foreground/background classification, clustering of novel points, and trajectory estimation over time in one or more camera views [7, 15].

Color-based approaches to background modelling have difficulty with illumination variation due to changing lighting and/or video projection. To overcome this problem, several researchers have supported the use of background models based on stereo range data [7, 14]. Unfortunately, most of these systems are based on computationally intense, exhaustive stereo disparity search.

We have developed a system that can perform dense, fast range-based tracking with modest computational complexity. We apply ordered disparity search techniques to prune most of the disparity search computation during foreground detection and disparity estimation, yielding a fast, illumination-insensitive 3D tracking system. Details of our system are presented in [6]. Our system reports the 3-D position of people moving about an environment equipped with an array of stereo cameras.

3.3 Vision-guided acoustic volume selection

We perform both audio localization and beamforming with a large, ceiling-mounted microphone array. Localization uses information from both audio and video, while beamforming uses only the audio data and the results of the localization pro-

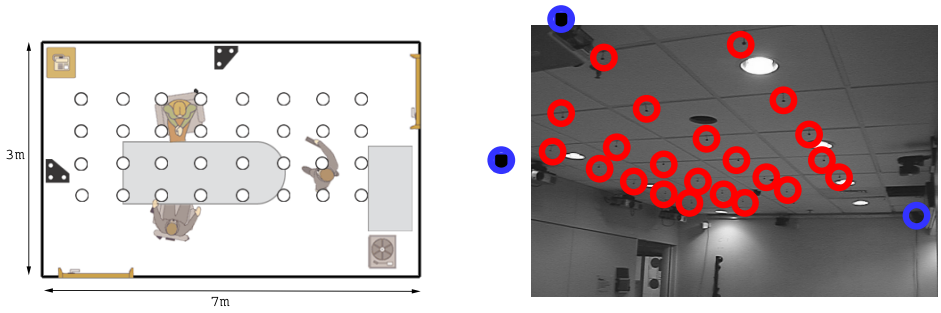


Figure 2: The test environment. On the left is a schematic view of the environment with stereo cameras represented by black triangles and microphones represented by empty circles. On the right is a photograph of the environment with microphones and camera locations highlighted.

cessing. A large array gives the ability to select a *volume* of 3-D space, rather than simply form a 2-D beam of enhanced response as anticipated by the standard array localization algorithms. However, the usual assumption that of constant target signal-to-noise ratio (SNR) across the array does not hold when the array geometry is large (array width on same scale as target distance.)

Our system uses the location estimate from the vision tracker as the initial guess from which to begin a gradient ascent search for a local maximum in beam output power. Beam power is defined as the integral over a half-second window of the square of the output amplitude. The vision tracker is accurate to within less than one meter. Gradient ascent to the nearest local maximum can therefore be expected to converge to the location of the speaker of interest when no other speakers are very close by.

For small microphone arrays, the relative SNRs of the individual channels do not vary significantly as a function of source location. This is, however, not true for larger microphone arrays. For our array, which is roughly 4 meters across, we must take into account the fact that some elements will have better signals than others. Specifically, if we assume that we have signals x_1 and x_2 which are versions of the unit-variance desired signal, s, that have been contaminated by unit-variance uncorrelated noise, we can analyze the problem as follows:

$$x_1 = a_1 s + n_1$$
$$x_2 = a_2 s + n_2$$

In this model, the signal to noise ratios of x_1 and x_2 will be a_1^2 and a_2^2, respectively. Their optimal linear combination will be of the form $y = bx_1 + x_2$. Because of the uncorrelated noise assumption, the SNR of this combination will be

$$SNR(y) = \frac{(ba_1 + a_2)^2}{b^2 + 1}$$

By taking the derivative of this expression with respect to b and setting the result equal to zero, one finds that the optimal value of b is:

$$b = \frac{a_1}{a_2} = \frac{SNR(x_1)}{SNR(x_2)}$$

Because of the symmetry of the signals, this result implies that, in general, individual elements' signals should be scaled by a constant proportional to the square root of their SNRs.

Ideally, we would like to have complete knowledge of the strengths and statistical relationships among the noise signals at the individual sensors. This information is not easy to obtain, but because of our large array and multiple stereo cameras, it is easy for us to use our location estimate to weight individual channels assuming a $1/r$ attenuation due to the spherical spreading of the source. Assuming $1/r$ attenuation from a source to each microphone, we have $a_n = 1/r_n$ in the above equations, so the optimal weighting factor for channel n is $1/r_n$. This is intuitively appealing since it means that microphones far from the source contribute relatively little to the array output.

4 Results

Our test environment, depicted in Figure 2, is a conference room equipped with 32 omnidirectional microphones spread across the ceiling and 2 stereo cameras on adjacent walls.

The audio and video subsystems were calibrated independently, and for our experiments, we performed a joint calibration by finding the least-squares best-fit alignment between the two coordinate systems.

Figure 1 is an example of what happens when multiple speakers are present in the room. Audio-only gradient ascent could easily find one of the undesirable local maxima. Because our vision-

	SNR (dB)
Distant microphone	−6.6
Video only	−4.4
Audio only (dominant speaker)	2.0
Audio-Video	2.3

Table 1: Audio-video localization performance.

Interferer:	None	-24 dB	-12 dB	0 dB
Lapel mic.	98	100	98	83
Mic. array	95	94	90	24
Distant mic.	78	73	38	1

Table 2: Sentence Recognition Rates (percent correct). Each recognition rate was calculated from 36 queries evenly divided between two male speakers. The close-talking microphone was clipped to the lapel of the speaker. The microphone array is as described above. The distant microphone is one array element from near the center of the room.

based tracker is accurate to within one meter, we can safely assume that we will find the correct local maximum even in the presence of interferers.

To validate our localization and source separation techniques, we ran an experiment in which two speakers spoke simultaneously while one of them moved through the room. We tracked the moving speaker with the stereo tracker and processed the corresponding audio stream using three different localization techniques. For each, we used a reference signal collected with a close-talking microphone to calculate a time-averaged SNR (Table 1). For performance comparison we use the signal from a single distant microphone near the center of the room. This provides no spatial selectivity, but for our scenario it tends to receive the desired speech more strongly than the interfering speech. The SNR for the single microphone case is negative because of a combination of the interfering speaker and diffuse noise from the room's ventilation system.

To evaluate the microphone array's effects on recognition rates for automated speech recognition (ASR), we connected our system to the MIT Spoken Language Systems (SLS) Group's JUPITER weather information system [26]. We had two male speakers issue each of nine weather-related queries from two different locations in the room. As collected, the data contains quiet but audible noise from the ventilation system in the room. To evaluate the results under noisier conditions, additional noise was added to these signals. The results are shown in Table 2.

The beamformed signal from the microphone array was in all cases superior to the single distant microphone. The distant microphone, which was approximately 1.5m from the speaker, yielded recognition rates that were too low to be useful in our current environment.

Our rough estimates of the signal power to noise power ratios for both the close-talking microphone and the distant microphone are about 10 dB. This suggests that in our scenario with no interferer, the primary benefit of the microphone array is that it reduces the proportion of reverberant energy in the signal.

The 0 and -12 dB interferer significantly degraded the performance of the array. We are currently working on adaptive null-steering algorithms that should improve performance in the presence of stronger interferers such as this.

These experiments demonstrate that audio-video localization is superior to video alone in our environment. We believe our approach improves upon audio-only localization in cases where there are multiple simultaneous speakers and the reverberant energy is nearly equal or greater than the direct path energy. The initial position estimate provided by video localization reduces the amount of computation required compared to an unconstrained audio-only search.

5 Single multi-modal sensor domain

In the single (multimodal) sensor domain geometry is less useful, and array processing impossible; in this case we instead exploit audiovisual joint statistics to localize speakers. We adopt the paradigm of looking at a single camera view, and seeing what information from a single microphone can tell us about that view (and vice-versa.)

We propose an independent cause model to capture the relationship between generated signals in each individual modality. Using principles from information theory and nonparametric statistics we show how an approach for learning maximally informative joint subspaces can find cross-modal correspondences. We analyze the graphical model of multi-modal generation and show under what conditions related subcomponents of each signal have high mutual information.

Non-parametric statistical density models can be used to measure the degree of mutual information in complex phenomena [12] which we apply to audio/visual data. This technique simultaneously learns projections of images in the video sequence *and* projections of sequences of periodograms taken from the audio sequence. The projections are computed adaptively such that the video and audio projections have maximum mutual information (MI).

We first review the basic method for audio-visual fusion and information theoretic adaptive methods, see [9] for full details. We present our probabilistic model for cross-modal signal generation, and show how audio-visual correspondences can be found by identifying components with maximal mutual information. In an experiment comparing the audio and video of every combination of a group of eight users, our technique was able to perfectly match the corresponding audio and video for each user.

Finally, we show a new result on a monocular speaker segmentation task where we segment the audio between several speakers seen by the camera. These results are based purely on the instantaneous cross-modal mutual information between the *projections* of the two signals, and do not rely on any prior experience or model of user's speech or appearance.

5.1 Probabilistic models of audio-visual fusion

We consider multimodal scenes which can be modelled probabilistically with one joint audio-visual source and distinct background interference sources for each modality. Each observation is a combination of information from the joint source, and information from the background interferer for that channel. In contrast to the array processing case, we explicitly model visual appearance variation, not just 3-D geometry. We use a graphical model (Figure 3) to represent this relationship. In the diagrams, B represents the joint source, while A and C represent single modality background interference. Our purpose here is to analyze under which conditions our methodology should uncover the underlying cause of our observations.

Figure 3a shows an independent cause model for our typical case, where $\{A, B, C\}$ are unobserved random variables representing the causes of our (high-dimensional) observations in each modality $\{X^a, X^v\}$. In general there may be more causes and more measurements, but this simple case can be used to illustrate our algorithm. An important aspect is that the measurements have dependence on only one common cause. The joint statistical model consistent with the graph of figure 3a is $P(A, B, C, X^a, X^v) = P(A)P(B)P(C)P(X^a|A, B)P(X^v|B, C)$.

Given the independent cause model a simple application of Bayes' rule (or the equivalent graphical manipulation) yields the graph of figure 3b which is consistent with $P(A, B, C, X^a, X^v) = P(X^a)P(C)P(A, B|X^a)P(X^v|B, C)$, which shows that information about X^a contained in X^v is conveyed through the *joint* statistics of A and B. The consequence being that, in general, we cannot disambiguate the influences that A and B have on the measurements. A similar graph is obtained by conditioning on X^v. Suppose decompositions of the measurement X^a and X^v *exist* such that the following joint densities can be written: $P(A, B, C, X^a, X^v) = P(A)P(B)P(C)P(X^a_A|A)P(X^a_B|B)P(X^v_B|B)P(X^v_C|C)$ where $X^a = [X^a_A, X^a_B]$ and $X^v = [X^v_B, X^v_C]$. An example for our specific application would be segmenting the video image (or filtering the audio signal). In this case we get the graph of Figure 3c and from that graph we can extract the Markov chain which contains elements related only to B. Figure 3d shows equivalent graphs of the extracted Markov chain. As a consequence, there is no influence due to A or C.

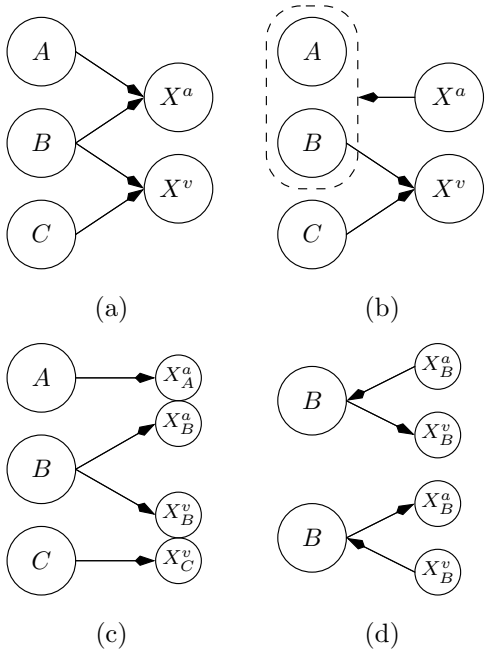


Figure 3: Graphs illustrating the various statistical models exploited by the algorithm: (a) the independent cause model - X^a and X^v are independent of each other conditioned on $\{A, B, C\}$, (b) information about X^a contained in X^v is conveyed through *joint* statistics of A and B, (c) the graph implied by the existence of a separating function, and (d) two equivalent Markov chains which can be extracted from the graphs *if* the separating functions can be found.

Of course, we are still left with the formidable task of finding a decomposition, but given the decomposition it can be shown, using the data processing inequality [5], that the following inequality

holds:

$$I(X_B^a, X_B^v) \leq I(X_B^a, B) \quad (1)$$
$$I(X_B^a, X_B^v) \leq I(X_B^v, B) \quad (2)$$

More importantly, these inequalities hold for functions of X_B^a and X_B^v (e.g. $Y^a = f(X^a; h_a)$ and $Y^v = f(X^v; h_v)$). Consequently, by maximizing the mutual information between $I(Y^a; Y^v)$ we must necessarily increase the mutual information between Y^a and B and Y^v and B. The implication is that fusion in such a manner discovers the underlying cause of the observations, that is, the joint density of $p(Y^a, Y^v)$ is strongly related to B. Furthermore, with an approximation, we can optimize this criterion without estimating the separating function directly. In the event that a perfect decomposition does not exist, it can be shown that the method will approach a "good" solution in the Kullback-Leibler sense.

From the perspective of information theory, estimating separate projections of the audio video measurements which have high mutual information makes intuitive sense as such features will be predictive of each other. The advantage is that the form of those statistics are not subject to the strong parametric assumptions (e.g. joint Gaussianity) which we wish to avoid.

We can find these projections using a technique that maximizes the mutual information between the projections of the two spaces. Following [10], we use a nonparametric model of joint density for which an analytic gradient of the mutual information with respect to projection parameters is available. In principle the method may be applied to any function of the measurements, $Y = f(X; h)$, which is differentiable in the parameters h (e.g. as shown in [10]). We consider a linear fusion model which results in a significant computational savings at a minimal cost to the representational power (largely due the nonparametric density modelling of the output):

$$\begin{bmatrix} y_1^v \cdots y_N^v \\ y_1^a \cdots y_N^a \end{bmatrix} = \begin{bmatrix} h_v^T & 0^T \\ 0^T & h_a^T \end{bmatrix} \begin{bmatrix} x_1^v \cdots x_N^v \\ x_1^a \cdots x_N^a \end{bmatrix} \quad (3)$$

where $x_i^v \in \Re^{N_v}$ and $x_i^a \in \Re^{N_a}$ are lexicographic samples of images and periodograms, respectively, from an A/V sequence. The linear projection defined by $h_v^T \in \Re^{M_v \times N_v}$ and $h_a^T \in \Re^{M_a \times N_a}$ maps A/V samples to low dimensional features $y_i^v \in \Re^{M_v}$ and $y_i^a \in \Re^{M_a}$. Treating x_i's and y_i's as samples from a random variable our goal is to choose h_v and h_a to maximize the mutual information, $I(Y^a; Y^v)$), of the derived measurements.

Mutual information indicates the amount of information that one random variable conveys on average about another. The usual difficulty of MI as a criterion for adaptation is that it is an integral function of probability densities. Furthermore, in general we are not given the densities themselves, but samples from which they must be inferred. We use a second-order entropy approximation with a nonparametric density estimator such that the gradient terms with respect to the projection coefficients can be computed *exactly* by evaluating a finite number of functions at a finite number of sample locations in the output space as shown in [11, 12].

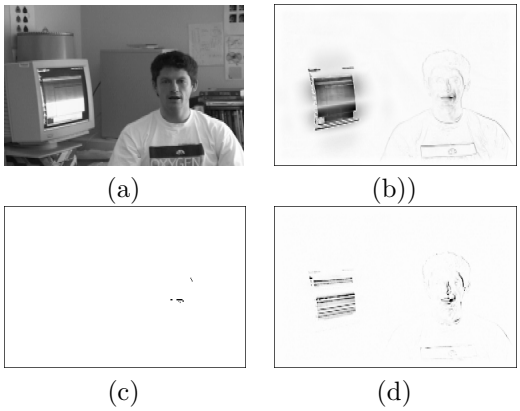


Figure 4: Video sequence contains one speaker and monitor which is flickering: (a) one image from the sequence, (b) pixel-wise image of standard deviations taken over the entire sequence, (c) image of the learned projection, h_v, (d) image of h_v for incorrect audio

This method requires that the projection be differentiable, which it is in our case. Additionally some form of capacity control is necessary as the method results in a system of underdetermined equations. To address this problem we impose an L_2 penalty on the projection coefficients of h_a and h_v [9]. Furthermore, we impose the criterion that if we consider the projection h_v as a filter, it has low output energy when convolved with images in the sequence (on average). This constraint is the same as that proposed by Mahalanobis *et al* [16] for designing optimized correlators the difference being that in their case the projection output was designed explicitly while in our case it is derived from the MI optimization in the output space. For the full details of this method see [10, 9].

5.2 Single microphone and camera experiments

Our motivating scenario for this application is a group of users interacting with an anonymous handheld device or kiosk using spoken commands. Given a received audio signal, we would like to verify whether the person speaking the command is

in the field of view of the camera on the device, and if so to localize which person is speaking.

Simple techniques which check only for the presence of a face (or moving face) would fail when two people were looking at their individual devices and one spoke a command. Since interaction may be anonymous, we presume no prior model of the voice or appearance of users are available to perform the verification and localization.

We collected audio-video data from eight subjects. In all cases the video data was collected at 29.97 frames per second at a resolution of 360x240. The audio signal was collected at 48000 KHz, but only 10Khz of frequency content was used. All subjects were asked to utter the phrase "How's the weather in Taipei?". This typically yielded 2-2.5 seconds of data. Video frames were processed as is, while the audio signal was transformed to a series of periodograms. The window length of the periodogram was 2/29.97 seconds (i.e. spanning the width of two video frames). Upon estimating projections the mutual information between the projected audio and video data samples is used as the measure of consistency. All values for mutual information are in terms of the maximum possible value, which is the value obtained (in the limit) if the two variables are uniformly distributed and perfectly predict one another. In all cases we assume that there is not significant head movement on the part of the speaker during the utterance of the sentence. While this assumption might be violated in practice one might account for head movement using a tracking algorithm, in which case the algorithm as described would process the images after tracking.

Figure 4a shows a single video frame from one sequence of data. In the figure there is a single speaker and a video monitor. Throughout the sequence the video monitor exhibits significant flicker. Figure 4b shows an image of the pixel-wise standard deviations of the image sequence. As can be seen, the energy associated with changes due to monitor flicker is greater than that due to the speaker. Figure 6a shows the associated periodogram sequence where the horizontal axis is time and the vertical axis is frequency (0-10 Khz). Figure 4c shows the coefficients of the learned projection when fused with the audio signal. As can be seen the projection highlights the region about the speaker's lips. Figure 5a shows results from another sequence in which there are two people. The person on the left was asked to utter the test phrase, while the person on the right moved their lips, but did not speak. This sequence is interesting in that a simple face detector would not

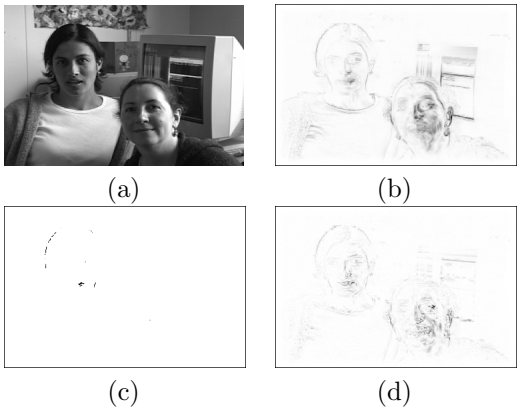


Figure 5: Video sequence containing one speaker (person on left) and one person who is randomly moving their mouth/head (but not speaking): (a) one image from the sequence, (b) pixel-wise image of standard deviations taken over the entire sequence, (c) image of the learned projection, h_v, (d) image of h_v for incorrect audio.

be sufficient to disambiguate the audio and video stream. Figure 5b shows the pixel variance as before. There are significant changes about both subjects lips. Figure 5c shows the coefficients of the learned projection when the video is fused with the audio and again the region about the correct speaker's lips is highlighted.

In addition to localizing the audio source in the image sequence we can also check for consistency between the audio and video. Such a test is useful in the case that the person to which a system is visually attending is not the person who actually spoke. Having learned a projection which optimizes MI in the output feature space, we can then estimate the resulting MI and use that estimate to quantify the audio/video consistency.

Using the sequences of figure 4 and 5 we compared the fusion result when using a separately recorded audio sequence from another speaker. The periodogram of the alternate audio sequence is shown in figure 6b. Figures 4d and 5d show the resulting h_v when the alternate audio sequence is used. In the case that the alternate audio was used we see that coefficients related to the video monitor increase significantly in 6d while energy is distributed throughout the image of 5d. For figure 4 the estimate of mutual information was 0.68 relative to the maximum possible value for the correct audio sequence. In contrast when compared to the periodogram of 6b, the value drops to 0.08 of maximum. For the sequence of figure 5, the estimate of mutual information for the correct sequence was 0.61 relative to maximum, while it drops to 0.27 when the alternate audio is used.

Data was collected from six additional subjects

(a)

(b)

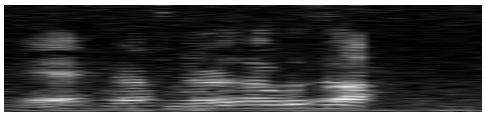


Figure 6: Gray scale magnitude of audio periodagrams. Frequency increases from bottom to top, while time is from left to right. (a) audio signal for image sequence of figure 4. (b) alternate audio signal recorded from different subject.

	a1	a2	a3	a4	a5	a6	a7	a8
v1	**0.68**	0.19	0.12	0.05	0.19	0.11	0.12	0.05
v2	0.20	**0.61**	0.10	0.11	0.05	0.05	0.18	0.32
v3	0.05	0.27	**0.55**	0.05	0.05	0.05	0.05	0.05
v4	0.12	0.24	0.32	**0.55**	0.22	0.05	0.05	0.10
v5	0.17	0.05	0.05	0.05	**0.55**	0.05	0.20	0.09
v6	0.20	0.05	0.05	0.13	0.14	**0.58**	0.05	0.07
v7	0.18	0.15	0.07	0.05	0.05	0.05	**0.64**	0.26
v8	0.13	0.05	0.10	0.05	0.31	0.16	0.12	**0.69**

Table 3: Summary of results over eight video sequences. The columns indicate which audio sequence was used while the rows indicate which video sequence was used. In all cases the correct audio/video pair have the highest relative MI score.

for this experiment, and each video sequence was compared to each audio sequence. (No attempt was made to temporally align the mismatched audio sequences at a fine scale, but they were coarsely aligned). Table 3 summarizes the results. The previous sequences correspond to subjects 1 and 2 in the table. In every case the matching audio/video pairs exhibited the highest mutual information after estimating the projections.

Finally, we present a new experiment demonstrating how this method can segregate speech of users in front of a kiosk. In concert with a face detection module, it is possible to detect which user is speaking and whether they are facing the camera. The audiovisual mutual information method is able to match the visual speech motion with the acoustic signal, and ignore confounding motions of the other user's head or other motions in the scene. Figure 7 shows the result tracking two users speaking in turns in front of a single camera and microphone, and detecting which is most likely to be speaking based on the measured audiovisual consistency.

6 Future Work

We have shown separately how geometric and statistical approaches can be used to solve audiovisual segmentation tasks and enable untethered conversational interaction. The geometric approach used 3-D tracking and array processing, and ignored appearance variation. The statistical approach used a mutual information analysis of appearance and spectral variation, and ignored 3-D geometry. While each approach is already valuable in the intended domain, it is clear that they are orthogonal and would benefit from combination. We are currently exploring an integrated approach that combines geometric and statistical insights in a common source separation algorithm. In addition, we are implementing a 3-D tracking algorithm which uses a symmetric approach to audio and video cues, rather than always using the video to initialize the audio search as reported above. We expect these results to be available for discussion at the June CLASS workshop meeting.

References

[1] S. Becker. *An Information-theoretic Unsupervised Learning Algorithm for Neural Networks.* PhD thesis, University of Toronto, 1992.

[2] U. Bub, M. Hunke, and A. Waibel. Knowing who to listen to in speech recognition: Visually guided beamforming. In *1995 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1995.

[3] M. Casey, W. Gardner, and S. Basu. Vision steered beam-forming and transaural rendering for the artificial life interactive video environment, (alive). In *99th Convention of the Audio Engineering Society*, 1995.

[4] M. Collobert, R. Feraud, G. LeTourneur, O. Bernier, J. E. Viallet, Y. Mahieux, and D. Collobert. Listen: a system for locating and tracking individual speakers. In *2nd International Conference on Face and Gesture Recognition*, 1996.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, 1991.

[6] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb. Plan-view trajectory estimation with dense stereo background models. In *2001 International Conference on Computer Vision*, 2001.

[7] T. Darrell, G. G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *IJCV*, (37(2)):199–207, June 2000.

[8] G. Deco and D. Obradovic. *An Information Theoretic Approach to Neural Computing.* Springer-Verlag, New York, 1996.

[9] J. W. Fisher III and T. Darrell. Probabalistic models and informative subspaces for audiovisual correspondence. In *to appear in Proceedings ECCV 2002.*, 2000.

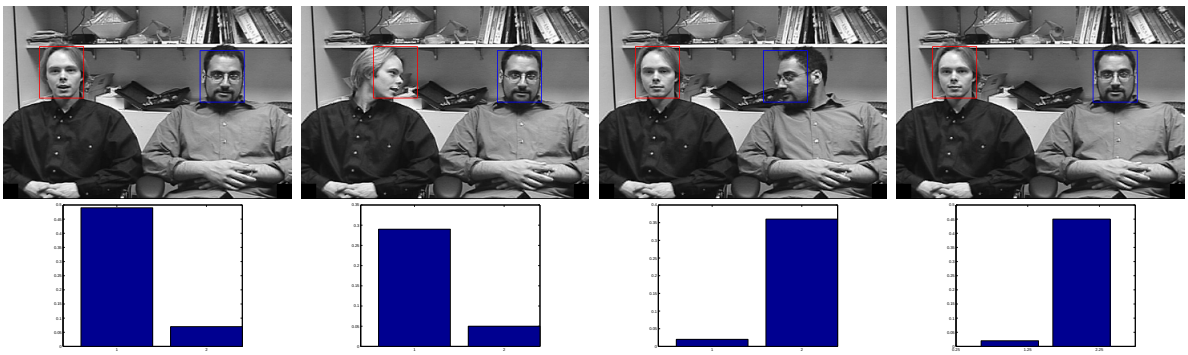


Figure 7: Top row presents four frames from a video sequence with two speakers in front of a single camera and microphone. Audiovisual consistency is measured using a mutual information criteria. In the first two frames the left person is speaking, while in the last two the right person is speaking. The consistency measure shown in the bottom row for each frame correctly detects who is speaking.

[10] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems 13*, 2000.

[11] J. W. Fisher III and J. C. Principe. Entropy manipulation of arbitrary nonlinear mappings. In J. Principe, editor, *Proc. IEEE Workshop, Neural Networks for Signal Processing VII*, pages 14–23, 1997.

[12] J. W. Fisher III and J. C. Principe. A methodology for information theoretic feature extraction. In A. Stuberud, editor, *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1998.

[13] J. Hershey and J. Movellan. Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, and K.-R. Mller, editors, *Advances in Neural Information Processing Systems 12*, pages 813–819, 1999.

[14] Y. A. Ivanov, A. F. Bobick, and J. Liu. Fast lighting independent background subtraction. *IJCV*, 2000.

[15] J. Krumm, S. Harris, B. Meyers, B. Brummit, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *3rd IEEE Workshop on Visual Surveillance*, 2000.

[16] A. Mahalanobis, B. Kumar, and D. Casasent. Minimum average correlation energy filters. *Applied Optics*, 26(17):3633–3640, 1987.

[17] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. Towards unrestricted lipreading. In *Second International Conference on Multimodal Interfaces (ICMI99)*, 1999.

[18] M. Plumbley. On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR. 78, Cambridge University Engineering Department, UK, 1991.

[19] M. Plumbley and S. Fallside. An information-theoretic approach to unsupervised connectionist models. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of the 1988 Connectionists Models Summer School*, pages 239–245. Morgan Kaufman, San Mateo, CA, 1988.

[20] H. F. Silverman, W. R. Patterson, and J. L. Flanagan. The huge microphone array. *IEEE Concurrency*, pages 36–46, Oct. 1998.

[21] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, 2000.

[22] B. D. V. Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, Apr. 1988.

[23] M. Viberg and H. Krim. Two decades of statistical array processing. In *31st Asilomar Conference on Signals, Systems, and Computers*, 1997.

[24] C. Wang and M. Brandstein. Multi-source face tracking with audio and visual data. In *IEEE International Workshop on Multimedia Signal Processing*, 1999.

[25] G. Wolff, K. V. Prasad, D. G. Stork, and M. Hennecke. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In *Proc. of Neural Information Proc. Sys. NIPS-6*, pages 1027–1034, 1994.

[26] V. Zue, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington. Jupiter: a telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96, 2000.

An Adaptive Dialogue System Using Multimodal Language Acquisition

Sorin DUSAN
CAIP, Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854, U.S.A.
sdusan@caip.rutgers.edu

James FLANAGAN
CAIP, Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854, U.S.A.
jlf@caip.rutgers.edu

Abstract

Our objective in developing natural human-computer interfaces is to allow for an adaptive dialogue based on learning new linguistic knowledge. This paper presents a dialogue system capable of adapting its language by learning new words, phrases, sentences, and their semantics from users. The acquisition of the new linguistic knowledge at syntactic and semantic levels is done using multiple modalities, including speaking, typing, pointing, touching or showing. The language knowledge is permanently stored in a rule grammar and a semantic database. Both can be updated periodically with newly acquired language.

Keywords: dialogue systems, adaptive dialogue, language acquisition, language understanding, multimodal interaction.

Introduction

A goal of current computer interfaces is to allow for a more natural, intelligent and easy human-computer interaction. One way of achieving this objective is by incorporating spoken dialogue into the interfaces. Increasingly, computers have integrated audio hardware for recording and playing sounds. This facilitates implementation of speech interfaces.

The speech modality complements and supplements the standard input-output methods based on keyboard, mouse and display. Other input devices, such as pen tablet and CCD camera, enhance further the interaction between users and system.

A spoken dialogue system requires the implementation of speech technologies that include automatic speech recognition (ASR), text-to-speech (TTS), speech understanding and dialogue management. Dialogue systems based on unconstrained vocabulary of course offer the most natural interaction, but they are more difficult to implement than those based on constrained vocabulary stored in a rule grammar.

A disadvantage of using dialogue systems based on rule grammars is that the developer cannot pre-program the rule grammar to account for all language preferences of users. Users find dialogue systems easier and more natural if they can change or adapt the allowed vocabulary and grammar according to their preferences.

To acquire language, computers need to learn linguistic knowledge at two levels: the *surface* level, represented by the syntax of these units; and the *deep* level, represented by the meaning of new linguistic units. In a series of studies of language acquisition based on connectionist methods, presented by Gorin (1995), new words or phrases are acquired at the *surface* level and their corresponding meanings are determined by probabilistic associations with pre-programmed semantic actions.

A study focusing on the acquisition of linguistic units and their primitive semantics from raw sensory data was published by Roy (1999). In that study the system learned new language by making associations between speech sounds representing words and their semantic representation acquired from a video camera.

Another study, published by Oates (2001), focused on discovering useful linguistic-semantic structures from raw sensory data. The goal was to enable a robot to discover associations between words and different semantic representations obtained from a video camera.

A method of acquiring new linguistic units and their semantics using multiple input modalities was introduced by Dusan and Flanagan (2001).

In this paper we present a computer dialogue system capable of adapting its vocabulary and grammar by learning new words, phrases, sentences and their semantic representation. The learning is accomplished from user input over a multimodal interface. The dialogue system is suitable for command and control applications in which the vocabulary of the dialogue can be expanded and personalised by users according to their preferences.

1 Adaptive Dialogue System

For a spoken dialogue interface, the adaptation means primarily being able to understand new words and sentences and this can be accomplished by learning new linguistic knowledge. The adaptation of the computer dialogue system presented in this paper is based on supervised language learning. We adopted a constrained-grammar dialogue method because this is suitable for command-and-control applications on a computer.

Our adaptive dialogue system is supported by a computer with a multimodal interface based on a microphone and speakers for speech, a keyboard for typing, a mouse for pointing, a pen tablet for drawing and touching, a CCD camera for image capturing and a display for graphics and text.

The adaptive dialogue system interprets users' utterances according to allowed sentence structures stored in the rule grammar, and executes different actions according to information stored in the semantic database. The adaptive dialogue system contains a speech recognition engine and a text-to-speech engine. The rule grammar and the semantic database are stored in two different files on a hard disk from which they are loaded into computer memory. When the adaptive system detects unknown words and the user provides the corresponding semantic representation, the system dynamically updates the rule grammar and the semantic database with the new linguistic knowledge. At the end of the application, the user has the option to save permanently the updated rule grammar and semantic database in corresponding files on the hard disk.

Adaptation of the dialogue to include new vocabulary and grammar takes place in real time during user-computer interaction. The learning method is that of supervised learning in which the user teaches the computer the semantic representation of new linguistic units. In addition to adapting the dialogue by learning new linguistic units and their semantics, the user can adapt the system's vocabulary by using synonyms or different names for already known terms.

2 Language Knowledge Representation

As mentioned, systems acquire language knowledge at two different levels: surface level and deep level. At the surface level, language is represented by knowledge of the vocabulary, syntax and grammar. We store this knowledge in a rule grammar. At the deep level, language is represented by knowledge of the meanings of linguistic units. We store this knowledge in a semantic database.

2.1 Rule Grammar

Grammar represents a specification of the allowed sentence structures in a language. A rule grammar defines the allowed sentence structures by a set of production rules. A context-free grammar (CFG) consists of a single start symbol and a set of rules.

We specify allowed sentence structures in a rule grammar, organised as a context-free grammar in a form of a semantic grammar, De Mori (1999). In this form the nonterminal symbols represent semantic classes of concepts, such as *colors*, *fruits* and *geometric shapes*, and the terminal symbols represent concept words such as *yellow*, *apple* and *rectangle*.

The rule grammar can be dynamically updated by adding new words or phrases in semantic classes or by adding new production rules. A linguistic unit integrated into a semantic class has a corresponding semantic object stored in the semantic database.

2.2 Semantic Database

Interpretation of utterances for performing necessary actions is based on semantic knowledge stored in the semantic database. This database contains a set of semantic objects that describe the meaning (or a semantic representation) for each concept stored in each class in the rule grammar. This semantic database can be dynamically updated with new

objects consisting of semantic representations of new linguistic units.

The semantic objects are created using the concept of object-oriented programming and they are instances of classes. The semantic representation stored in such an object defines the computer knowledge and interpretation of the corresponding linguistic unit. For example, the semantic object corresponding to the word *blue* included in the semantic class *colors*, has semantic representation defined by the RGB attributes (0, 0, 255). These attributes represent all computer knowledge regarding the meaning of this color. Another example is the semantic object for the word *square*. In this case the object contains a pointer to a regular polygon and an attribute equal to 4 representing the number of sides. All characteristics of a *regular polygon* are thus inherited by the *square* semantic object.

The semantic representation necessary to build these objects is either pre-programmed or acquired from the user through multiple input modalities.

3 Multimodal Language Acquisition

In our system acquisition of language consists of acquiring new vocabulary and corresponding semantics using multiple input modalities. A detailed block diagram of the system is given in Fig. 1.

The user communicates with the system by voice and the utterances are converted to text strings by an automatic speech recognition (ASR) engine and then analysed by a Language Understanding module, containing a Parser, a Command Processor, a Rule Grammar and a Semantic Database. Initially the ASR uses a language model derived from the Rule Grammar. The allowed utterances are converted to text at ASR output 1. Then they are parsed and executed by the Command Processor or forwarded to the Dialog Processor to provide appropriate answers through synthetic voice. The Command Processor also displays the results on the screen. The Dialogue Processor module includes a Text-To-Speech synthesizer and a Dialogue History necessary to solve ambiguities from the context of the dialogue. When the user's utterances contain unknown words or phrases, the ASR engine does not provide any text at its output 1. In this case the

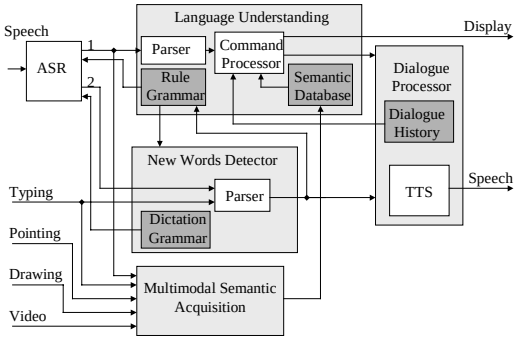


Figure 1: Multimodal Language Acquisition

ASR switches the language model to one derived from the Dictation Grammar. Then these utterances are converted to text strings at ASR output 2 and are applied to the New Words Detector. This module contains a Dictation Grammar and a Parser. The Dictation Grammar contains a very large vocabulary of words and allows the ASR to recognise more unconstrained utterances. The Parser analyses these utterances according to all allowed words stored in the Rule Grammar and detects in these utterances unknown words or phrases. Upon the detection of new linguistic units, this module issues a signal to the Dialogue Processor that asks the user by synthetic voice to provide a semantic representation of the new words or phrases.

The user can provide a semantic representation using multiple modalities. For each new linguistic unit the semantic representation is captured by the Multimodal Semantic Acquisition module that creates a new semantic object. After the user has provided semantics for the new words or phrases, the new linguistic units are stored in the rule grammar and the corresponding semantic objects are stored in the semantic database.

Another means to acquire language is by teaching the computer by typing a whole new sentence and the corresponding computer action. The new sentence is stored in the rule grammar and the computer action must be based on a combination of known actions. For example, one can type in the new sentence 'Double the radius' and the corresponding computer action '{radius} {multiplication} {2}', where the word 'double'

is unknown, but the words ‘the’, ‘radius’, ‘multiplication’ and ‘2’ are known.

4 Experiments

To demonstrate and evaluate the method of adaptive dialogue based on multimodal language acquisition we developed a simple application on a personal computer system with multimodal input-output devices consisting of microphone and speakers, keyboard, mouse, pen tablet and stylus, a CCD camera and a graphic display. The whole application was written in JAVA. The rule grammar was specified using Java Speech Grammar Format (JSGF). We used the automatic speech recognition and speech synthesis engines from the IBM ViaVoice commercial package.

The system permits a dialogue between user and computer to create, move, rotate and delete graphic objects on the screen. The initial allowed dialogue was stored in a rule grammar in a set of 25 production rules and 23 nonterminal symbols representing classes of concepts such as, *colors*, *actions*, *names*, *display variables*, etc.

The system can easily learn synonyms, new user names, new colors by pointing the mouse on a color palette or new graphic objects by drawing with a stylus on a pen tablet. The new language knowledge can be built upon already known knowledge. Fig. 2 shows a graphic screen for a session in which the user taught the computer the graphic representations for the terms: *hair*, *face*, *left eye*, *right eye*, *nose*, *left ear*, *right ear*, *mouth* and *head*. Each new word or phrase was spoken and the computer asked for semantic representations which were provided by the user by drawing the corresponding graphics. After all primitive terms were taught, the user created a composition of these graphical elements representing a head and then taught the computer the word *head*. The second head in the picture was then created by the user by just saying ‘Create a head’ and simultaneously pointing the cursor with the mouse to the desired location. The nine concepts were taught in about 8 minutes, but most of this time was spent in drawing.

Figure 2: Language acquisition example screen

Conclusion

We present a computer dialogue system capable of adaptation and personalization by teaching the computer new linguistic units and their semantics using multiple input modalities. The system can expand its vocabulary by adding new words or phrases in the rule grammar and by creating and storing into a semantic database new objects containing the corresponding semantic representation. An alternative method of dialogue adaptation is by typing new sentences and their executable actions. The dialogue system can also be personalised by users by providing synonyms or different proper names to already-known terms.

Acknowledgements

This research was supported by the National Science Foundation under the KDI project, grant NSF IIS-98-72995.

References

Allen Gorin (1995) On automated language acquisition, Journal of the Acoustical Society of America, 97 (6), pp. 3441-3461

Deb K. Roy (1999) Learning Words from Sights and Sounds: A Computational Model, Ph.D. Thesis, MIT

Tim Oates (2001) Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning, Ph.D. Thesis, MIT

Sorin Dusan and James L. Flanagan (2001) Human Language Acquisition by Computers, in Proceedings of the International Conference on Robotics, Distance Learning and Intelligent Communication Systems, WSES/IEEE, Malta, pp. 387-392

Renato De Mori (1999) Recognizing and Using Knowledge Structures in Dialog Systems, In Proceedings of IEEE-ASRU99, Keystone CO, pp. 297-307

Effective interaction with talking animated agents in dialogue systems

Björn GRANSTRÖM
CTT, KTH
SE 10044 Stockholm, Sweden
bjorn@speech.kth.se

David HOUSE
CTT, KTH
SE 10044 Stockholm, Sweden
davidh@speech.kth.se

Abstract

This paper describes activities at CTT using the potential of animated talking agents to increase the efficiency of communication. Our motivation for moving into audiovisual output is to investigate the advantages of multimodality in human-system communication. While the mainstream character animation area has focussed on the naturalness and realism of the animated agents, our primary concern has been the possible increase of intelligibility and efficiency of interaction, resulting from the addition of a talking face.

Introduction

Spoken dialogue systems, which strive to take advantage of the effective communication potential of human conversation, need in some way to embody the conversational partner. A talking animated agent provides the user with an interactive partner whose goal is to take the role of the optimal human agent. This is the agent who is ready and eager to supply the user with a wealth of information, can smoothly navigate through varying and complex sources of data and can ultimately assist the user in a decision making process through the give and take of conversation. One way to achieve believability is through the use of a talking head which transforms information through text into speech, articulator movements, speech related gestures and conversational gestures.

The talking head developed at KTH is based on text-to-speech synthesis. Audio speech synthesis is generated from a text representation in synchrony with visual articulator movements of the lips, tongue and jaw. Linguistic information in the text is used to generate visual cues for relevant prosodic categories such as prominence, phrasing and emphasis. These cues generally take the form of eyebrow and head movements which we have termed “visual prosody”. These types of visual cues with the addition of a smiling or frowning face are also used as conversational gestures to signal such things as positive or negative feedback, turntaking regulation, and the system’s internal state. In addition, the head can visually signal attitudes and emotions.

In the context of this paper, the talking head is primarily discussed in terms of applications in spoken dialogue systems which enable the user to access information and reach a decision through a conversational interface. Other useful applications include aids for the hearing impaired, educational software, stimuli for audiovisual human perception experiments, entertainment, and high-quality audio-visual text-to-speech synthesis for applications such as news reading. In this paper we will focus on two aspects of effective interaction: presentation of information and the flow of interactive dialogue.

Effectiveness in the presentation of information is crucial to the success of an interactive system. Information must be presented rapidly, succinctly and with high intelligibility. The use of the talking head aims at improving the intelligibility of speech synthesis through visual articulation and by providing the system with a visible location of the speech source to maintain the attention of the user. Important information is highlighted by prosodic enhancement and by the use of the agent’s gaze and visual prosody to create and maintain a common focus of attention.

The second issue of effective interaction focusses on facilitating the interactive nature of dialogue. In this area, the use of the talking head aims at increasing effectiveness by building on the user’s social skills to improve the flow of the dialogue and engage the user interactively.

Visual cues to feedback, turntaking and signalling the system's internal state (the thinking metaphor) are key aspects of effective interaction.

This paper presents a brief overview and technical description of the KTH talking head explaining what the head can do and how. Examples of experimental applications in which the head is used are then described, and finally, the two issues of intelligibility and communication interaction are discussed and exemplified by results from applications and perceptual evaluation experiments.

1 Technical description of the talking head

Animated synthetic talking faces and characters have been developed using a number of different techniques and for a variety of purposes during the past two decades. Our approach is based on parameterised, deformable 3D facial models, controlled by rules within a text-to-speech framework (Carlson & Granström, 1997) The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account (Beskow, 1995). We employ a generalised parameterisation technique to adapt a static 3D-wireframe of a face for visual speech animation (Beskow, 1997). Based on concepts first introduced by Parke (1982), we define a set of parameters that will deform the wireframe by applying weighted transformations to its vertices. One critical difference from Parke's system, however, is that we have de-coupled the model definitions from the animation engine, thereby greatly increasing flexibility.

The models are made up of polygon surfaces that are rendered in 3D using standard computer graphics techniques. The surfaces can be articulated and deformed under the control of a number of parameters. The parameters are designed to allow for intuitive interactive or rule-based control. For the purposes of animation, parameters can be roughly divided into two (overlapping) categories: those controlling speech articulation and those used for non-articulatory cues and emotions. The articulatory parameters include jaw rotation, lip rounding, bilabial occlusion, labiodental occlusion and tongue tip elevation. The non-articulatory category includes eyebrow raising, eyebrow shape, smile, gaze direction and head orientation. Furthermore, some of the articulatory parameters such as jaw rotation can be useful in signalling non-verbal elements such as certain emotions. The display can be chosen to show only the surfaces or the polygons for the different components of the face. The surfaces can be made (semi)transparent to display the internal parts of the model. The model presently contains a relatively crude tongue model primarily intended to provide realism as seen from the outside, through the mouth opening. A full 3D model of the internal speech organs is presently being developed for integration in the talking head (Engwall, 2001). This capability of the model is especially useful in explaining non-visible articulations in the language learning situation (Cole et al., 1999) In Figure 1 some of the display options are illustrated.

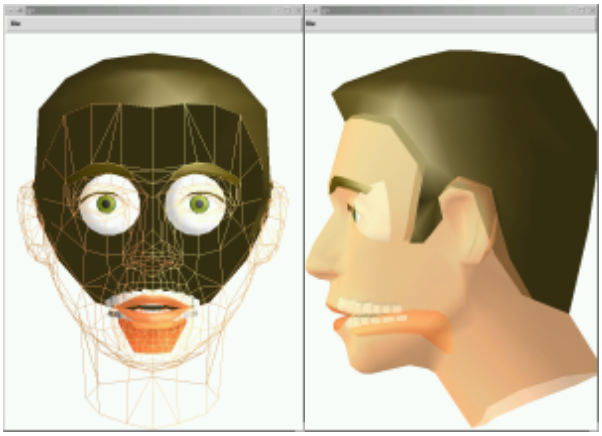


Figure 1. Different display possibilities for the talking head model. Different parts of the model can be displayed as polygons or smooth (semi)transparent surfaces to emphasise different parts of the model.

For stimuli preparation and explorative investigations, we have developed a control interface that allows fine-grained control over the trajectories for acoustic as well as visual parameters. The interface is implemented as an extension to the WaveSurfer application (www.speech.kth.se/wavesurfer + Sjölander & Beskow, 2000) which is a tool for recording, playing, editing, viewing, printing, and labelling audio data.

The interface makes it possible to start with an utterance synthesised from text, with the articulatory parameters generated by rule, and

then interactively edit the parameter tracks for F0, visual (non-articulatory) parameters as well as the durations of individual segments in the utterance to produce specific cues. An example of the user interface is shown in Figure 2. In the top box a text can be entered in Swedish or English. This creates a phonetic transcription that can then be edited. On pushing "Synthesize", rule generated parameters will be created and displayed in different panes below. The selection of parameters is user controlled. The lower section contains segmentation and the acoustic waveform. A talking face is displayed in a separate window. The acoustic synthesis can be exchanged for a natural utterance and synchronised to the face synthesis. This is useful for different experiments on multimodal integration and has been used in the Synface/Teleface project (see below). In language learning applications it could be used to add to the naturalness of the tutor's voice in cases when the acoustic synthesis is judged to be inappropriate.

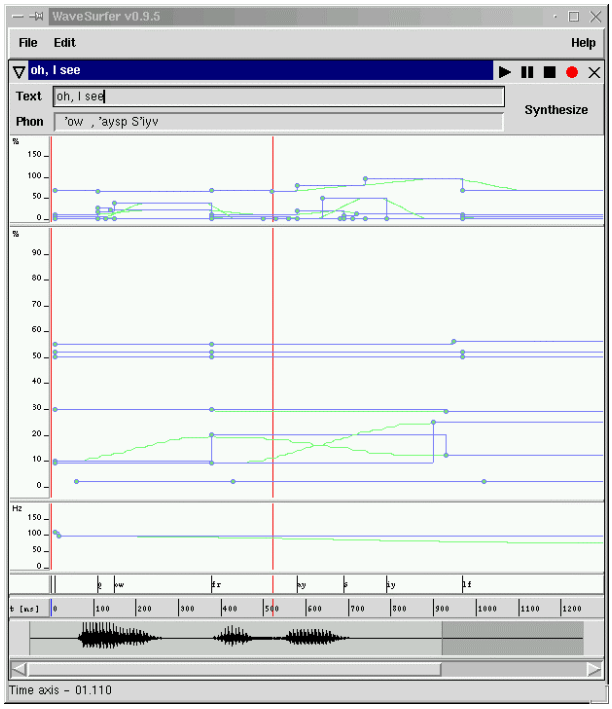


Figure 2. The Speech Surfer user interface for parametric manipulation of the multimodal synthesis.

The parametric manipulation tool is used to experiment with and define different gestures. A gesture library is under construction, containing procedures with general emotion settings and non-speech specific gestures as well as some procedures with linguistic cues. We are at present developing an XML-based representation of visual cues that facilitates description of the visual cues at a higher level.

2 Experimental applications

During the past decade a number of experimental applications using the talking head have been developed at KTH. Four examples which will be discussed here are the Waxholm demonstator system designed to provide tourist information on the Stockholm archipelago, the Synface project which is a visual hearing aid, the August project which was a dialogue system in public use, and the Adapt multimodal real-estate agent.

2.1 The waxholm demonstator

The first KTH demonstrator application, which we named WAXHOLM, gives information on boat traffic in the Stockholm archipelago. It references timetables for a fleet of some twenty boats from the Waxholm company connecting about two hundred ports (Bertenstam et al., 1995)

Besides the dialogue management and the speech recognition and synthesis components, the system contains modules that handle graphic information such as pictures, maps, charts, and timetables. This information can be presented as a result of the user-initiated dialogue.

The Waxholm system can be viewed as a micro-world, consisting of harbours with different facilities and with boats that you can take between them. The user gets graphic feedback in the form of tables complemented by speech synthesis. In the initial experiments, users were given a scenario with different numbers of subtasks to solve. A problem with this approach is that the users tend to use the same vocabulary as the text in the given scenario. We also observed that the user often did not get enough feedback to be able to decide if the system had the same interpretation of the dialogue as the user.

To deal with these problems a graphical representation that visualises the Waxholm micro-world was implemented. An example is shown in Figure 3. One purpose of this was to give the subject an idea of what can be done

with the system, without expressing it in words. The interface continuously feeds back the information that the system has obtained from the parsing of the subject's utterance, such as time, departure port and so on. The interface is also meant to give a graphical view of the knowledge the subject has secured thus far, in the form of listings of hotels and so on.

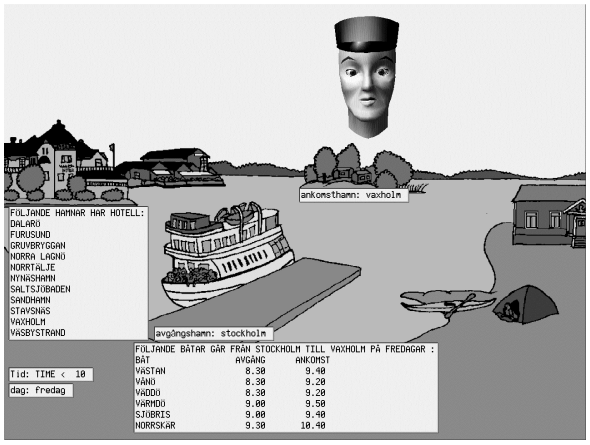


Figure 3. The graphical model of the WAXHOLM micro-world.

The visual animated talking agent is an integral part of the system. This is expected to raise the intelligibility of the system's responses and questions. Furthermore, the addition of the face into the dialogue system has many other exciting implications. Facial non-verbal signals can be used to support turntaking in the dialogue, and to direct the user's attention in certain ways, e.g. by letting the head turn towards time tables, charts, etc. that appear on the screen during the dialogue. The dialogue system also provides an ideal framework for experiments with non-verbal communication and facial actions at the prosodic level, as discussed above, since the system has a much better knowledge of the discourse context than is the case in plain text-to-speech synthesis.

To make the face more alive, one does not necessarily have to synthesise meaningful non-verbal facial actions. By introducing semi-random eyeblinks and very faint eye and head movements, the face looks much more active, and becomes more pleasant to watch. This is especially important when the face is not talking.

2.2 The Synface/Teleface project

The speech intelligibility of talking animated agents, as the ones described above, has been tested within the Teleface project at KTH (Beskow et al., 1997; Agelfors et al., 1998). The project has recently been continued/expanded in a European project, Synface (Granström, Karlsson & Spens, 2002). The project focuses on the usage of multi-modal speech technology for hearing-impaired persons. The aim of the first phase of the project was to evaluate the increased intelligibility hearing-impaired persons experience from an auditory signal when it is complemented by a synthesised face. In this case, techniques for combining natural speech with lip-synchronised face synthesis have been developed. A demonstrator of a system for telephony with a synthetic face that articulates in synchrony with a natural voice is currently being implemented (see Figure 4).

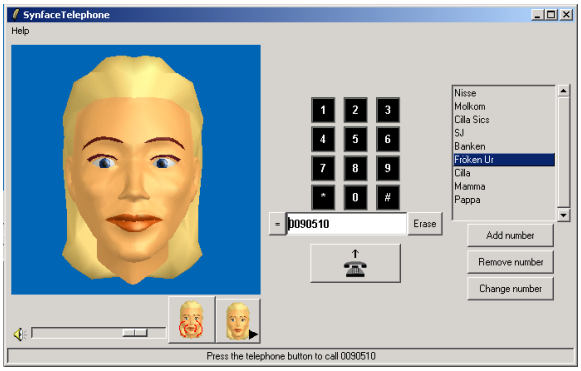


Figure 4. Telephone interface for SYNFACE.

2.3 The August system

The Swedish author, August Strindberg, provided inspiration to create the animated talking agent used in a dialogue system that was on display during 1998 as part of the activities celebrating Stockholm as the Cultural Capital of Europe (Gustafson et al., 1999). This dialogue system made it possible to combine several domains, thanks to the modular functionality of the architecture. Each domain has its own dialogue manager, and an example based topic spotter is used to relay the user utterances to the appropriate dialog manager. In this system, the animated agent "August" presents different tasks such as taking the visitors on a trip through the Department of Speech, Music and Hearing, and giving street directions and also presenting short

excerpts from the works of August Strindberg, when waiting for someone to talk to.

August was placed, unattended in a public area of Kulturhuset in the centre of Stockholm. One challenge is this very open situation with no explicit instructions being given to the visitor. A simple visual "visitor detector" makes August start talking about one of his knowledge domains.

2.4 The Adapt multimodal realestate agent

The practical goal of the AdApt project is to build a system in which a user can collaborate with an animated agent to solve complicated tasks (Gustafson et al., 2000). We have chosen a domain in which multimodal interaction is highly useful, and which is known to engage a wide variety of people in our surroundings, namely, finding available apartments in Stockholm. In the AdApt project, the agent has been given the role of asking questions and providing guidance by retrieving detailed authentic information about apartments. The user interface can be seen in Figure 5

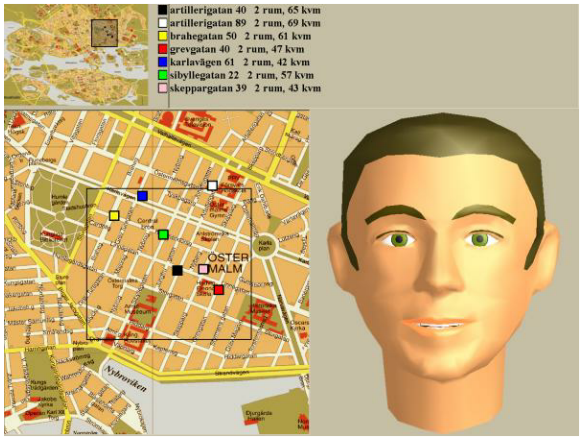


Figure 5 . The agent Urban in the AdApt apartment domain.

Because of the conversational nature of the AdApt domain, the demand is great for appropriate interactive signals (both verbal and visual) for encouragement, affirmation, confirmation and turntaking (Cassell et al., 2000; Pelachaud, Badler & Steedman, 1996). As generation of prosodically grammatical utterances (e.g. correct focus assignment with regard to the information structure and dialogue state) is also one of the goals of the system it is important to maintain modality consistency by simultaneous use of both visual and verbal prosodic and conversational cues (Nass & Gong, 1999). As described in Section 1, we are at present developing an XML-based representation of such cues that facilitates description of both verbal and visual cues at the level of speech generation. These cues can be of varying range covering attitudinal settings appropriate for an entire sentence or conversational turn or be of a shorter nature like a qualifying comment to something just said. Cues relating to turntaking or feedback need not be associated with speech acts but can occur during breaks in the conversation. Also in this case, it is important that there is a one-to-many relation between the symbols and the actual gesture implementation to avoid stereotypic agent behaviour. Currently a weighted random selection between different realizations is used.

3 Effectiveness in intelligibility and information presentation

One of the more striking examples of improvement and effectiveness in speech intelligibility is taken from the Synface project which aims at improving telephone communication for the hearing impaired (Agelfors et al., 1998). The results of a series of tests using VCV words and hearing impaired subjects showed a significant gain in intelligibility when the talking head was added to a natural voice. With the synthetic face, consonant identification improved from 29% to 54% correct responses. This compares to the 57% correct response result obtained by using the natural face. In certain cases, notably the consonants consisting of lip movement (i.e. the bilabial and labiodental consonants), the response results were in fact better for the synthetic face than for the natural face. This points to the possibility of using overarticulation strategies for the talking face in these kinds of applications. Recent results indicate that a certain degree of overarticulation can be advantageous in improving intelligibility (Beskow, Granström & Spens, 2002)

Similar intelligibility tests have been run using normal hearing subjects where the audio signal was degraded by adding white noise (Agelfors et al., 1998). Similar results were obtained. For example, for a synthetic male voice, consonant identification improved from 31% without the

face to 45% with the face. While the visual articulation is most probably the key factor contributing to this increase, we can speculate that the presence of visual information of the speech source can also contribute to increased intelligibility by sharpening the focus of attention of the subjects. Although this hypothesis has not been formally tested, it could be useful to test it generally in many different applications.

Another quite different example of the contribution of the talking head to information presentation is taken from the results of perception studies in which the percept of emphasis and syllable prominence is enhanced by eyebrow and head movements. In an early study restricted to eyebrows and prominence (Granström et al., 1999) it was shown that raising the eyebrows alone during a particular syllable resulted in an increase in prominence judgments for the word in question by nearly 30%. In a later study, it was shown that eyebrows and head movement can serve as independent visual cues for prominence, and that synchronization of the visual movement with the audio speech syllable is an important factor (House et al., 2001). Head movement was shown to be somewhat more salient for signalling prominence as eyebrow movement could be potentially misinterpreted as supplying non-linguistic information such as irony.

A third example of information enhancement by the visual modality is to be found in the Waxholm demonstrator and the Adapt system. In both these systems, the agent uses gaze to point to areas and objects on the screen, thereby strengthening the common focus of attention between the agent and the user. Although this type of information enhancement has not yet been formally evaluated in the context of these systems, it must be seen as an important potential for improving the effectiveness of interaction.

Finally, an important example of the addition of information through the visual modality is to be found in the August system. This involved adding mood, emotion and attitude to the agent. To enable display of the agent's different moods, six basic emotions similar to the six universal emotions defined by Ekman (1979) were implemented (Figure 6), in a way similar to that described by Pelachaud, Badler & Steedman (1996). Appropriate emotional cues were assigned to a number of utterances in the system, often paired with other gestures.

4 Effectiveness in interaction

The use of a believable talking head can trigger the user's social skills such as using greetings, addressing the agent by name, and generally socially chatting with the agent. This was clearly shown by the results of the public use of the August system during a period of six months (Bell & Gustafson, 1999)). These promising results have led to more specific studies on visual cues for feedback (Granström et al., 2002) in which smile, for example, was found to be the strongest cue for affirmative feedback. Further detailed work on turntaking regulation, feedback seeking and giving and the signalling of the system's internal state will enable us to improve the gesture library available for the animated talking head and continue to improve the effectiveness of multimodal dialogue systems.

Figure 6. August showing different emotions (from top left to bottom right): Happiness, Anger, Fear, Surprise, Disgust and Sadness

Acknowledgement

The research reported here was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

References

Agelfors, E., J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K.-E. Spens & T. Öhman "Synthetic faces as a lipreading support". In: Proceedings of ICSLP'98, Sydney, Australia, 1998.

Bell L & Gustafson J.(1999). Utterance types in the August System . Proc from IDS '99

Bertenstam J., J. Beskow, M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, J. Högberg, R. Lindell, L. Neovius, A. de Serpa-Leitao, L. Nord & N. Ström.(1995) "The Waxholm system - a progress report", In: Proceedings of Spoken Dialogue Systems, Vigsø, Denmark, 1995.

Beskow J (1995). Rule-based Visual Speech Synthesis, In Proceedings of Eurospeech '95, 299-302, Madrid, Spain.

Beskow J., Dahlquist M., Granström B., Lundeberg M., Spens K.-E. & Öhman T. "The teleface project - multimodal speech communication for the hearing impaired". In: Proceedings of Eurospeech '97, Rhodos, Greece, 1997.

Beskow, Granström, & Spens (2002) Articulation strength - Readability experiments with a synthetic talking face TMH-QPSR vol. 44, Stockholm: KTH, 97-100.

Beskow, J. "Animation of Talking Agents", In: Proceedings of AVSP'97, ESCA Workshop on Audio-Visual Speech Processing, Rhodes, Greece, 1997.

Carlson R and Granström B (1997). Speech Synthesis, In Hardcastle W and Laver J (eds) The Handbook of Phonetic Sciences, 768-788, Oxford: Blackwell Publishers Ltd.

Cassell J., T. Bickmore, L. Campbell, V. Hannes & H. Yan. "Conversation as a System Framework: Designing Embodied Conversational Agents". In: Cassell J., J. Sullivan, S. Prevost & E. Churchill (Eds). Embodied Conversational Agents, Cambridge MA: The MIT Press, 2000.

Cole R, Massaro D W, de Villiers J, Rundle B, Shobaki K, Wouters J, Cohen M, Beskow J, Stone P, Connors P, Tarachow A and Solcher D (1999). New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. In Proceedings of ESCA/Socrates Workshop on Method and Tool Innovations for Speech Science Education (MATISSE), 45-52, London: University College London.

Ekman P. "About brows: Emotional and conversational signals". In: von Cranach M., K Foppa, W. Lepinies & D. Ploog. (Eds). Human ethology: Claims and limits of a new discipline: Contributions to the Colloquium, 169-248, Cambridge: Cambridge University Press, 1979.

Engwall O. "Making the Tongue Model Talk: Merging MRI & EMA Measurements", In: Proc of Eurospeech 2001, 261-264, 2001.

Granström B, House D & Lundeberg M (1999). Prosodic cues in multi-modal speech perception. Proc of ICPhS-99, 655-658

Granström, B., House, D. and Swerts, M.G. 'Multimodal feedback cues in human-machine interactions.' In Bernard Bel & Isabelle Marlien (eds.), Proceedings of the Speech Prosody 2002 Conference, 11-13 April 2002. 347-350.Aix-en-Provence: Laboratoire Parole et Langage.

Granström, Karlsson & Spens (2002) SYNFACE – a project presentation TMH-QPSR vol. 44, Stockholm: KTH, 93-96.

Gustafson J, Lindberg N & Lundeberg M (1999). The August spoken dialogue system. Proc of Eurospeech'99, 1151-1154

Gustafson, J., L. Bell, J. Beskow, J. Boye, R. Carlson, J. Edlund, B. Granström, D. House & M. Wirén "AdApt - a multimodal conversational dialogue system in an apartment domain". In: Proceedings of ICSLP 2000, (2) 134-137. Beijing, China, 2000.

House, D., J. Beskow and B. Granström. 'Timing and interaction of visual cues for prominence in audiovisual speech perception.' In Proceeding of Eurospeech 2001, 387-390. Aalborg, Denmark.

Nass C. & L. Gong. "Maximized modality or constrained consistency?" In: Proceedings of AVSP'99, 1-5, Santa Cruz, USA, 1999.

Parke F I (1982). Parameterized models for facial animation, IEEE Computer Graphics, 2(9), 61-68.

Pelachaud C., N.I. Badler & M. Steedman. "Generating Facial Expressions for Speech". Cognitive Science 28, 1-46, 1996.

Sjölander, K. and Beskow, J. (2000). "WaveSurfer - an Open Source Speech Tool", in Proceedings of ICSLP 2000, Bejing, China, October 2000

Analysing Multi-Modal Communication: Repair-Based Measures of Communicative Co-ordination

Patrick G.T. Healey and Mike Thirlwell
Information, Media, and Communication Research Group,
Department of Computer Science,
Queen Mary University of London
{ph,miket}@dcs.qmul.ac.uk

Abstract

Few techniques are available for analysing the effectiveness of multimodal interaction. Psycholinguistic models have the potential to fill this gap however existing approaches have some methodological and practical limitations. This paper proposes a technique based on the conversation analytic model of breakdown and repair. The rationale for the approach is presented and a protocol for coding repair is introduced. The potential application of this approach to the analysis of mutli-modal interaction is described.

Keywords: Computer-Mediated Communication, Evaluation, Conversation Analysis.

1 Introduction

Communication has a central role in many human activities, even for tasks and technologies that are ostensibly individual (Suchman, 1987; Heath, 2000; Nardi and Miller, 1991; Nickerson and Adams, 1995). Communication is central to the implicit and explicit co-ordination of routine activities and to peoples' responses to unexpected contingencies. Significant commerical effort is directed toward the development of applications that are explicitly designed to support communication. For example, desktop 'messenger' and chat applications and more complex integrated conferencing tools. It is also recognised that even for applications not directly designed to support interaction the extent to which they impede or facilitate communication is often critical to their success (Heath and Luff, 1992; Hughes, Randall and Shapiro, 1992; Bowers, Button and Sharrock, 1995). Conceptual and empirical analysis of human communication should have much to contribute to the design and use of these technologies. However, there are few techniques either for identifying the communication related requirements of a given activity or for evaluating the impact different technologies have on the effectiveness of communication.

Ethnomethodological studies of workplace interaction have provided detailed descriptions of some of the practices by which individuals adjust their patterns of communication to sustain collaborative activity (Heath and Luff, 1992; Hughes, Randall and Shapiro, 1992; Bowers, Button and Sharrock, 1995). For example, one recurrent observation is the use of 'outlouds', a class of utterances that have a broadcast character which helps to manipulate the visibility of activities to members of a team without being addressed to any specific individual or requiring any specific acknowledgment. Despite the elegance of the empirical observations, workplace studies have a problematic relationship to design (Hughes, Randall and Shapiro, 1992; Button and Dourish, 1996). The analyses they provide are retrospective in character and do not support systematic comparisons across different technologies or situations.

The models and techniques developed in the psycholinguistic tradition explicitly aim to quantify communicative phenomena and develop explanations that support predictive

generalisations. The simplest application of this approach has been to use structural measures of interaction such as turn-taking, interruptions, backchannels, and gaze (O'Conaill, Whittaker and Wilbur, 1993; OMalley, Langton, Anderson, Doherty-Sneddon, and Bruce, 1996). Although these measures support direct comparison between some aspects of, for example, video-mediated interaction and face-to-face exchanges, the coding categories are coarse and do not take account of communicative function. For example, the category of 'interruptions' is formulated as the occurrence of overlapping speech where there has been no signal that a speaker is relinquishing the floor. This conflates accidental overlap, disruptive or competitive interventions and cooperative interventions such as collaborative completions. As a result the distribution of e.g., numbers of turns, lengths of turn, and interruptions are consistent with a number of possible interpretations (Anderson et al, 1997). Observations such as the finding that video-mediated communication leads participants to use more turns and words than they would when face-to-face are thus difficult to interpret. It might indicate either that video adds something to the interaction or that participants are working to compensate for its limitations. Two psycholinguistic approaches that have improved on structural analyses are Dialogue Games analysis and the Collaborative Model of Dialogue.

1.1 Dialogue Games Analysis

Problems with the interpretation of simple structural analyses have been addressed to some extent by techniques which analyse communicative function directly. For example, Anderson et al (1997) compare the performance of subjects carrying out a map drawing task under different media conditions. This task has a reliable dialogue coding system developed for it which characterises the functional structure of the dialogues, analysing each utterance as a move (e.g., instruct, explain, check, clarify, query-w, align) and structured sequences of moves as dialogue games (Kowtko, Isard and Doherty, 1991). (Anderson et al, 1997) have shown how this approach improves on structural analyses by analysing the profile of moves used to carry out the task either face-to-face, using audio only or using video-mediated communication. Unlike the structural analyses this provides a means of isolating the communicative functions that are preserved or impeded under different conditions of mediated interaction.

Although it supports systematic comparisons between media, this kind of functional analysis also has limitations. The coding system was designed to exhaustively classify utterance function in a particular collaborative task, the map task. As a result the move types are tailored to the transactional character of the task and although it does appear to generalise well to some other information exchange tasks it is unclear whether the coding scheme is adequate for qualitatively different kinds of exchange such as competitive negotiation. A practical limitation is that because the coding system is exhaustive, every utterance is analysed and coded. This is labour intensive and has the consequence that the sensitivity of the coding system is traded-off against its coverage (Carletta, Isard, Isard, Kowtko, Doherty-Sneddon and Anderson, 1996). An additional concern is that the functional analysis doesn't discriminate between incidental exchanges and those necessary for the activity at hand. The coding system includes all exchanges regardless of whether they are related to the task, the weather, or incidental gossip and anecdotes.

1.2 The Collaborative Model of Dialogue

A more readily generalised approach to modeling communication that also improves on simple structural analyses is the collaborative model of dialogue (CM) developed by Clark and co-workers (Clark, 1996; Clark and Wilkes-Gibbs, 1986; Clark, 1989). This model is based on an account of the process through which people build up their common ground during interaction. The basic principle is that the parties to an interaction only consider an utterance, or other communica-

tive act, to be successful where some positive evidence of its acceptance or 'grounding' has been obtained. The grounding principle is modified according to the types of evidence necessary to secure understanding in a given case and the degree of effort required to achieve it. Two of the most important qualifications are that; a) interlocutors always seek to reduce the joint, as opposed to individual, effort necessary to successfully ground a communicative act and b) interlocutors will attempt to ground a contribution only up to a criterion level (the grounding criterion) which is adjusted according to circumstances.

This apparatus has been applied to characterising the properties of different media. Clark and Brennan (1991) identify a set of 8 constraints on grounding that derive from the signal characteristics of different communicative media (Copresence, Visibility, Audibility, Cotemporality, Simultaneity, Sequentiality, Reviewability and Revisability). These constraints alter the ease with which particular types of feedback can be provided and, consequently, the ease with which grounding can be achieved. The constraints are linked to the process of grounding by reference to the costs they exact on grounding techniques (Formulation costs, Production costs, Reception costs, Understanding Costs, Start-up costs, Delay costs, Asynchrony costs, Speaker Change costs, Fault costs, Repair costs). During interaction, individuals must make a trade-off between the different types of action they can undertake in order to ground a particular contribution and their relative costs in a particular medium. For example, where turn taking costs are high individuals may invest more in the construction of each utterance and less in attempts at concurrent feedback.

The CM has also been applied to the analysis of system feedback. The CM distinguishes a number of levels at which an action is considered complete. For example, an utterance may be perceived but not understood, or it may be understood but the action it proposes is not undertaken. These action levels are ordered according to the principle that feedback which indicates completion at a higher level presupposes completion at a lower level: if I comply with your request then this is evidence that I have also heard and understood it. Consequently, the current state of the common ground with respect to communicative action can vary depending on the degree of grounding that has been secured. The maintenance of context in interaction with a system can be supported by giving feedback which signals the level of grounding that a particular action has achieved (Brennan, 1988) . The level of feedback given can be modulated by the risks associated with possible misunderstandings.

There are some important limitations to these applications of the CM model. One problem is that it provides only a limited analysis of situations in which contributions fail to secure acceptance. Where this occurs there is a general expectation that some repair will ensue, for example through reformulation of the contribution in a way that is acceptable to the addressee(s). In this situation a number of possible types of reformulation e.g., alternative descriptions, installment descriptions and trial references, are distinguished but the pattern of choice amongst these types, and their relationship to the success of repairs, is not addressed in the CM. Although it is explicitly acknowledged that processes of conversational repair play a critical role in sustaining the mutual-intelligibility of interactions (Brennan, 1988; Clark, 1996), no specific mechanism is provided for dealing with it.

A more practical limitation on the application of the CM to the analysis of interaction is its relative underspecification. As Clark and Brennan (1991) note the media constraints and costs invoked to explain particular patterns of communication in different media are essentially heuristic. The list of costs and constraints are neither exhaustive nor exclusive and there is no means of quantifying the factors necessary to provide more precise analyses of the possible trade-offs involved. In order to make a systematic comparison of the costs and benefits imposed by different me-

dia some quantification of the communicative effort invested in an interaction would be required. Without some means of comparing the grounding criteria being employed in different cases i.e., of estimating the grounding criterion, the rationale behind particular patterns of communicative response cannot be determined. A pattern of small, installment, contributions is consequently consistent both with a situation in which, say, turn costs are low or a situation in which the grounding criterion is high.

2 Breakdown and Repair

The position developed in this paper is that the problems with functional and CM analyses of communication identified above can be addressed by focusing on the analysis of breakdown and repair in communication. Conceptually, the distinguishing feature of this approach is that it is concerned only with those parts of an interaction in which communicative trouble arises. A detailed framework for characterising these situations has been developed in the conversation analytic (CA) tradition (Sacks, Schegloff and Jefferson, 1974; Schegloff, 1987, 1992). Before describing the potential application of this framework to the analysis of mediated interactions, it is important to set out the basic CA repair framework and then clarify the concept of communicative problem it invokes.

Two basic aspects to the CA model are of particular relevance to the present paper. The structural or procedural elements and the analysis of specificity. Structurally, the CA repair framework distinguishes between three things; who initiates a repair, where in the turn taking structure it occurs, and the subsequent trajectory of the repair to completion. For example, self-initiated repair occurs where the speaker identifies a problem with one of their own turns in a conversation. Other-initiated repair occurs in situations in which someone signals a problem with another participant's turn. The point at which the problem is addressed, the 'repair', is also classified in terms of 'self' and 'other'. Self repair occurs where the person who produced the problematic utterance also addresses the problem, regardless of whether they signaled that it was problematic. Other repair occurs where someone addresses a putative problem in someone else's turn. Four positions are distinguished in which a problem can be signaled or addressed. First position repair occurs in the turn in which a problem occurs, second position repair takes place in the next turn that occurs as a response to the problem turn. Third position repair-initiation occurs in the next turn that occurs as a response to the second position and so on (Schegloff, 1992, 1987).

Repair initiations are also distinguished according to the specificity with which they localise a problem. Sacks, Schegloff and Jefferson (1974) proposed a non-exhaustive ordering of other-initiation types according to their power. The most general kind of initiation is a "huh?" or "what?" which signals that the utterer has a problem but gives almost no clues about its precise character. This is followed, in order of increasing specificity, by a 'wh' question, such as "who?" or "what". In this case the nature of the signaled problem is clearer and could potentially be associated with a particular sub-part of the problematic turn. More specific still are a partial repeat plus a 'wh' question or just a partial repeat. This type of reprise clarification provides information about the specific point in the original turn that caused the problem. The strongest form of repair initiation in Sacks et. al.'s ordering is a full paraphrase or reformulation prefaced by "you mean ". In this situation the recipient of the turn has successfully recognised and parsed what was said but wishes to test a possible interpretation of it with the original speaker. Sacks, Schegloff and Jefferson (1974) note that there is a preference for using the strongest or most specific type of initiation available in any given case. This is supported by the observed tendency to interrupt weaker initiations with stronger initiations and, where several initiations occur in sequence, for an increase in strength of initiator as they progress.

2.1 A Repair-based Analysis

The CA repair framework can be adapted to the comparative analysis of technologically mediated communication. In contrast to the approaches discussed above this framework focuses only on those junctures where something goes wrong in an interaction. The CA approach to breakdown and repair does not involve appeal to a model of what is 'actually' being communicated or a theory of error. This theoretical orientation is inherited from the Ethnomethodological and Phenomenological roots of CA (see Taylor (1992) for further discussion). For example, Schegloff (1992) states that:

> "adequacy of understanding and intersubjectivity is assessed not against some general criterion of meaning or efficacy (such as convergent paraphrase) and not by 'external' analysts, but by the parties themselves vis--vis the exigencies of the circumstances in which they find themselves." (Schegloff 1992, p.1338)

The question of whether a breakdown in communication occurs because of some genuine or objectively verifiable misunderstanding is explicitly suspended (Garfinkel, 1967). The focus instead is on analysing the situations which the interlocutors *treat* as problematic, independently of accounts of what is really being transacted in a given exchange or of whether a problem is really resolved by a repair. Consequently, the analysis is not concerned with whether a turn was in some sense correctly formulated or contains accurate information, but only with whether it was treated as intelligible by the participants themselves. For example, a request for clarification that does not signal a problem with the intelligibility of a preceding utterance is not a repair in the present sense. Moreover, a complaint about, say, audibility when using voice over IP does not count as a communication problem unless there is evidence that the recipient of the complaint had trouble understanding it.

This orientation has two potential practical benefits. Firstly, the analysis of communicative coordination is separated from analysis of the task domain. It is not necessary for the analyst to have a theory of what task people are engaged with or how it is carried out. The patterns of repair type and trajectory can be analysed independently of the transaction involved. In contrast to some functional schemes of analysis such as dialogue games, the analysis should thus be applicable to a variety of different kinds of communicative interaction.

The second potential benefit is that it promises to improve the sensitivity of the analysis by concentrating attention on those exchanges in which communicative problems occur. This claim trades on the assumption that, relative to other kinds of exchange, the frequency with which problems are signaled and addressed is moderated by their perceived importance to the coherence of the interaction. This assumption is based on the observation in the CA literature that turns which initiate repair are avoided if possible, especially those that signal problems with another participants contribution (Sacks, Schegloff and Jefferson, 1974). A focus on repair should thus help to filter out those exchanges which are incidental to the participants' purposes from those which are essential.

2.2 The Repair Protocol

In order to exploit the CA model in the comparative analysis of interaction it is necessary to develop a coding protocol that operationalises and specifies criteria for identifying the instances of repair of each type. This involves putting the CA framework to an unintended use. The CA repair model is based on the detailed analysis of spoken conversations and was developed in a tradition in which statistical generalisations are specifically eschewed (Schegloff, 1992), although see Frohlich, Drew and Monk (1994). The intention here is to exploit the procedures for signaling and addressing communicative problems described by CA. Inevitably some of the subtlety of the original observations is lost by

doing this. For example, the protocol does not capture 4th position repairs because they are too rare to be of service in making systematic comparisons.

The complete repair protocol is shown in Figure 3 in the appendix. It is constructed so that an analyst takes each turn in an exchange and tests it against the criteria specified in each box. These follow a binary decision format and are formulated to be as simple as possible. Because a given turn may contain more than one repair related event the protocol is applied recursively to a turn until no further repairs are detected.

Although built on empirical studies of verbal interactions, the protocol is designed to be modality neutral. It aims to capture repair phenomena across a variety of modalities including, for example, graphical and gestural interaction. As a result it refers to initiators rather than speakers and, following Clark (1989), contributions, and modifications to contributions, rather than turns. It has also been designed to avoid reliance on clearly identifiable sequences of turns. This is important for situations such as text and whiteboard based interaction where turn sequence is less reliably maintained than in verbal interaction. The protocol does require, however, that analysts can identify what, if any, preceding parts of the interaction a contribution may be addressed to.

The protocol has been designed to be used by people who have no specific knowledge of the body of conversation analytic research that it draws on. However, one potentially count-intuitive aspect of this heritage is that the protocol is not concerned with whether a turn was in some sense correct but only with whether it was treated as intelligible by the participants. For example, a question for clarification that does not signal a problem with the intelligibility of a preceding utterance is not a repair in the present sense. This approach is incorporated into the protocol through the use of retrospective criteria for determining the type, for example, of a position two (P2) or position three (P3) repair by assessing whether they occurs in response to a *prior* turn. Additionally, if a participant makes a statement that is an error from the analyst's point of view this will not be coded as a communication problem unless it is treated as such by the participants.

2.3 Measuring Communicative Co-ordination

Repair *per se* is not necessarily an indicator of lower communicative coherence. It could, for example, reflect greater efforts to understand exactly what is being said or reflect a more complex exchange. Instead of making overall comparisons of the frequency of repair, the proposal is to use the structure and distribution of specific types of breakdown and repair to provide indices of communicative coordination. The protocol has yet to be systematically tested and this is the subject of ongoing work. This section illustrates the potential of this approach to analyse multimodal communication by suggesting some of the potential measures of communicative coordination it could provide.

The simplest index that this approach can provide is a measure of the difficulty of producing a contribution in a particular medium. In conversation a significant proportion of communicative problems relate to problems with articulating an utterance. Analogous problems arise in text chat where typos are frequently a problem, and in drawing where problems with the execution of shapes and letters are common. In the protocol problems of this kind are classified as Articulation problems and their frequency of occurrence provides a basic index of the difficulty of externalising a contribution in a particular medium. Because the protocol captures only those 'typos' or 'disfluencies' that the initiator of a contribution chooses to correct, it reflects the participants estimate of the impact of the articulation problem on the effectiveness of the interaction. It can also be used to index the grounding criteria (see above) that an individual employs. For example, if we hold task and media constant the frequency of Articulation repairs should be proportional to the effort being invested in making them-

selves understood.

A second index that the protocol can provide is a measure of the effect of a medium on the difficulty of formulating a contribution. In the protocol this is captured by the position 1, self-initiated, self-repairs (P1,SI,SR). These are repairs in which the initiator makes modifications that, that unlike typos, alter the possible meaning of their contribution during production. For example a referring expression may be rephrased by replacing "he" with "she", or part of a drawing may be erased and redrawn before being presented as complete. It might be expected that media which produce a persistent representation of a contribution, e.g., text chat or email, should, all things being equal, lead to more Formulation repairs than those that produce a transient representations, for example speech. Alternatively, if we hold the medium constant then the frequency of Formulation repairs should vary as a function of the cognitive load the task places on participants.

Perhaps the most interesting potential measures are those which promise to directly index the communicative load imposed by different media and tasks. One way in which this could be addressed is by assessing the frequency of, for example, position two and position three repair initiations under different task and media conditions. All things being equal, if a particular medium alters the intelligibility of interaction in some situation then this should be reflected in the frequency of repair initiations. More subtle distinctions can also be made. Arguably, self-initiated, self-repair in position three is indicative of high communicative co-ordination since it depends on sensitivity to a recipient's interpretation of one of the initiator's preceding utterances. Measures like this could provide for characterisation of the relative communicative 'transparency' of different media.

One further interactional measure can be derived from the analysis of the specificity of the problems encountered. The ability of interlocutors to efficiently localise and deal with a problem provides an index of their communicative coherence. One possibility provided directly by the CA framework is to exploit the ranking of initiation types according to their power to locate a 'repairable' as discussed above. However, the notions of paraphrase and 'wh question do not generalise to non-verbal interaction. If it is assumed that, all things being equal, more severe problems will require more extensive repairs then analysis can focus on the amount of the preceding material that is replaced. For verbal exchanges this could be indexed by the proportion of words altered or amended in the repair. For graphical exchanges it can be indexed by the proportion of a drawing or sketch that is revised.

3 Discussion

The present proposal is that a repair-based analysis can provide useful operationalisations of several aspects of communicative coordination. The discussion of specific measures is however speculative. It is also an open question whether the categories proposed by conversation analysts, and abstracted in the protocol, can be reliably identified by independent judges. Although there is a large body of research which has applied the CA analyses to a variety of examples, interjudge agreement between different individuals has not, to our knowledge, been assessed. This is a prerequisite for the applications proposed here and must be evaluated in future work.

The present claim is that a repair-based approach can provide a more effective analysis of communicative coordination than existing applications of psycholinguistic techniques. One important gap in the current protocol is that it doesn't currently capture problems with the handover of turns. Anecdotal observations of remote mutli-modal communication suggest that this is an important class of communiction problem, particularly where there are more than two participants.

An interesting corollary of the present analysis is that communicative coherence should be enhanced by situations or technologies that make visible as much of the structure of each individual's contribution as possible. This follows from the observation that repairs are ini-

tiated and effected by manipulating the structure of preceding contributions. To give a concrete example, in current implementations of text chat it is difficult to signal a problem with preceding turns. It is difficult both to identify the preceding turn itself and to identify what elements of the turn were problematic. Users usually have to repeat the problematic material together with some identifier in order to effect a repair initiation. Shared whiteboards, by contrast, support much simpler devices. For example, users can circle or underline problematic contributions directly. The original contribution is thus more easily operated on on a whiteboard than in text chat. Media or environments that allow users to manipulate the structure of each others contributions should, on this view, provide more effective support for co-ordinating understanding.

Acknowledgments

This work was generously supported by Avaya research Laboratories under the "Mixed Mode Communication" Research project. This paper supersedes an earlier, unpublished, version presented at the 1999 AAAI fall symposium "Psychological Models of Communication in Collaborative Systems". We gratefully acknowledge Marcus Colman, Anjum Khan, Ioannis Spyradakis and Sylvia Wilbur for their contributions to the development of this work.

References

Anderson, A. H., O'Malley, C., Doherty-Sneddon, G., Langton, S., Newlands, A., Mullin, J., Fleming, A. M., & Velden, J. Van der. (1997). The impact of vmc on collaborative problem solving: An analysis of task performance, communicative process and user satisfaction. In K. E. Finn, A. Sellen, & S. B. Wilbur (Eds.), *Video-mediated communication* (pp. 133–155). Mahwah, New Jersey: Lawrence Earlbaum Associates.

Bowers, J., Button, G., & Sharrock, W. (1995). Workflow from within and without: technology and cooperative work on the print industry shop floor. In *Proceedings of the fourth european conference on computer-supported cooperative work* (pp. 51–66). New York: ACM Press.

Brennan, S. (1988). The grounding problem in conversations with and through computers. In S. R. Fussell & R. J. Kreuz (Eds.), *Social and cognitive approaches to interpersonal communication* (pp. 201–225). Mahwah : Lawrence Erlbaum Associates.

Button, D., & Dourish, P. (1996). Technomethodology: Paradoxes and possiblities. In *Proceedings of chi'96* (pp. 19–26). New York: ACM Press.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1996). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, *23*(1), 13–31.

Clark, H., & Brennan, S. (1991). Grounding in communication. In L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–150). American Psychological Association.

Clark, H. H. (1996). *Using language.* Cambridge: Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*, 259–294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Frohlich, D., Drew, P., & Monk, A. (1994). The management of repair in human-computer interaction. *Human-Computer Interaction*, *9*, 385–425.

Garfinkel, H. (1967). *Studies in ethnomethodology.* Englewood Cliffs: Prentice Hall.

Heath, C., & Luff, P. (1992). Collaboration and control: Crisis management and multimedia technology in london underground control rooms. *Computer Supported Cooperative Work*, *2*.

Heath, C., & Luff, P. (2000). *Technology in action.* Cambridge: Cambridge University Press.

Hughes, J., Randall, D., & Shapiro, D. (1992). Faltering from ethnography to design. In *Cscw '92: Proceedings of the conference on computer-supported cooperative work* (pp. 115–122).

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*, 265–288.

Kowtko, J. C., Isard, S. D., & Doherty, G. M. (1991). Conversational games within dialogue. In *Proceedings of the espirit workshop on discourse coherence.*

Nardi, B., & Miller, J. (1991). Twinkling lights and nested loops: distributed problem solving and spreadsheet development. *International Journal of Man-Machine Studies*, *34*, 161–184.

O'Conaill, B., Whittaker, S., & Wilbur, S. (1993). Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, *8*, 389–428.

OMalley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., & Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, *8*(2), 177–192.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, *50*, 696–735.

Schegloff, E. A. (1987). Some sources of misunderstanding in talk-in-interaction. *Linguistics*, *25*, 201–218.

Schegloff, E. A. (1992). Repair after the next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, *97*(5), 1295–1345.

Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication.* Cambridge: Cambridge University Press.

Taylor, T. J. (1992). *Mutual misunderstanding: Scepticism and the theorizing of language and interpretation.* London: Routledge.

Appendix: Repair Protocol

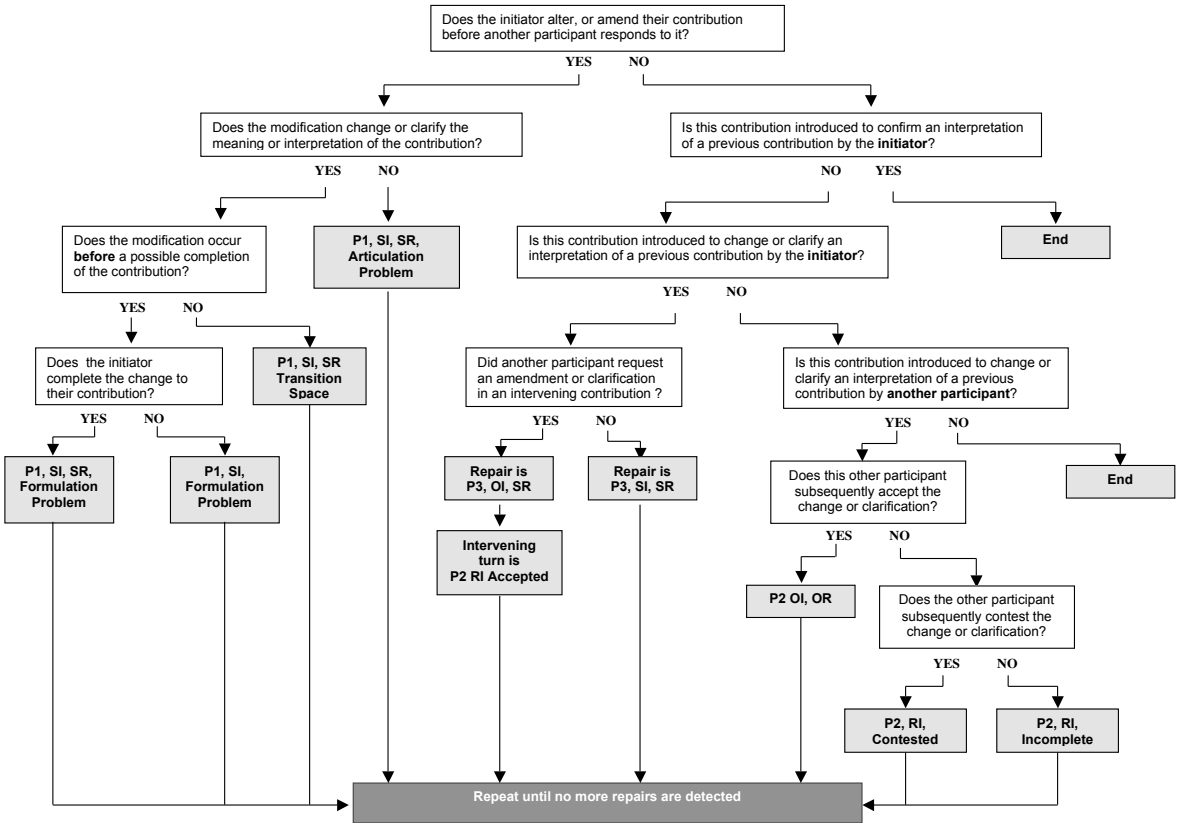


Figure 1: Repair Analysis Protocol

Experimenting with the Gaze of a Conversational Agent

Dirk HEYLEN **Ivo VAN ES** **Anton NIJHOLT** **Betsy VAN DIJK**
Computer Science, University of Twente
PoBox 217
7500 AE Enschede, The Netherlands,
{heylen,es,anijholt,bvdijk}@cs.utwente.nl

Abstract

We have carried out a pilot experiment to investigate the effects of different eye gaze behaviors of a cartoon-like talking face on the quality of human-agent dialogues[1]. We compared a version of the talking face that roughly implements some patterns of human-like behavior with two other versions. We called this the optimal version. In one of the other versions the shifts in gaze were kept minimal and in the other version the shifts would occur randomly. The talking face has a number of restrictions. There is no speech recognition, so questions and replies have to be typed in by the users of the systems. Despite this restriction we found that participants that conversed with the optimal agent appreciated the agent more than participants that conversed with the other agents. Conversations with the optimal version proceeded more efficiently. Participants needed less time to complete their task.

Introduction

Research on embodied conversational agents is carried out in order to improve models and implementations simulating aspects of human-like conversational behavior as best as possible. Ultimately, one would like the synthetic characters that one is building to be believable, trustworthy, likeable, human- and life-like. This involves, amongst other things, having the character display the appropriate signs of a changing mood, a recognisable personality and a rich emotional life. The actions that have to be carried out by agents in dialogue situations include the obvious language understanding and generation tasks, knowing how to carry out a conversation and all the types of conversational acts this involves (openings, greetings, closings, repairs, asking a question, acknowledging, back-channeling, etc.) and also using all the different modalities, including body-language (posture, gesture, and facial expressions).

Although embodied conversational agents are still far from perfect, some agents have already been developed that can perform quite a few of the functions that were listed above to a reasonable extent and that can be useful in practical applications like tutoring (Cassell, 2001).

In our research laboratory we started to develop spoken dialogue systems some years ago. We focused on an interface to a database containing information on performances in the local theatres. Through natural language dialogue, people could obtain information about performances and order tickets. A second step involved reconstructing one of the theatres in 3D using VRML and design a virtual human, Karin, that embodies this dialogue systems. We first focused the attention on several aspects of the multi-modal presentation of information (Nijholt and Hulstijn, 2000). We combined presentation of the information through the dialogue system with traditional desktop ways of presentation through tables, pop-up menus and we combined natural language interaction with keyboard and mouse input. We wanted our basic version to be web-accessible which, for reasons of efficiency, forced us at that time to leave out the speech recognition interface from this version. We have moved on to implement other types of embodied conversational agents that are designed to carry out other tasks like navigating the user through the virtual environment or agents that act as tutors. Besides the work we did on building other types of agents we have also tried to

[1] Short 2 page papers related to this experiment were submitted to the CHI 2002 conference (Minneapolis) and AVI (Trento) and accepted for presentation. We have benefitted greatly from comments made by anonymous reviewers to these versions.

explore in more depth different cognitive and affective models of agents, including symbolic BDI models as well as neural network models. We have also worked on extending their communicative skills. Current work, as summarised in Heylen et al. (2001), is concerned with several aspects of non-verbal behavior including facial expressions, posture and gesture, and gaze (which is the topic of this paper).

In the next section of this paper we will discuss some aspects of the function of gaze in face-to-face conversations between humans and in mediated forms. Next we describe our experiment and discuss the outcome.

1. Functions of (mutual) gaze

The function of gaze in human-human, face-to-face dialogues has been studied quite extensively (see Argyle and Cook (1976) and many other publications mentioned in the references). The way speakers and hearers seek or avoid mutual eye contact, the function of looking to or away from the interlocutor, the timing of this behavior in relation to aspects of discourse and information structure have all been investigated in great detail and certain typical patterns have been found to occur. In these investigations a lot of parameters like age, gender, personality traits, and aspects of interpersonal relationships like friendship or dominance have been considered.

Gaze has been shown to serve a number of functions in human-human interaction (Kendon, 1990). It helps to regulate the flow of conversation and plays an important role in ensuring smooth turn-taking behavior. Speakers, for instance, have the tendency to gaze away from listeners at potential turn-taking positions when they want to keep on talking. Listeners show continued attention when gazing at the speaker. Duration and types of gaze communicate the nature of the relationship between the interlocutors.

In trying to build life-like and human-like software agents that act as talking heads which humans can interact with as if they were talking face-to-face with another human, one is forced to consider the way the agents look away and towards the human interlocutor. This has been the concern of several researchers on embodied conversational agents and on other forms of mediated communication as in teleconferencing systems that make use of avatars, for instance. Previous research was mostly concerned with trying to describe an accurate computational model of gaze behavior. Evaluations of the effects of gaze on the quality of interactions in mediated conversation (mostly avatars instead of autonomous agents) have been carried out by Vertegaal (1999), Garau et al. (2001), Colburn et al. (2000) and Thórisson and Cassell (1996), amongst others. These papers have shown that improving gaze behavior of agents or avatars in human-agent or human-avatar communication has noticeable effects on the way communication proceeds. This made us curious about our own situation with the agent Karin. We wondered whether implementing some kind of human-like rules for gaze behavior would have any effects given her somewhat limited dialogue functionality, her cartoon-like face, the somewhat unnatural way of input that lets users type in their questions only instead of using speech and the fact that the face is only one modality amongst others that is used to present information. We therefore set up our experiment which is further described in Section 3.

1.1 Human to Human

The amount of eye contact in a human-human encounter varies widely. Some of the sources of this variation as well as some typical patterns that occur have been identified. Women, for instance, are found to engage in eye contact more than men. Cultural differences account for part of the variation as well.

When people in a conversation like each other or are cooperating there is more eye contact. When personal or cognitively demanding topics are discussed eye contact is avoided. Stressing the fact that the following figures are only averages and that wide variation is found, Argyle (1993) provides the following statistics on the percentage of time people look at one another in dyadic (two-person) conversations.

Individual gaze	60 %
While listening	75 %
While talking	40 %
Eye-contact	30 %

Among the common subjective interpretations of eye contact have been found friendship, sexual

attraction, hate and a struggle for dominance. Gaze levels are also higher in those who are extroverted, dominant or assertive, and socially skilled. People who look more tend to be perceived more favourably, other things being equal, and in particular as competent, friendly, credible, assertive and socially skilled (Kleinke, 1987). Besides these more psychological or emotional signal functions of gaze, looking to the conversational partner also plays an important part in regulating the interaction. The patterns in turn taking behavior and the relation to (mutual) gaze have been the subject of several investigations. In our experiment we wanted to focus the attention on the way appropriate rules of gazing of the agent would improve the quality of the conversation. However, we also wanted to see whether the different patterns that we had chosen would affect the way our agent was liked or disliked.

Studying the patterns in gaze and turn-taking behavior, Kendon (1990) was one of the first to look with some detail at how gaze behaviour operates in dyadic conversations. He distinguishes between two important functions of an individual's perceptual activity in social interaction. By looking or not looking, a person can control the degree of monitoring his interlocutor and this choice can also have regulatory/expressive functions.

Argyle and Dean (1972) report that in all investigations where this has been studied it has been found that there is more eye contact when the subject is listening than when he is speaking (cf. the table above). Furthermore people look up at the end of their turn and/or at the end of phrases and look away at the start of (long) utterances, not necessarily resulting in mutual gaze or eye contact. The patterns in gaze behaviour are explained by a combination of principles. Speakers that start longer utterances tend to look away to concentrate on what they are saying, avoiding distraction, and to signal that they are taking the floor and do not want to be interrupted. At the end of a turn, speakers tend to look up to monitor the hearer's reaction and to offer the floor.

In Cassell et al. (1999), the relation between gaze, turn-taking, and information structure is investigated in more detail. The empirical analysis shows the general pattern of looking away and looking towards the hearer at turn-switching positions. The main finding reported in this paper, is that if the beginning of a turn starts with the thematic part (the part that links the utterance with previously uttered or contextualised information), then the speaker will always look away and when the end of the turn coincides with a rhematic part (that provides new information), than the speaker will always look towards the hearer at the beginning of the rhematic part. In general, beginnings of themes and beginnings of rhemes are important places where looking away and looking towards movements occur.

1.2 Mediated Conversation

Several researchers have investigated the effects of implementing gaze behavior in conversational agents or in other forms of mediated conversation. In videoconfering for instance, avatars may be used to represent the users.

Vertegaal (1999) describes the GAZE groupware system in which participants are represented by simple avatars. Eye-tracking of the participants informs the direction in which the avatars appear to look at each other on the screen (see also Vertegaal et al., 2001).

Garau et al. (2001) describe an experiment with dyadic conversation between humans in 4 mediated conditions: video, audio-only, random-gaze avatars and informed gaze avatars (gaze was related to conversational flow). The experiment showed that the random-gaze avatar did not improve on audio-only communication, whereas the informed gaze-avatar significantly outperformed audio-only on a number of response measures.

Colburn et al. (2000) also describe some experiments in conversations between humans and avatars in a video-conferencing context. One of the questions they asked was whether users that interact with an avatar will act in ways that resemble human-human interaction or whether the knowledge that they are talking to an artificial agent counteracts natural reactions. In one experiment they changed the gaze behavior of avatars during a conversation. It appears from this and similar experiments that participants while not consciously aware of the differences in the avatar’s gaze behavior will still react differently (subliminally).

In the context of embodied conversational agents, rules for gaze behavior of agents have

been studied by Cassell et al. (1994, 1999). Algorithms and architectures for controlling the non-verbal behavior, including gaze, of agents are also presented in Chopra et al. (2001) and Novick et al. (1996). These have focussed mainly on getting the appropriate computational models instead of on evaluation. Previous work on evaluation in this respect is reported in Thórisson and Cassell (1996). They found that conversations with a gaze informed agent increased ease/believability and efficiency compared to a content-only agent and an agent that produced content and emotional emblems.

In our pilot experiment described in the next section, we were not so much interested in the precise rules or the architecture of the system implementing the rules, but rather in the effects on dialogue quality that a simple implementation of the patterns might have. Some of the factors that we wanted to look into are the efficiency of interactions, the way people judge the character of the agents and how they rate the quality of the conversation in general.

Although the work on evaluation of gaze behavior has not been concerned to any great extent with autonomous embodied conversational agents, the evaluation work on human-controlled avatars and mediated conversation seemed to provide a promise for reasonable effects in mediated conversations with agents in general and even with our agent Karin whom users have to interact with by typing in their utterances and who presents information also in the form of tables.

2 Our experiment

In our experiment we compared three versions of Karin that differed with respect to gaze behavior. We had 48 participants each carry out two ticket reservation tasks with one version of Karin. After they had finished, they filled out a questionnaire. Together with some other measures (such as the time it took them to complete the tasks) this data was used to evaluate the implementations on a number of factors.

2.1 Participants, task and procedure

The 48 participants in our experiment were all graduate students of the University of Twente, aged between 18 and 25, two thirds were male and one third female. These participants were randomly assigned to one of the three conditions, taking care that the male/female ratio was roughly the same for each.

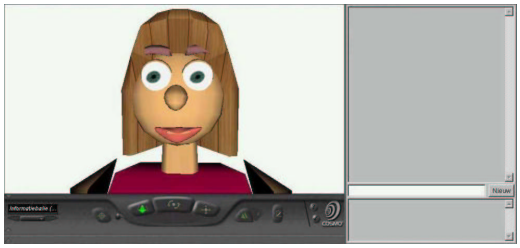
The participants were given the task to make reservations for two concerts. During the execution of the task they were left alone in a room monitored by two cameras. After they finished the task they filled out a questionnaire. The questionnaire together with the notes taken when observing the particpants through the camera and the time it took for the participants to complete the task were used to evaluate the differences between the three versions of the agent.

2.2 Versions

In the following table a part of a typical conversation is given with indications of how Karin turns her eyes away and towards the human participant. We show the optimal and suboptimal version. In the “direct” behavior, Karin turns her eyes downwards, towards a table that contains information about the performances.

		Optimal	Sub
K	Hello, I’m Karin.	Avert	Gaze
	What can I help you with?	Gaze	Gaze
S	Hi. When is the next concert of X?		
K	Just a moment, while I look it up. There are 27 concerts.	Avert	Gaze
	Take a look at the table	Direct	Direct
	For the dates.	Gaze	Gaze
S	I want to book tickets for the concert on November 7.		
K	You want to make a reservation for the Lunch series.	Avert	Gaze
	I have the following information for this series:	Gaze	Gaze
	20 guilders normal rate.	Avert	Gaze
	How many tickets do you want?	Gaze	Gaze

In the optimal version Karen will avert her eyes at the beginning of a turn for a short period and then start gazing again. In general Karin's replies are quite short. But some consist of some longer sequences, for instance, when she repeats the information she has so far and also adds a question to initiate the next step in the reservation. This is illustrated by the last reply. In that case, Karin will avert her eyes from the speaker to indicate that she is not ready yet and does not want the user to take the turn. We have tried to time eye-movements and information-structure in accordance with the rules described by Cassell et al. (1999).



We introduced a second version, in which Karin will only stop looking at the user when she directs the users with her eyes to the table with the performances. Eye-movements are therefore limited severely in this version. In the third version a random eye-movement action was chosen at each position at which a specific eye-movement change could occur in the optimal version.

2.3 Measures

In general, we wanted to find out whether participants talking to the optimal version of Karin were more satisfied with the conversation than the other participants. We distinguished between several factors that could be judged: *ease of use*, *satisfaction*, *involvement*, *effiency*, *personality/character*, *naturalness* (of eye and head movements) and *mental load*. Most of the measures were judgements on a five point Likert scale (<agree>/<disagree>). A selection of the questions asked is presented below. Some factors were evaluated by taking other measures into account. The time it took to complete the tasks was used, for instance, to measure efficiency. We asked participants some questions about the things said in the dialogue to judge differences in attention (mental load).

Satisfaction

I <liked> / <didn't like> talking to Karin
It takes Karin too long to respond
The conversation had a clear structure
I like ordering tickets this ways

Ease of Use

It is easy to get the right information
It was clear what I had to ask/say
It took a lot of trouble to order tickets

Involvement

I think I looked at Karin about as often as I look to interlocutors in normal conversations
Karin keeps her distance
It was always clear when Karin finished speaking

Personality

I trust Karin
Karin is a friendly person
Karin is quite bad tempered

We were not sure whether participants would be influenced a lot by the differences in the gaze behavior. However, if there were any effects, we assumed that the optimal version would be most efficient, in that it signals turn-taking mimicking human patterns.

2.4 Results

Efficiency was analyzed using a one-way ANOVA test. A significant difference was found between the three groups ($F(2,45)=3.80$, $p<.05$). For means and corresponding standard deviations see the table below. To find out which version was most efficient, the groups were compared two by two using t-tests (instead of post-hoc analysis). The optimal version was found to be significantly more efficient than the subobtimal version ($t(30)=-2.31$, $p<.05$, 1-tailed) and the random version ($t(30)=-2.64$, $p<.01$). No significant difference (at 5% level) was found between the suboptimal and the random version.

The main effect of the experimental conditions on the other factors was analyzed using the Kruskal-Wallis test. Answers to questions were recoded such that for all factors the best possible score was 1 and the worse score was 5. The results are summarized in the table. The table shows significant differences between the versions for ease of use, satisfaction and naturalness of head movement and a marginally significant difference for personality.

The main effects of experimental condition: means and standard deviations (in parentheses) of the factor scores and the results of the Kruskal-Wallis test

Factors	Opti	Sub	Ran	X^2
Ease of use	2.55 (1.31)	3.05 (1.30)	2.66 (1.17)	12.09**
Satisfaction	2.33 (1.20)	2.74 (1.29)	2.79 (1.20)	9.63**
Involvement	3.08 (1.35)	3.47 (1.28)	3.47 (1.17)	3.53
Personality	2.46 (1.21)	2.79 (1.27)	2.79 (1.14)	5.62†
Natural head movement	1.31 (.62)	1.31 (.55)	1.63 (.61)	11.66**
Natural eye movement	1.13 (.39)	1.13 (.49)	1.29 (.58)	3.34
Mental load	2.54 (1.27)	3.02 (1.31)	2.63 (1.20)	3.93
Efficiency	6.88 (2.00)	8.88 (2.83)	9.56 (3.56)	-

† $p<.10$ * $p<.05$ ** $p<.01$

Two by two comparisons using Mann-Whitney tests pointed out that on the factor *ease of use* the optimal version was significantly better than the suboptimal version (U=6345, $p<.001$). Users of the optimal version were more *satisfied* than users of the suboptimal and the random version (resp. U=5140, $p<.05$ and U=4913.5, $p<.01$). On the factor *personality* the optimal version was better than the random version (U=5261.5, $p<.05$) and marginally better than the suboptimal version (U=5356.5, $p<.10$). Both the optimal and the suboptimal agent *moved* their *head* more naturally than the random agent (resp. U=805.5, $p<.01$ and U=823.5, $p<.01$). The *eye movements* were found to be marginally better in the optimal version than in the random version (U=1006, $p<.10$). On the factor *mental load* the difference between the optimal version and the suboptimal version was marginally significant (U=910, $p<.10$). The other comparisons yielded no significant differences.

3 Discussion

The table clearly shows that the optimal version performs best overall. We can thus conclude that even a crude implementation of gaze patterns in turn-taking situations has significant effects. Not only do participants like the optimal version best, they also perform the tasks much faster and tend to be more involved in the conversation. The more natural version is preferred above a version in which the eyes are fixed almost constantly and a version in which the eyes may move as much as in the optimal situation but do not follow the conventional patterns of gaze.

To measure satisfaction participants were asked to rate how well they liked Karin and how they felt the conversation went in general besides some other questions that relate directly or indirectly to what can be called satisfaction. The participants of the optimal version were not only more satisfied with their version, but they also related more to Karin than the participants of the other versions did as they found her to be more friendly, helpful, trustworthy, and less distant. The differences between the optimal and the suboptimal version seem to correspond to patterns observed in human-human interaction. In the suboptimal version, Karin looks at the visitor almost constantly. Although in general it is the case that people who look more tend to be perceived more favourably, as mentioned above (Kleinke, 1987), in this case the suboptimal version in which Karin looks at the participants the most of all the versions is not the preferred one. This, however, is in line with a conclusion of Argyle et al. (1974) who point out that continuous gaze can result in negative evaluation of a conversation partner. This is probably the major explanation behind the negative effect on how Karin is perceived as a person in this version. Note that Karin still looks at participants quite a lot in the optimal version as she only looks away at beginning of turns and at potential turn-taking positions when she wants to keep the turn, otherwise she will look at the listener while speaking. She also looks towards the interlocutor while listening. She therefore seems to have found an adequate equilibrum in gazing a lot to be liked but not too much.

When participants have to evaluate how natural the faces behave it appears that the random version scored lower than the other versions but no differences could be noted between the optimal and suboptimal version. Making “the right” head and eye movements or almost no movements are both conceived of as being equally natural, whereas random movements are judged less natural. What is interesting, however, is that these explicit judgements on the life-likeness of the behavior of the agents do not

reflect directly judgments on other factors. The random version may be rated as less natural than the others but in general it does not perform worse than the suboptimal version. For the factor ease of use it is judged even significantly better than the suboptimal version. Does this mean that having regular movements of the eyes instead of almost fixed eyes is the important cue here? On the other hand, the difference in this rating (which is gotten from judgments on questions like "does it take Karin long to respond", "was it easy to order tickets") is not in line with the real amount of time people actually spent on the task. Though the random version is judged easy to use, it takes the participants using it the most time to complete the tasks.

The optimal version is clearly the most efficient in actual use. This gain in efficiency might be a result of the transparancy of turn-taking signals; i.e. the flow of conversation may have improved as one would assume when regulators like gaze work appropriately. But the gain might also have been a result, indirectly, of the increased involvement in the conversation of the participants that used the optimal version. Whatever is cause or effect is difficult to say. We have an indication that the different gaze patterns had some impact not just on overall efficiency but also on the awareness of participants about when Karin was finishing her turn. We have some rough figures on the number of times participants started their turn before Karin was finished with hers. In almost all of these cases this slowed down the task, because participants would have to redo change their utterance midway.

	Opt	Sub	Ran
Often/Regularly		5	4
Sometimes	4	2	3
Never	12	9	9

These figures are not conclusive, but give an indication that at least in the optimal version, participants paid more attention to Karin than in the other versions.

4 Conclusion

In face-to-face conversations between human interlocutors, gaze is an important factor in signalling interpersonal attitudes and personality. Gaze and mutual gaze also function as indicators that help in guiding turn-switching. In the experiment that we have conducted, we were interested in the effects of implementing a simple strategy to control eye-movements of an artificial agent at turn-taking boundaries.

The crude rules that we have used are sufficient to effect significant improvements in communication between humans and embodied conversational agents. So, therefore, the effort to investigate and implement human-like behavior in artificial agents seems to be well worth the investment.

References

M. Argyle (1993) *Bodily Communication*. Routledge, second edition.

M. Argyle, M. Cook (1976) *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge.

M. Argyle, J. Dean (1972) Eye contact, distance and affiliation. Reprinted in. J. Laver, S. Hutcheson (eds.) *Communication in Face to Face Interaction*, Penguin [1962 original] (p. 155-171).

M. Argyle, L. Lefebre, M. Cook (1974) The Meaning of Five Patterns of Gaze. In: *European Journal of Social Psychology*, 4(2) (p.125-136).

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, M. Stone (1994). Animated Conversation. Rule Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. In: *Computer Graphics* (p. 413-420).

J. Cassell, O. Torres, S. Prevost (1999). Turn Taking vs. Discourse Structure, in *Machine Conversations* (p. 143-154).

J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.) (2000) Embodied Conversational Agents, MIT Press.

S. Chopra-Khullar, N.I. Badler (1999). Where to look? Automating attending behaviors of virtual human characters. In: *Proceedings of Autonomous Agents*. Seattle.

R.A. Colburn, M.F. Cohen, S.M. Drucker (2000) Avatar Mediated conversational interfaces. Microsoft Technical Report. MSR-TR-2000-81. July 2000.

M. Garau, M. Slater, S. Bee, M.A. Sasse (2001) The impact of eye gaze on communication using humanoid avatars. In: *CHI* 2001 (p. 309-316).

D. Heylen, A. Nijholt & M. Poel (2001) Embodied agents in virtual environments: The Aveiro project. In: Proceedings European Symposium on

Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Tenerife, Spain, December 2001, Verlag Mainz, Wissenschaftsverlag Aachen, 110-111.

A. Kendon (1990) Some functions of gaze direction in two-person conversation. Reprinted in: *Conducting Interaction*, Cambridge University Press, Cambridge (p. 51-89).

C.L. Kleinke (1987). Gaze and Eye Contact: a research review. In: *Psychological Bulletin*, 100 (p. 78-100).

A. Nijholt, J. Hulstijn (2000) Multimodal Interactions with Agents in Virtual Worlds. In: N. Kasabov (ed.) *Future Directions for Intelligent Information Systems and Information Science*, Physica-Verlag, (p. 148-173).

D.G. Novick, B. Hansen, K. Ward (1996) Coordinating Turn-Taking with Gaze. In: *Proceedings ICSLP*.

K.R. Thórisson, J. Cassell (1996) Why Put an Agent in a Body: the importance of communicative feedback in human-humanoid dialogue. Presented at Lifelike Computer Characters, Utah, October 1996.

R. Vertegaal (1999) The GAZE Groupware system: Mediating Joint Attention in Multiparty Communication and Collaboration. In: *Proceedings of CHI'*99, Pittsburgh, ACM Press (p. 294-301).

R. Vertegaal, R. Slagter, G. van der Veer, A. Nijholt (2001) Eye Gaze Patterns in Conversation. There is more to conversational agents than meets the eyes. In: *Proceedings of CHI 2001 Anyone. Anywhere*. ACM.

FORM: An Extensible, Kinematically-based Gesture Annotation Scheme

Craig Martell
Department of Computer and Information Sciences
and Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104 USA
cmartell@unagi.cis.upenn.edu

Abstract

Annotated corpora have played a critical role in speech and natural language research; and, there is an increasing interest in corpora-based research in sign language and gesture as well. We present a non-semantic, geometrically-based annotation scheme, FORM, which allows an annotator to capture the kinematic information in a gesture just from videos of speakers. In addition, FORM stores this gestural information in Annotation Graph format—allowing for easy integration of gesture information with other types of communication information, e.g., discourse structure, parts of speech, intonation information, etc.[1]

1 Introduction

FORM[2] is an annotation scheme designed both to describe the kinematic information in a gesture, as well as to be extensible in order to add speech and other conversational information.

Our goal is to build an extensible corpus of annotated videos in order to allow for general research on the relationship among the many different aspects of conversational interaction. Additionally, further tools and algorithms to add these annotations and evaluate inter-annotator agreement will be developed. The end result of this work will be a corpus of annotated conversational interaction, which can be:

- extended to include new types of information concerning the same conversations; as new tag-sets and coding schemes are developed—discourse-structure or facial-expression, for example—new annotations could easily be added;
- used to test scientific hypotheses concerning the relationship of the paralinguistic aspects of communication to speech and to meaning;
- used to develop statistical algorithms to automatically analyze and generate these paralinguistic aspects of communication (e.g., for Human-Computer Interface research).

2 Structure of FORM[3]

FORM is designed as a series of tracks representing different aspects of the gestural space. Generally, each independently moved part of the body has two tracks, one track for Location/Shape/Orientation, and one for Movement. When a part of the body is held without movement, a Location object describes its position and spans the amount of time the position is held. When a part of the body is in

[1] This presentation is a shortened version of (Martell, 2002)

[2] The author wishes to sincerely thank Adam Kendon for his input on the FORM project. He has provided not only suggestions as to the direction of the project, but also his unpublished work on a kinematically-based gesture annotation scheme was the FORM project's starting point (Kendon, 2000).

[3] The author wishes to acknowledge Jesse Friedman and Paul Howard in this section. Most of what is written here is from their "Code Book" section of http://www.ldc.upenn.edu/Projects/FORM/.

motion, Location objects with no time period are placed at the beginning and end of the movement to show where the gesture began and ended. Location objects spanning no period of time are also used to indicate the Location information at critical points in certain complex gestures.

An object in a movement track spans the time period in which the body part in question is in motion. It is often the case that one part of the body will remain static while others move. For example, a single hand shape may be held throughout a gesture in which the upper arm moves. FORM's multi-track system allows such disparate parts of single gestures to be recorded separately and efficiently and to be viewed easily once recorded. Once all tracks are filled with the appropriate information, it is easy to see the structure of a gesture broken down into its anatomical components.'

At the highest level of FORM are groups. Groups can contain subgroups. Within each group or subgroup are tracks. Each track contains a list of attributes concerning a particular part of the arm or body. At the lowest level (under each attribute), all possible values are listed. Described below are the tracks for the Location of the Right or Left UpperArm.

Right/Left Arm

Upper Arm (from the shoulder to the elbow).

Location

UPPER ARM LIFT (from side of the body)

- no lift
- 0-45
- approx. 45
- 45-90
- approx. 90
- 90-135
- approx. 135
- 135-180
- approx. 180

RELATIVE ELBOW POSITION: The upper arm lift attribute defines a circle on which the elbow can lie. The relative elbow position attribute indicates where on that circle the elbow lies. Combined, these two attributes provide full information about the location of the elbow and reveal total location information (in relation to the shoulder) of the upper arm.

- extremely inward
- inward
- front
- front-outward
- outward (in frontal plane)
- behind
- far behind

The next three attributes individually indicate the direction in which the biceps muscle is pointed in one spatial dimension. Taken together, these three attributes reveal the orientation of the upper arm.

BICEPS: INWARD/OUTWARD

- none
- inward
- outward

BICEPS: UPWARD/DOWNWARD

- none
- upward
- downward

BICEPS: FORWARD/BACKWARD

- none
- forward
- backward

OBSCURED: This is an binary attribute which allows the annotator to indicate if the attributes and values chosen were "guesses" necessitated by visual occlusion. This attribute is present in each of FORM's tracks.

Again, we have only presented the Location tracks for the Right or Left Arm UpperArm group.The full "Code Book" can be found at http://www.ldc.upenn.edu/Projects/FORM/. Listed there are all the Group, Subgroup, Track, Attribute and Value possibilities.

3 Annotation Graphs

In order to allow for maximum extensibility, FORM uses annotation graphs (AGs) as its logical representation[4]. As described in (Bird and Liberman, 1999), annotation graphs are a formal framework for "representing linguistic annotations of time series data." AGs do this by extracting away from the physical-storage layer, as well as from application-specific formatting, to provide a "logical layer for annotation systems." An annotation graph is a collection arcs and nodes which share a common timeline, that of a video tape, for example. Each node represents a timestamp and each arc represents some linguistic event spanning the time between the nodes. The arcs are labeled with both attributes and values, so that the arc given by the 4-tuple (1,5,Wrist Movement,Side-to-side) represents that there was side-to-side wrist movement between timestamp 1 and timestamp 5. The advantage of using annotation graphs as the logical representation is that it is easy to combine heterogeneous data—as long as they share a common time line. So, if we have a dataset consisting of gesture-arcs, as above, we can easily extend this dataset by adding more arcs representing discourse structure, for example, simply by adding other arcs which have discourse-structure attributes and values. Again, this allows different researchers to use the same linguistic data for many different purposes, while, at the same time, allowing others to explore the correlations between the different phenomena being studied.

4 Preliminary Inter-Annotator Agreement Results

Preliminary results from FORM show that with sufficient training, agreement among the annotators can be very high. Table 2 shows preliminary interannotator agreement results from a FORM pilot study.[5] The results are for two trained annotators for approximately 1.5 minutes of Jan24-09.mov, the video from Figure 1. For this clip, the two annotators agreed that there were at least these 4 gesture excursions. One annotator found 2 additional excursions. Precision refers to the decimal precision of the time stamps given for the beginning and end of gestural components. The *SAME* value means that all time-stamps were given the same value. This was done in order to judge agreement with having to judge the exact beginning and end of an excursion factored out. *Exact* vs. *No-Value* percentage refers to whether both the attributes and values matched exactly or whether just the attributes matched exactly. This distinction is included because a gesture excursion is defined as all movement between two rest positions of the arms and hands. For an excursion, the annotators have to judge both which parts of the arms and hands are salient to the movement (e.g., upper-arm lift and rotation, as well as forearm change in orientation and hand/wrist position) as well as what values to assign (e.g., the upper-arm lifted 15-degrees and rotated 45-degrees). So, the *No-Value%* column captures the degree to which the annotators agree just on the structure of the movement, while *Exact%* measures agreement on both structure and values.

The degree to which inter-annotator agreement varies among these gestures might suggest difficulty in reaching consensus. However, the results on *intra*-annotator agreement studies demonstrate that a single annotator shows similar variance when doing the same video-clip at different times. Table 3 gives the intra-annotator results for one annotator annotating the first 2 gesture excursions of Jan24-09.mov.

For both sets of data, the pattern is the same:

- the less precise the time-stamps, the better the results;

[4] Cf. (Martell, 2002) for a more complete discussion of FORM's use of AGs

[5] Essentially, all the arcs for each annotator are thrown into a bag. Then all the bags are combined and the intersection is extracted. This intersection constitutes the overlap in annotation, i.e., where the annotators agreed. The percentage of the intersection to the whole is then calculated to get the scores presented.

Gesture Excursion	Precision	Exact%	No-Value%
1	2	3.41	4.35
	1	10.07	12.8
	0	29.44	41.38
	SAME	56.92	86.15
2	2	37.5	52.5
	1	60	77.5
	0	75.56	94.81
	SAME	73.24	95.77
3	2	0	0
	1	19.25	27.81
	0	62.5	86.11
	SAME	67.61	95.77
4	2	10.2	12.06
	1	25.68	31.72
	0	57.77	77.67
	SAME	68.29	95.12

Table 1: Inter-Annotator Agreement on Jan24-09.mov

Gesture Excursion	Precision	Exact%	No-Value%
1	0	5.98	7.56
	1	20.52	25.21
	0	58.03	74.64
	SAME	85.52	96.55
2	2	0	0
	1	25.81	28.39
	0	89.06	95.31
	SAME	90.91	93.94

Table 2: Intra-Annotator Agreement on Jan24-09.mov

- *No-Value%* is significantly higher than *Exact%*.

It is also important to note that Gesture Excursion 1 is far more complex than Gesture Excursion 2. And, in both simple and complex gestures, inter-annotator agreement is approaching intra-annotator agreement. Notice, also, that for Excursion 2, inner-annotator agreement is actually better than intra-annotator agreement for the first two rows. This is a result of the difficulty for even the same person over time to precisely pin down the beginning and end of a gesture excursion. Although the preliminary results are very encouraging, all of the above suggests that further research concerning training and how to judge similarity of gestures is necessary. Visual information may need very different similarity criteria.

5 Conclusion: Applications to HLT and HCI?

We plan to augment FORM to include richer paralinguistic information (Head/Torso Movement, Transcription/Syntactic Information, and Intonation/Pitch Information). This will create a corpus that allows for research that heretofore we have been unable to do. It will facilitate experiments that we predict will be useful for speech recognition, as well as other Human-Language Technologies (HLT). As an example of similar research, consider the work of Francis Quek et al. (Quek and others, 2001). They have been able to demonstrate that gestural information is useful in helping with automatic detection of discourse transition. However, their results are limited by the amount of kinematic information they can gather with their video-capture system. Further, an augmented-FORM corpus will contain much more specific data and will allow for more fine-grained analyses than is currently feasible.

Additionally, knowing the relationships among the different facets of human conversation will allow for more informed research in Human-Computer Interaction (HCI). If one of the goals of HCI is to have better immersive-training, then it will be imperative that we understand the subtle connections among the paralinguistic aspects of interaction. A virtual human, for example, would be much better if it were able to understand, and act in accordance with, all of our communicative quirks

Having an extensible corpus such as we describe in this paper is a first-step that will allow many researchers, across many disciplines, to explore these and other useful ideas.

References

Steven Bird and Mark Liberman. 1999. A formal framework for linguistic annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, Pennsylvania. http://citeseer.nj.nec.com/article/bird99formal.html.

Adam Kendon. 2000. Suggestions for a descriptive notation for manual gestures. Unpublished.

Craig Martell. 2002. Form: An extensible, kinematically-based gesture annotation scheme. In *Proceeding of the International Conference on Language Resources and Evaluation.* European Language Resources Association. http://www.ldc.upenn.edu/Projects/FORM.

Francis Quek et al. 2001. Gestural origo and loci-transitions in natural discourse segmentation. Technical Report VISLab-01-12, Department of Computer Science and Engineering, Wright State University. http://vislab.cs.wright.edu/Publications/QueBMH01.html.

The Psychology and Technology of Talking Heads In Human-Machine Interaction

Dominic W. Massaro
University of California
Santa Cruz, CA 95060 U.S.A.
1-831-459-2330
FAX 1-831-459-3519
massaro@fuzzy.ucsc.edu

Abstract

Given the value of visible speech, our persistent goal has been to develop, evaluate, and apply animated agents to produce accurate visible speech. The goal of our recent research has been to increase the number of agents and to improve the accuracy of visible speech. Perceptual tests indicted positive results of this work. Given this technology and the framework of the fuzzy logical model of perception (FLMP), we have developed computer-assisted speech and language tutors for deaf, hard of hearing, and autistic children. Baldi, as the conversational agent, guides students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness. The results indicate that the psychology and technology of Baldi holds great promise in language learning and speech therapy.

Introduction

The face presents visual information during speech that is critically important for effective communication. While the auditory signal alone is adequate for communication, visual information from movements of the lips, tongue and jaws enhance intelligibility of the acoustic stimulus (particularly in noisy environments). Moreover, speech is enriched by the facial expressions, emotions and gestures produced by a speaker (Massaro, 1998). The visual components of speech offer a lifeline to those with severe or profound hearing loss. Even for individuals who hear well, these visible aspects of speech are especially important in noisy environments. For individuals with severe or profound hearing loss, understanding visible speech can make the difference in effectively communicating orally with others or a life of relative isolation from oral society (Trychin, 1997).

Our persistent goal has been to develop, evaluate, and apply animated agents to produce accurate visible speech. These agents have a tremendous potential to benefit virtually all individuals, but especially those with hearing problems (> 28,000,000 in the USA), including the millions of people who acquire age-related hearing loss every year (http://www.nidcd.nih.gov/health/hb.htm), and for whom visible speech takes on increasing importance. One of many applications of animated characters allows the training of individuals with hearing loss to "read" visible speech, and thus facilitate face-to-face oral communication in all situations (educational, social, work-related, etc). These enhanced characters can also function effectively as language tutors, reading tutors, or personal agents in human machine interaction.

For the past ten years, my colleagues and I have been improving the accuracy of visible speech produced by an animated talking face - Baldi (Massaro, 1998, chapters 12-14). Baldi has been used

effectively to teach vocabulary to profoundly deaf children at Tucker-Maxon Oral School in a project funded by an NSF Challenge Grant (Barker, 2002; Massaro et al., 2000). The same pedagogy and technology has been employed for language learning with autistic children (Massaro & Bosseler, 2002). While Baldi's visible speech and tongue model probably represent the best of the state of the art in real-time visible speech synthesis by a talking face, experiments have shown that Baldi's visible speech is not as effective as human faces. Preliminary observations strongly suggest that the specific segmental and prosodic characteristics are not defined optimally. One of our goals, therefore, is to significantly improve the communicative effectiveness of synthetic visual speech.

1 Facial Animation and Visible Speech Synthesis

Visible speech synthesis is a sub-field of the general areas of speech synthesis and computer facial animation (Chapter 12, Massaro, 1998, organizes the representative work that has been done in this area). The goal of the visible speech synthesis in the Perceptual Science Laboratory (PSL) has been to develop a polygon (wireframe) model with realistic motions (but not to duplicate the musculature of the face to control this mask). We call this technique terminal analogue synthesis because its goal is to simply use the final speech product to control the facial articulation of speech (rather than illustrate the physiological mechanisms that produce it). This method of rendering visible speech synthesis has also proven most successful with audible speech synthesis. One advantage of the terminal analogue synthesis is that calculations of the changing surface shapes in the polygon models can be carried out much faster than those for muscle and tissue simulations. For example, our software can generate a talking face in real time on a commodity PC, whereas muscle and tissue simulations are usually too computationally intensive to perform in real time (Massaro, 1998). More recently, image synthesis, which joins together images of a real speaker, has been gaining in popularity because of the realism that it provides. These systems also are not capable of real-time synthesis because of their computational intensity.

Our own current software (Cohen & Massaro, 1993; Cohen et al., 1996; Cohen et al., 1998; Massaro, 1998) is a descendant of Parke's software and his particular 3-D talking head (Parke, 1975). Our modifications over the last 6 years have included increased resolution of the model, additional and modified control parameters, three generations of a tongue (which was lacking in Parke's model), a new visual speech synthesis coarticulatory control strategy, controls for paralinguistic information and affect in the face, alignment with natural speech, text-to-speech synthesis, and bimodal (auditory/visual) synthesis. Most of our current parameters move vertices (and the polygons formed from these vertices) on the face by geometric functions such as rotation (e.g. jaw rotation) or translation of the vertices in one or more dimensions (e.g., lower and upper lip height, mouth widening). Other parameters work by scaling and interpolating between two different face subareas. Many of the face shape parameters--such as cheek, neck, or forehead shape, and also some affect parameters such as smiling--use interpolation. Our animated talking face, Baldi, can be seen at: http://mambo.ucsc.edu.

We have used phonemes as the basic unit of speech synthesis. In this scheme, any utterance can be represented as a string of successive phonemes, and each phoneme is represented as a set of target values for the control parameters such as jaw rotation, mouth width, etc. Because speech production is a continuous process involving movements of different articulators (e.g., tongue, lips, jaw) having mass and inertia, phoneme utterances are influenced by the context in which they occur by a process called coarticulation. In our visual speech synthesis algorithm

(Cohen & Massaro, 1993; Massaro, 1998, chapter 12), coarticulation is based on a model of speech production using rules that describe the relative dominance of the characteristics of the speech segments. In our model, each segment is specified by a target value for each facial control parameter. For each control parameter of a speech segment, there are also temporal dominance functions dictating the influence of that segment over the control parameter. These dominance functions determine independently for each control parameter how much weight its target value carries against those of neighboring segments, which will in turn determine the final control values.

Baldi's synthetic tongue is constructed of a polygon surface defined by sagittal and coronal b-spline curves. The control points of these b-spline curves are controlled singly and in pairs by speech articulation control parameters. There are now 9 sagittal and 3 * 7 coronal parameters that are modified to mimic natural tongue movements. The tongue, teeth, and palate interactions during speaking require an algorithm to prevent the tongue from going into rather than colliding with the teeth and palate. To ensure this, we have developed a fast collision detection method to instantiate the appropriate interactions. Two sets of observations of real talkers have been used to inform the appropriate movements of the tongue. These include 1) three dimensional ultrasound measurements of upper tongue surfaces and 2) EPG data collected from a natural talker using a plastic palate insert that incorporates a grid of about a hundred electrodes that detect contact between the tongue and palate at a fast rate (e.g. a full set of measurements 100 times per second). These measurements were made in collaboration with Maureen Stone at John Hopkins University. Minimization and optimization routines are used to create animated tongue movements that mimic the observed tongue movements (Cohen et al., 1998).

2 Recent Progress in Visible Speech Synthesis

Important goals for the application of talking heads are to have a large gallery of possible agents and to have highly intelligible and realistic synthetic visible speech. Our development of visible speech synthesis is based on facial animation of a single canonical face, called Baldi (see Figure 1; Massaro, 1998).

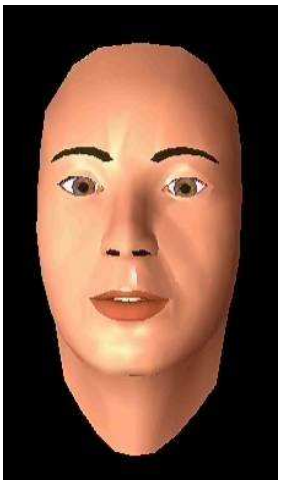


Figure 1. Picture of Baldi, our computed animated talking head.

Although the synthesis, parameter control, coarticulation scheme, and rendering engine are specific to Baldi, we have developed software to reshape our canonical face to match various target facial models. To achieve realistic and accurate synthesis, we use measurements of facial, lip, and tongue movements during speech production to optimize both the static and dynamic accuracy of the visible speech. This optimization process is called minimization because we seek to minimize the error between the empirical observations of real human speech and the speech produced by our synthetic talker (Cohen, Beskow, & Massaro, 1998; Cohen, Clark, & Massaro, 2001; Cohen, Clark, & Massaro, 2002).

2.1 Improving the Static Model

A Cyberware 3D laser scanning system is used to enroll new citizens in our gallery of talking heads. A laser scan of a new target head produces a very high polygon count representation. Rather than trying to animate this high-resolution head (which is impossible to do in real-time with current hardware), our software uses these data to reshape our canonical head to take on the shape of the new target head. In this approach, facial landmarks on both the laser scan head and the generic Baldi head are marked by a human operator. Our canonical head is then warped until it assumes as closely as possible the shape of the target head, with the additional constraint that the landmarks of the canonical face move to positions corresponding to those on the target face.

2.1.1 Improving the Dynamic Model

To improve the intelligibility of our talking heads, we have developed software for using dynamic 3D optical measurements (Optotrak) of points on a real face while talking. In one study, we recorded a large speech database with 19 markers affixed to the face of DWM at important locations (see Figure 2).

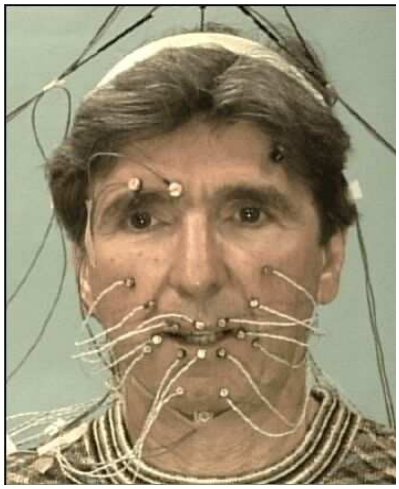


Figure 2. Frame from video used in the recording of the data base and in the evaluation.

Fitting of these dynamic data occurred in several stages. To begin, we assigned points on the surface of the synthetic model that best correspond to the Optotrak measurement points. In the training, the Optotrak data were adjusted in rotation, translation, and scale to best match the corresponding points marked on the synthetic face.

The data collected for the training consisted of 100 CID sentences recorded by DWM speaking in a fairly natural manner. In the first stage fit, for each time frame (30 fps) we automatically and iteratively adjusted 11 facial control parameters of the face to get the best fit (the least sum of squared distances) between the Optotrak measurements and the corresponding point locations on the synthetic face. In the second stage fit, the goal was to tune the segment definitions (parameter targets, dominance function strengths, attack and decay rates, and peak strength time offsets) used in our coarticulation algorithm (Cohen & Massaro, 1993) to get the best fit with the parameter tracks obtained in the first stage fit. We first used Viterbi alignment on the acoustic speech data of each sentence to obtain the phoneme durations used to synthesize each sentence. Given the phonemes and durations, we used our standard parametric phoneme synthesis and coarticulation algorithm to synthesize the parameter tracks for all 100 CID sentences. These were compared with the parameter tracks obtained from the first stage fit, the error computed, and the parameters adjusted until the best fit was achieved.

3 Perceptual Evaluation

We carried out a perceptual recognition experiment with human subjects to evaluate how well this improved synthetic talker conveyed speech information relative to the real talker. To do this we presented the 100 CID sentences in three conditions: auditory alone, auditory + synthetic talker, and auditory + real talker. In all cases there was white (speech band) noise added to the audio channel. Each of the 100 CID

sentences was presented in each of the three modalities for a total of 300 trials. Each trial began with the presentation of the sentence, and subjects then typed in as many words as the could recognize. Students in an introductory psychology course served as subjects.

Figure 3 shows the proportion of correct words reported as a function of the initial consonant under the three presentation conditions. There was a significant advantage of having the visible speech, and the advantage of the synthetic head was equivalent to the original video. Overall, the proportion of correctly reported words for the three conditions was 0.22 auditory, 0.43 synthetic face, and 0.42 with the real face.

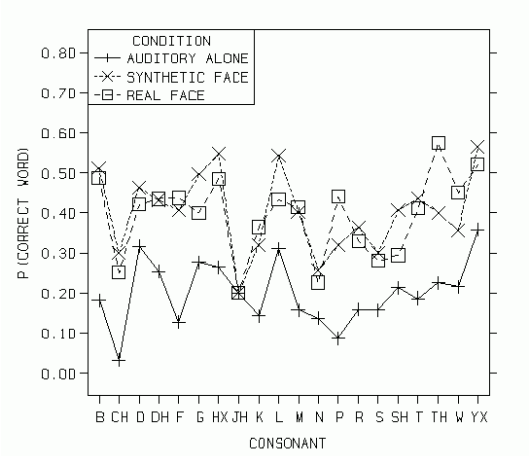


Figure 3. Proportion words correct as a function of initial consonant of all words in the test sentences for auditory alone, synthetic and real face conditions.

The results of the current evaluation study, using the stage 1 best fitting parameters is encouraging. In studies to follow, we'll be comparing performance with visual TTS synthesis based on the segment definitions from the stage 2 fits, both for single segments, context sensitive segments, and also using concatenation of diphone sized chunks from the stage 1 fits. In addition, we will be using a higher resolution canonical head with many additional polygons and an improved texture map.

4 Early History of Speech Science

Speech science evolved as the study of a unimodal phenomenon. Speech was viewed as a solely auditory event, as captured by the seminal speech-chain illustration of Denes and Pinson (1963). This view is no longer viable as witnessed by a burgeoning record of research findings. Speech as a multimodal phenomenon is supported by experiments indicating that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech. Many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves intelligibility in these situations. Visible speech also is an important communication channel for individuals with hearing loss and others with specific deficits in processing auditory information.

We have seen that the number of words understood from a degraded auditory message can often be doubled by pairing the message with visible speech from the talker's face. The combination of auditory and visual speech has been called super-additive because their combination can lead to accuracy that is much greater than accuracy on either modality alone. Our participants, for example, would have performed very poorly given just the visual speech alone. Furthermore, the strong influence of visible speech is not limited to situations with degraded auditory input. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. For example, if the ambiguous auditory sentence, *My bab pop me poo brive*, is paired with the visible sentence, *My gag kok me koo grive*, the perceiver is likely to hear, *My dad taught me to drive*. Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro, 1998).

There are several reasons why the use of auditory and visual information

together is so successful. These include a) robustness of visual speech, b) complementarity of auditory and visual speech, and c) optimal integration of these two sources of information. Speechreading, or the ability to obtain speech information from the face, is robust in that perceivers are fairly good at speech reading even when they are not looking directly at the talker's lips. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro, 1998, Chapter 14).

Complementarity of auditory and visual information simply means that one of the sources is strong when the other is weak. A distinction between two segments robustly conveyed in one modality is relatively ambiguous in the other modality. For example, the place difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the voicing difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. Two complementary sources of information make their combined use much more informative than would be the case if the two sources were non-complementary, or redundant (Massaro, 1998, pp. 424-427).

The final reason is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner. There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion in which both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from each modality to perform as efficiently as possible. Many different empirical results have been accurately predicted by a model that describes an optimally efficient process of combination (Massaro, 1998). We now describe this model.

5 Fuzzy Logical Model of Perception

The fuzzy logical model of perception (FLMP), shown in Figure 4, assumes necessarily successive but overlapping stages of processing. The perceiver of speech is viewed as having multiple sources of information supporting the identification and interpretation of the language input. The model assumes that 1) each source of information is evaluated to give the continuous degree to which that source supports various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives (Massaro et al., 2001, in press, a, b).

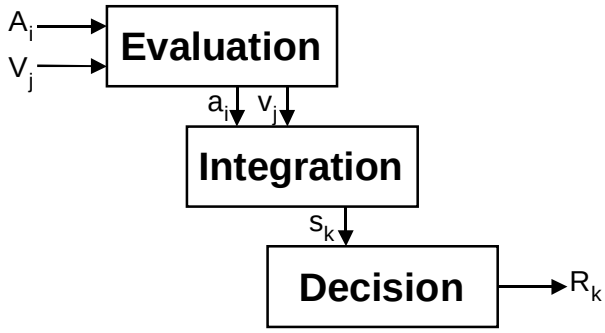


Figure 4. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, s_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

The paradigm that we have developed permits us to determine how visible speech is processed and integrated with other sources of information. The results also inform us about which of the many potentially functional cues are actually used by human observers (Massaro, 1987, Chapter 1). The systematic variation of properties of the speech signal combined with the quantitative test of models of speech perception enables the investigator to test the psychological validity of different cues. This paradigm has already proven to be effective in the study of audible, visible, and bimodal speech perception (Massaro, 1987, 1998). Thus, our research strategy not only addresses how different sources of information are evaluated and integrated, but can uncover what sources of information are actually used. We believe that the research paradigm confronts both the important psychophysical question of the nature of information and the process question of how the information is transformed and mapped into behavior. Many independent tests point to the viability of the FLMP as a general description of pattern recognition. The FLMP is centered around a universal law of how people integrate multiple sources of information. This law and its relationship to other laws is developed in detail in Massaro (1998). The FLMP is also valuable because it motivates our approach to language learning.

Baldi can display a midsagital view, or the skin on the face can be made transparent to reveal the internal articulators. The orientation of the face can be changed to display different viewpoints while speaking, such as a side view, or a view from the back of the head (Massaro 1999, 2000). The auditory and visual speech can also be independently controlled and manipulated, permitting customized enhancements of the informative characteristics of speech. These features offer novel approaches in language training, permitting one to pedagogically illustrate appropriate articulations that are usually hidden by the face. This technology has the potential to help individuals with language delays and deficits, and we have been utilizing Baldi to carry out language tutoring with deaf children and children with autism.

6 Language Learning

As with most issues in social science, there is no consensus on the best way to teach or to learn language. There are important areas of agreement, however. One is the central importance of vocabulary knowledge for understanding the world and for language competence in both spoken language and in reading. There is empirical evidence that very young children more easily form conceptual categories when category labels are available than when they are not (Waxman & Kosowski, 1990). There is also evidence that there is a sudden increase in the rate at which new words are learned once the child knows about 150 words. Grammatical skill also emerges at this time (Marchman & Bates, 1994). Even children experiencing language delays because of specific language impairment benefit once this level of word knowledge is obtained. It follows that increasing the pervasiveness and effectiveness of vocabulary learning offers a huge opportunity for improving conceptual knowledge and language competence for all individuals, whether or not they are disadvantaged because of sensory limitations, learning disabilities, or social condition. Finally, it is well-known that vocabulary knowledge is positively correlated with both listening and reading comprehension (Anderson & Freebody, 1981).

Another area of agreement is the importance of time on task; learning and retention are positively correlated with the time spent learning. Our technology offers a platform for unlimited instruction, which can be initiated when and wherever the child and/or supervisor chooses. Baldi and the accompanying lessons are perpetual. Take, for example, children with autism, who have irregular sleep patterns. A child could conceivably wake in the middle of the night

and participate in language learning with Baldi as his or her friendly guide.

Several advantages of utilizing a computer-animated agent as a language tutor are clear, including the popularity of computers and embodied conversational agents with children with autism. A second advantage is the availability of the program. Instruction is always available to the child, 24 hours a day 365 days a year. Furthermore, instruction occurs in a one-on-one learning environment for the students. We have found that the students enjoy working with Baldi because he offers extreme patience, he doesn't become angry, tired, or bored, and he is in effect a perpetual teaching machine.

Our Language Tutor, Baldi, encompasses and instantiates the developments in the pedagogy of how language is learned, remembered and used. Education research has shown that children can be taught new word meanings by using drill and practice methods (e.g., McKeown et al., 1986; Stahl, 1983). It has also been convincing demonstrated that direct teaching of vocabulary by computer software is possible, and that an interactive multimedia environment is ideally suited for this learning (Wood, 2001). As cogently observed by Wood (2001), "Products that emphasize multimodal learning, often by combining many of the features discussed above, perhaps make the greatest contribution to dynamic vocabulary learning. Mulitimodal features not only help keep children actively engaged in their own learning, but also accommodate a range of learning styles by offering several entry points: When children can see new words in context, hear them pronounced, type them into a journal, and cut and paste an accompanying illustration (or create their own), the potential for learning can be dramatically increased." Following this logic, many aspects of our lessons enhance and reinforce learning. For example, the existing program and planned modifications make it possible for the student to 1) Observe the words being spoken by a realistic talking interlocutor (Baldi), 2) See the word as written as well as spoken, 3) See visual images of referents of the words or view an animation of a meaningful scene, 4) Click on or point to the referent, 5) Hear himself or herself say the word, 6) Spell the word by typing, observe the word used in context, and 7) Incorporate the word into his or her own speech act.

Other benefits of our program include the ability to seamlessly meld spoken and written language, provide a semblance of a game-playing experience while actually learning, and to lead the child along a growth path that always bridges his or her current "zone of proximal development."

6.1 Description of Vocabulary Wizard and Player

The Vocabulary Wizard is a set of formatted programs permitting authoring abilities to create vocabulary training in a language tutorial program. The wizard interface incorporates Baldi, synthesized speech, and images of the vocabulary items. The visual images were imported to create the vocabulary-training and program in which parts of the visual image were associated with spoken words or phrases. Figure 5 shows a view of the screen in a prototypical application.

In this application, the students learn to identify prepositions such as inside, next to, in front of, etc. Baldi asks the student to "click on the bear inside of the box". An outlined region in orange designates the selected region. The faces in the left-hand corner of the figure are the "stickers", which show a happy or a sad face as feedback for correct and incorrect responses. Processing information presented via the visual modality reinforces learning (Courchesne, et al. 1994) and is consistent with the TEEACH (Schopler et al., 1995) suggestion for visually presented material for educating children with autism.

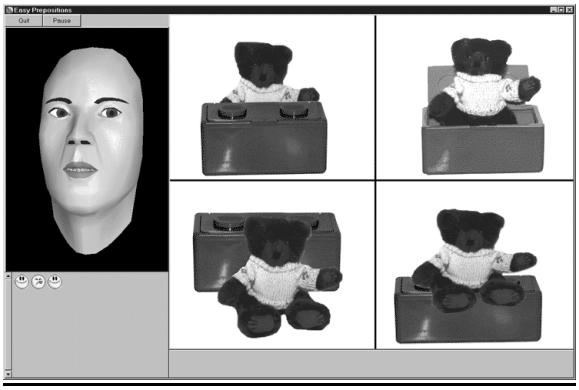


Figure 5. A prototypical Vocabulary Wizard illustrating the format of the tutors. Each application contains Baldi, the vocabulary items and written text and captioning (optional), and "stickers". In this application the students learn to identify prepositions. For example, Baldi says "show me the bear inside of the box". The student clicks on the appropriate region and visual feedback in the form of stickers (the happy and sad faces) are given for each response

All of the exercises required the children to respond to spoken directives such as "click on the little chair", or "find the red fox". These images were associated with the corresponding spoken vocabulary words (see appendix for vocabulary examples). The items became highlighted whenever the mouse passed over that region. The student selected his or her response by clicking the mouse on one of the designated areas.

The Vocabulary Wizard consists of 5 application modules. These modules are pretest, presentation, perception practice, production, and post-test. The Wizard is equipped with easily changeable default settings that determine what Baldi says and how he says it, the feedback given for responses, the number of attempts permitted for the student per section, and the number of times each item is presented. The program automatically creates and writes all student performance information to a log file stored in the student's directory.

6.1.1 Research on the educational impact of animated tutors:

Research has shown that this pedagogical and technological program is highly effective for both children with hearing loss and children with autism. These children tend to have major difficulties in acquiring language, and they serve as particularly challenging tests for the effectiveness of our pedagogy. There are recent research reports on the positive results of employing our animated tutor to teach both children with hearing loss (Barker, 2002) and children with autism (Bosseler & Massaro, 2002).

Improving the vocabulary of hard of hearing children

It is well-known that hard of hearing children have significant deficits in vocabulary knowledge. In many cases, the children do not have names for specific things and concepts. These children often communicate with phrases such as "the window in the front of the car," "the big shelf where the sink is," or "the step by the street" rather than "windshield," "counter," or "curb" (Barker, 2002, citing Pat Stone). The vocabulary player has been in use at the Tucker Maxon Oral School in Portland, Oregon, and Barker (in press) evaluated its effectiveness. Students were given cameras to photograph objects at home and surroundings. The pictures of these objects were then incorporated as items in the lessons. A given lesson had between 10 and 15 items. Students worked on the items about 10 minutes a day until they reached 100% on the posttest. They then moved on to another lesson. About one month after each successful (100%) posttest, they were retested on the same items. Ten girls and nine boys the "upper school" and the "lower school" participated in the applications. There were six deaf children and one hearing child between 8 and 10 years of age in the lower school. Ten deaf and two hearing children, between 11 and 14 years of age, participated from the upper school.

Figure 6 gives the results of these lessons for the children. The results are given for three stages of the study: Pretest, Posttest, and Retention after 30 days. The

items were classified as known, not known, and learned. Known items are those that the children already knew on the initial pretest before the first lesson. Not known items are those that the children did not know, as evidenced by their inability to identify these items in the initial pretest. Learned items are those that the children identified correctly in the posttest. Similar results were found for the younger age group. Students knew about one-half of the items without any learning, they successfully learned the other half of the items, and retained about one-half of the newly learned items when retested 30 days later. These results demonstrate the effectiveness of the language player for learning and retaining new vocabulary.

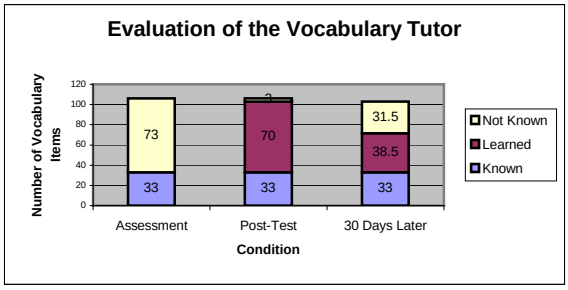


Figure 6. Results of word learning at the Tucker-Maxon Oral School using the vocabulary Wizard/Tutor. The results give the average number of words that were already known, the average number learned using the program, and the average number retained after 30 days. This outcome indicates significant vocabulary learning, with about 55% retention of new words after 30 days. Results from Barker (2002).

6.1.1.1 Improving the vocabulary of children with autism

Autism is a spectrum disorder characterized by a variety of characteristics, which usually include perceptual, cognitive, and social differences. Among the defining characteristics of autism, the limited ability to produce and comprehend spoken language is the most common factor leading to diagnosis (American Psychiatric Association, 1994). The language and communicative deficits extend across a broad range of expression (Tager-Flusberg, 1999). Individual variations occur in the degree to which these children develop the fundamental lexical, semantic, syntactic, phonological, and pragmatic components of language including those who fail to develop one or more of these elements of language comprehension and production.

Approximately one-half of the autistic population fails to develop any form of functional language (Tager-Flusberg, 2000). Within the population that does develop language, the onset and rate at which the children pass through linguistic milestones are often delayed compared to non-autistic children (e.g. no single words by age 2 years, no communicative phrases by age 3) (American Psychiatric Association, 1994). The ability to label objects is often severely delayed in this population as well as the deviant use and knowledge of verbs and adjectives. Van Lancker et. al. (1991) investigated the abilities of autistic and schizophrenic children to identify concrete nouns, non-emotional adjectives, and emotional adjectives. The results showed that the performance of children with autism was below controls in all three areas.

Despite the prevalence of language delays in autistic individuals, formalized research has been limited, partly due to the social challenges inherent in this population (Tager-Flusberg, 2000). Intervention programs for children with autism typically emphasize developing speech and communication skills (e.g. TEEACH, Applied Behavioral Analysis). These programs most often focus on the fundamental lexical, semantic, syntactic, phonological, and pragmatic components of language. The behavioral difficulties speech therapists and instructors encounter, such as lack of cooperation, aggression, and lack of motivation to communicate, create difficult situations that are not optimal for learning. Thus, creating motivational environments necessary to develop these language skills introduces many inherent obstacles (Tager-Flusberg, 2000).

In this study (Bosseler & Massaro, 2002), the Tutors were constructed and run

on a 600 MHz PC with 128 MB RAM hard drive running Microsoft Windows NT 4 with a Gforce 256 AGP-V6800 DDR graphics board. The tutorials were presented on a Graphic Series view Sonic 20" monitor. All students wore a Plantronics PC Headset model SR1. Students completed 2 sessions a week, a minimum of 2 lessons per session, and an average of 3, and sometimes as many as 8. The sessions lasted between 10 and 40 minutes. A total of 559 different vocabulary items were selected from the curriculum of both schools for a total of over 84 unique vocabulary lessons.

A series of observations by the experimenter (AB) during the course of each lesson led to many changes in the program, including the use of headsets, isolating the student from the rest of the class and removal of negative verbal feedback from Baldi (such as, "No, (user) that's not right". The students appeared to enjoy working with Baldi. We documented the children saying such things as "Hi Baldi" and "I love you Baldi". The stickers generated for correct (happy face) and incorrect (sad face) responses proved to be an effective way to provide feedback for the children, although some students displayed frustration when he or she received more than one sad face. The children responded to the happy faces by saying such things like "Look, I got them all right", or laughing when a happy face appeared. We also observed the students providing verbal praise to themselves such as "Good job", or prompting the experimenter to say "Good job" after every response. For the autistic children, several hundred vocabulary tutors were constructed, consisting of various vocabulary items selected from the curriculum of two schools. The children were administered the tutorial lessons until 100% accuracy was attained on the posttest module. Once 100% accuracy was attained on the final posttest module, the child did not see these lessons again until reassessment approximately 30 days later.

Figure 7 shows that the children learned many new words, grammatical constructions, and concepts, proving that the language tutors are a valuable learning environment for these children.

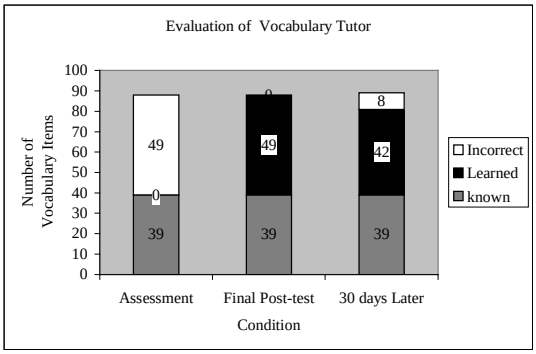


Figure 7. The mean observed proportion of correct identifications for the initial assessment, final posttest and reassessment for each of the seven students. Student 8 was omitted form this analysis because he left the program before we began reassessment. The results reveal these seven students were able to accurately identify significantly more words during the reassessment than the initial assessment.

In order to assess how well the children would retain the vocabulary items that were learned during the tutorial lesson, we administered the assessment test to the student at least 30 days following the final posttest. As can be seen in Figure 8, the students were able to recall 85% of the newly-learned vocabulary items at least 30 days following training.

Although all of the children demonstrated learning from initial assessment to final reassessment, the children might have been learning the words outside of our program, for example, from speech therapists, at home, or in their school curriculum. Furthermore, we questioned whether the vocabulary knowledge would generalize to new pictorial instances of the words. To address these issues we conducted a second experiment. Corroborating with the children's instructors and speech therapists, we gathered an assortment of vocabulary words that the children supposedly did not know. We used these words in the Horner and Baer (1978) single subject multiple probe design. We

randomly separated the words to be trained into three sets, established individual pre-training performance for each set of vocabulary items, and trained on the first set of words while probing performance for both the trained and untrained sets of words.

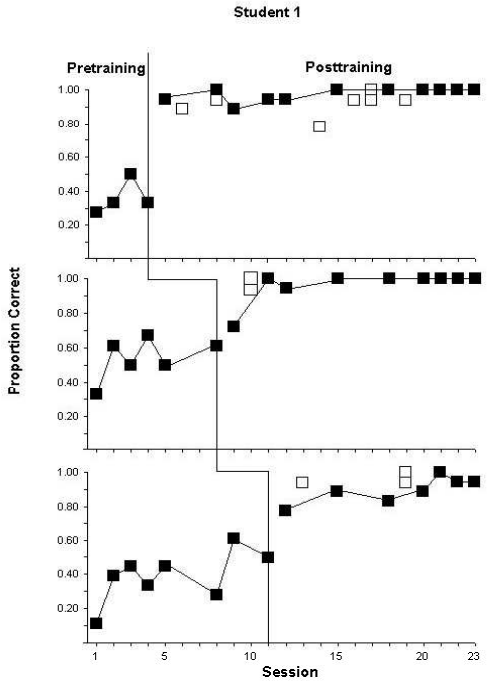


Figure 8. Proportion correct during the Pretraining, Posttraining, and Generalization for one of the six students. The vertical lines separate the Pretraining and Postraining conditions. Generalization results are given by the open squares. See text for additional details.

Once the student was able to attain 100% identification accuracy during a training session, a generalization probe to new instances of the vocabulary images was initiated. If the child did not meet the criterion, he or she was trained on these new images. Generalization training continued until the criterion was met, at which time training began on the next set of words. Probe tests continued on the original learned set of words and images until the end of the study. We continued this procedure until the student completed training on all three sets of words. Our goal was to observe a significant increase in identification accuracy during the post-training sessions relative to the pre-training sessions.

Figure 9 displays the proportion of correct responses for a typical student during the probe sessions conducted at pre-training and post-training for each of the three word sets. The vertical lines in each of the three panels indicates the last pre-training session before the onset of training. Some of the words were clearly known prior to training, and were even learned to some degree without training. As can be seen in the figure, however, training was necessary for substantial learning to occur. In addition, the children were able to generalize accurate identification to four instances of untrained images.

The goal of these investigations was to evaluate the potential of using a computer-animated talking tutor for children with language delays. The results showed a significant gain in vocabulary. We also found that the students were able to recall much of the new vocabulary when reassessed 30 days after learning. Followup research showed that the learning is indeed occurring from the computer program and vocabulary knowledge can transfer to novel images.

We believe that the children in our investigation profited from having the face and that seeing and hearing spoken language can better guide language learning than either modality alone. A direct test of this hypothesis would involve comparing learning with and without the face. Baldi can actually provide more information than a natural face. He can be programmed to display a midsagital view, or the skin on the face can be made transparent to reveal the internal articulators. The orientation of the face can be changed to display different viewpoints while speaking, such as a side view, or a view from the back of the head (Massaro, 1999). The auditory and visual speech can also be independently controlled and manipulated, permitting customized enhancements of the informative characteristics of speech. These features offer novel approaches in language training, permitting one to pedagogically illustrate appropriate articulations that are usually hidden by the face. More generally,

additional research should investigate whether the influence of several modalities on language processing provide a productive approach to language learning.

Acknowledgements

The research and writing of the paper were supported by grants from National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), Intel Corporation, the University of California Digital Media Program, the Cure Autism Now Foundation, and the University of California, Santa Cruz.

References

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders, DSM-IV* (4th ed.). Washington, DC: Author.

Barker, L., J (2002). Computer-Assisted Vocabulary Acquisition: The CSLU Vocabulary Tutor in Oral-Deaf Education. Journal of Deaf Studies and Deaf Education (in press).

Bosseler, A., & Massaro, D.W. (submitted). Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning in Children with Autism. *Journal of Autism and Developmental Disorders*, submitted.

Cohen, M. M. & Massaro, D. W. (1993) Modeling coarticulation in synthetic visual speech. In M. Thalmann & D. Thalmann (Eds.) *Computer Animation '93*. Tokyo: Springer-Verlag. http://mambo.ucsc.edu/psl/ca93.html

Cohen, M.M., Beskow, J., & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. In D. Burnham, J. Robert-Ribes, & E. Vatikiotis-Bateson (Eds.) Proceedings of Auditory Visual Speech Perception '98. (pp. 201-206). Terrigal-Sydney Australia, December, 1998. *AVSP '98* (December 4-6, 1998, Sydney, Australia).

Cohen, M.M., Clark, R., & Massaro, D.W. (2001). Animated speech: Research progress and applications. In D.W. Massaro, J. Light, K. Geraci (Eds.) *AVSP2001, Proceedings of Auditory-Visual Speech Processing, AVSP2001*, Santa Cruz, CA: Perceptual Science Laboratory, p. 201. *AVSP 200, (September 7-9, 2001, Aalborg, Denmark*).

Cohen, M.M., Clark, R., & Massaro, D.W. (2002). Training a talking head. Paper submitted to *the fourth International Conference on Multimodal Interfaces (ICMI'2002)*, Pittsburgh, 14-16 October 2002.

Cohen, M. M., Walker, R. L., & Massaro, D. W. (1996). Perception of synthetic visual speech. In D. G. Stork & M. E. Hennecke (eds.), *Speechreading by humans and machines* (pp. 153-168). New York: Springer.

Courchesne, E., Townsend, J., Ashoomoff, N.A., Yeung-Courchesne, R., Press, G., Murakami, J., Lincoln, A., James, H., Saitoh, O., Haas, R., & Schreibman, L. (1994). A new finding in autism: Impairment in shifting attention. In S. H. Broman & J. Grafman (Eds.), *Atypical cognitive deficits in developmental disorders: Implications for brain function* (pp. 101-137). Hillsdale NJ: Lawrence. Erlbaum.

Denes, P. B., & Pinson, E. N. (1963). The speech chain. *The physics and biology of spoken language.* New York: Bell Telephone Laboratories.

Massaro, D. W. (1987). *Speech perception by ear and eye: A Paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press.

Massaro, D.W. (1999). From theory to practice: Rewards and challenges. In Proceedings of the International Conference of Phonetic Sciences, San Francisco, CA, August.

Massaro, D.W. (2000). From "Speech is Special" to Talking Heads in Language Learning. In proceedings of *Integrating speech technology in the (language)*

learning and assistive interface, (InSTIL 2000) August 29-30.

Massaro, D. W. (in press a). *Speech Perception and Recognition* (Article 085) for the Encyclopedia of Cognitive Science.

Massaro, D. W. (in press b). Multimodal Speech Perception: A Paradigm for Speech Science In B. Granstrom, D. House, & I. Karlsson (Eds.) *Multilmodality in language and speech systems*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Massaro, D. W.; & Bosseler, A. (2002). A computer-animated tutor for vocabulary and language learning in children with autism. Paper submitted to the *7th International Conference on Spoken Language Processing*, 2002, Denver, Colorado, September 16-20.

Massaro, D. W.; Cohen, M. M.; Campbell, C. S.; Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review, 8 (1)*: p. 1-17.

McKeown, M., Beck, I., Omanson, R., & Pople, M. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20,* 522-535.

Parke, F. I. (1975). A model for human faces that allows speech synchronized animation. *Computers and Graphics Journal, 1*, 1-4.

Schopler, E., Mizibov, G. B., & Hearsey, K (1995). Structured teaching in the TEACCH system. In. E. Schopler & Mesibov (Eds.), *Learning and cognition in autism. Current issues in autism* (243-268). New York: Plenum Press.

Stahl, S. A. (1986). Three principals of effective vocabulary instruction. *Journal of Reading, 29*, 662-668.

Tager-Flusberg, H. (1999). A psychological approach to understanding the social and language impairments in autism. *International Review of Psychiatry*, 11, 355-334.

Tager-Flusberg, H (2000). Language development in children with autism. In L. Menn & N. Bernstein Ratner (Ed.), *Methods For Studying Language Production* (pp., 313-332). New Jersey: Mahwah.

Van Lancker, D., Cornelius, C., Needleman, R. (1991). Comprehension of Verbal Terms for Emotions in Normal, Autistic, and Schizophrenic Children. *Developmental Neuropsychology*, 7, 1-18.

Wood, J. (2001). Can software support children's vocabulary development? *Language Learning & Technology, 5*, 166-201

Creating multmodal, multilevel annotated corpora with TASX

Jan-Torsten Milde
Department of Linguistics and Literary Studies
University of Bielefeld, Germany
milde@coli.uni-bielefeld.de

Abstract

The paper describes the design and implementation of an XML-based corpus environment for multilevel annotated multimodal (language) data. The TASX-environment (TASX: Time Aligned Signal data eXchange format) constitutes a technical basis for all aspects of the corpus setup procedure: XML-based annotation of the multimodal data, transformation of non XML-annotations, and the web-based analysis and dissemination of the data.

1 Introduction

In this paper we describe ongoing research in the design and implementation of an XML-based corpus environment for complex annotated multimodal data.

The development of the corpus environment complements the LeaP[1] project, which explores the acquisition of prosody by both second language learners of German and English. In a period of two years a large set of audio and video recordings of second language learners' speech will be made and phonologically annotated. In addition, the form and function of gestures in non-native speech are analysed. It is hypotheized that transfer and interference from the native language as well as an adapted variability and frequency of gestures will occur. The alignment of gestures with prosodic features will be explored.

[1] http://www.spectrum.uni-bielefeld.de/LeaP/

From the collected data an XML-annotated multimodal corpus will be set up. The implementation model of the corpus engine is based on a client/server approach. For performance reasons the XML-annotated data can be stored in a relational database. The XSL-T-based transformation of the data is a server sided process. The TASX-environment presented here supports the complete corpus setup procedure: XML-based annotation of raw speech and video data, the transformation of non XML-data and the analysis and dissemination of the corpus.

The paper is organized in four sections. The underlying XML-based TASX format is explained and the components of the TASX-environment will be described in more detail. Finally, a short conslusion will be given.

2 The TASX format

A central aspect of our research is to explore up to which point current standard XML technology (XML, XSL-T, XSL-FO, XPATH, SVG, XQUERY) can be used to model multimodal corpora, to transform, query and distribute the content of such corpora and to perform adequate linguistic analysis. As a result, all linguistic data in our system is stored in an XML-based format called TASX: the *T*ime *A*ligned *S*ignal data e*X*change format.

A TASX-annotated corpus consists of a set of sessions, each one holding an arbitrary number of descriptive tiers, called layers. Each layer consists of a set of separated events. Each event stores some textual information (e.g. a syllable or a handform) and is linked to the primary audio data by two time stamps denoting the interval of

this event. Relations between events on different tiers can be encoded by defining links using the ID/IDREFS mechanism of XML. This is similar to the approach of stand-off markup as proposed by MATE (Dybkaer et al., June 1999), respectivly NITE (Carletta et al., 2002).

Finally, arbitrary meta-data can be assigned to the complete corpus, each session, each layer and each event. It might be sensible to extend the meta data description in a way that tree structured data can immediately be described by XML annotations. Currently we rather use the simpler version with linear structure. The following DTD fragment formalizes the TASX format:

```
<!-- corpus data -->
<!ELEMENT tasx (meta*,session+)>

<!ELEMENT session (meta*,layer+)>
<!ELEMENT layer (meta*,event+)>
<!ELEMENT event (#PCDATA,meta*)>

<!-- meta data -->
<!ELEMENT meta (desc*)>
<!ELEMENT desc (name,val)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT val (#PCDATA)>

<!-- atributes -->
<!ATTLIST session
        s-id CDATA #REQUIRED
       day CDATA #REQUIRED
        ref IDREF #IMPLIED
        month CDATA #REQUIRED
        year CDATA #REQUIRED>

<!ATTLIST layer
       l-id CDATA #REQUIRED
       ref IDREF #IMPLIED>

<!ATTLIST event
          e-id CDATA #REQUIRED
          start CDATA #REQUIRED
          end CDATA #REQUIRED
          ref IDREF #IMPLIED
          mid CDATA #IMPLIED
          len CDATA #IMPLIED>

<!ATTLIST meta
        m-id CDATA #REQUIRED
        ref IDREF #IMPLIED
        access CDATA #IMPLIED
        level CDATA #IMPLIED>
```

Despite of its simplicity, the TASX-format is powerful enough to encode most of the corpus annotation formats currently in use. Indeed a number of format transformation programms have been implemented. For example, in order to reconstruct the equivalent annotation graphs (Bird and Liberman, 1999) representation of a TASX annotated corpus, one only has to collect the time stamps encoded in the start and end attributes of the event tags, sort them and then produce the timeline. Finally the time stamps of the events have to be replaced by references to the timeline.

Currently a number of annotation tools are under development (e.g. Elan (Brugman and Wittenburg, 2001), AGTK (Bird et al., 2001), Anvil (Kipp, 2001), Exmaralada (Schmidt, 2001)), each of them designed for a specific target audience. Most of the tools are using Java as an implementation base and encode the linguistic data in a comparable way as proposed here (XML-based, using time spans to mark events, separating meta-data and content). As a result it becomes relativly easy to convert/generate TASX-annotated corpora into/from these formats.

2.1 The TASX-annotator and the corpus engine

The complete TASX-environment consists of:

- tools for the annotation of empirical language data (video and audio material),
- a simple meta-data editor
- programs for the transformation of various formats of linguistic standard software (Transcriber, Praat, ESPS/waves+, SyncWriter, Exmaralda etc.)
- a set of programs for linguistic analysis of the TASX-annotated data, and
- a corpus system for the distribution of language data via the internet, including interactive corpus query and multimodal data display in a standard web browser.

In the following sections these modules will be described in more detail (see also (Milde and Gut, 2001), (Milde and Gut, 2002)).

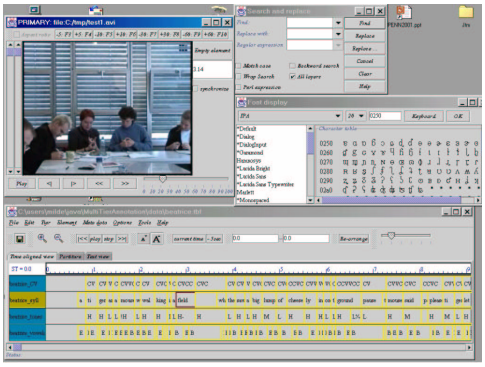


Figure 1: A screenshot of the TASX-annotator. In the bottom half, the main panel is visible, where the time aligned tier view has been selected. On top of the main window, the font selection panel is visible (showing some IPA characters). Above it, the find tool is shown. In the upper left corner the video display can been seen.

2.2 The TASX-annotator

The TASX-annotator is a central component of the TASX-environment. The tool allows the multilevel annotation and transcription of video (multi-channel) and audio data (see figure 1).

The programm is very user friendly and can be used without a high level of computer skills. It is possible to completely control the tool by either mouse *or* by keyboard shortcuts.

Video and audio playback can be controlled by a foot switch. Different data views are programmed (time-aligned partiture, word-aligned partiture, sequential text view) to make annotation as effective as possible.

The time aligned view is organized as a two dimensinal grid of infinite size. A layer is presented as a horizontal tier of events. The order of the layers is arbitrary and can be changed instantly. The user is able to define time intervals by dragging the mouse. Each time interval represents an event. The event is displayed as a graphical box which can be selected and moved with the mouse. The content of an event is entered in an additional text field. Any (unicode) font (e.g. IPA fonts, HamNoSys fonts etc.) available for the operating system can be used for the transcription. The user can choose font and fontpage from a table displaying all characters of the selected font. It is also possible to define a virtual keyboard which maps the given keystrokes to arbitrary characters of the target font.

A separate video playback window will be opened up for each video file making it possible to e.g. display multiple perspectives of the same scene. The video playback is synchronized with the transcription. For audio transcriptions an oszillogram is calculated and is displayed inside the main window.

In the text view the data can be manipulated in a standard text editor panel. The content of the editor represents the layer and each line represents an event. A list selection box allows switching between different layers. It is possible to transfer text from standard text editors, e.g. Microsoft Word, by cut and paste operations. In order to additionally speed up the transcription process, a word completion function has been implemented for the text view. Entering the initial letter of a word and consecutively pressing CTRL+L will bring up all words starting with this letter. Once the text is tranferred into the TASX-annotator, the events still have to be aligned with the primary audio and video data. Switching back to the time aligned view and moving the events with the mouse makes this task quite simple.

In the partiture view the data cannot be edited. In practice this means that the data is transformed into an HTML table and then displayed to the user. A number of different HTML formatted views have been designed. The views can also be saved to external files and loaded into standard web browsers.

One potential strength of the TASX-annotator is its manner of handling the export/import of XML based information. A standard way of solving this problem would be the implementation of a set of format specific XML parsers which construct the internal representation (e.g. JDom) of the XML file. While powerful integrated development systems such as *Sun's Forte for Java* make the

design of such XML handlers simpler, it still remains a complex task to implement such a parser. In the TASX-annotator we follow a different approach. The system integrates an XSL-T processor (saxon), making it easy to perform on the fly data transformations. The import of an XML-file is split into two steps: first an XSL-T stylesheet transforms the XML file into TASX, second another XSL-T stylesheet will transform the TASX file into a simple text oriented format. This format can be loaded efficiently.

A crucial problem when setting up larger corpora are inter anotator transcription errors. While the TASX-annotator is designed to be used by a single person, it still provides a number of routines to combine (merge), control and align annotations created by a larger team of people. We do not integrate more complex control functions. This contradicts our approach of clearly separating corpus creation from corpus analysis.

2.3 Transcoding tools

The development of tools for the TASX-environment is based on the concept that a re-implementation of functionalities already available in other language and speech processing software is not necessary. Established software systems such as Praat or ESPS/waves+ do not need to be duplicated.

TASX	import	export
Annotation graphs	XSL-T	XSL-T
Exmaralda	XSL-T/Java	Java/XSL-T
HTML-table	–	XSL-T
HTML-partiture	–	XSL-T
RTF	–	XSL-T/Java
Anvil	XSL-T	–
Praat-label	Perl/XSL-T	XSL-T
ESPS-label	Perl	XSL-T
ESPS-freq	Perl/XSL-T/Java	XSL-T
SyncWriter	Perl	–

Table 1: List of currently implemented transcoding tools. The table shows the programming languages used to implement the transcoders.

The TASX-environment therefore focuses on the development of transcoding filters from and into various formats. These include: Praat/freq, Praat/label, ESPS/waves+, ESPS/F0-analysis, Transcriber, annotation graphs stored in XML, SyncWriter and basic text formats (see table 1). In addition, filters for data import and export of the Exmaralda system (Schmidt, 2001) are available. Most of these components are implemented in Java, transformations are defined in XSL-T and a smaller number of additional tools is written in Perl (mainly to transform non-XML data).

2.4 Pause tracker

To speed up the annotation process a pause tracking programm has been developed. The programm separates speech from pauses and generates a TASX annotated XML document with two tiers, one holding all pause events, the other one holding all speech events.

The tracker uses Praat (Boersma, 2001) to perform the actual speech analysis. It simply calculates the pitch curve of the audio signal. If no pitch is detected, then non-speech is assumed, otherwise speech. In a second step, the results of this classification are combined to continuous stretches of pauses/speech. Finally the TASX conformant output is generated.

The pause tracker has shown to work quite reliably on a set of recording in different languages (Japanese, English, German, Saterfriesisch, French, Ega). Even if tracking is far from perfect, the human annotator gets a good pre-segmentation of the signal. This allows to move very quickly through the file, possibly performing minor adjustments to the boundaries or combining a set of separated events of one speaker.

While the pause tracker gives good results when doing conversational analysis it is not of much help for fine grained phonetic research. Here a tracking system for vowels and consonants would be very useful. Garcia et.al. are working on such a system (Garcia et al., 2002)

2.5 The corpus system

The main function of the corpus system constitutes the internet-based dissemination of

the corpus data. With the currently implemented interface it is also possible to inspect and query the speech corpus, to listen to the audio material and to display the graphic representation of the waveforms in a standard web browser. We make use of the built-in features of the web browser here. Furthermore, the PAX-tools (Gibbon and Trippel, 2001) for displaying the intonation contour, the intensity and the spectrogram of the selected regions in the audio file can be integrated.

When playing back the sound file, both the audio parts and the waveform images are generated automatically by a small Java servlet program. The servlet parses the XML-annotated corpus, extracts the time stamps of the relevant events and then cuts out the corresponding parts of the original sound file.

The corpus system is split into two larger subcomponents: the *information pool* and the *corpus engine* (see figure 2). The information pool stores the primary data (raw audio data) as well as the XML-annotated transcriptions of the audio files. The corpus engine consists of five subsystems:

1. Web-client: the interactive user interface is completly defined to run in a standard web browser. We are using HTML-query forms which activate services on the server side to generate XSL-T-filters processing the data. Waveforms are displayed using SVG. This will allow the user to select parts of the sound signal and to perfom more complex phonetic analyses.

2. Web-server: the web server distributes the corpus information in several standard formats (XML, HTML, PDF, SVG, WAV).

3. Servlet-engine: the servlet engine activates the suitable services on the server side (transformation of XML-annotated data, on-the-fly phonetic analysis, generation of graphics).

4. Servlets: a set of TASX/XML-aware servlets are used to transform the data in numerous ways: generating HTML to be displayed in the browser, generating PDF to be printed out, generating wavefiles and images of the waveforms. XSL-T and XSL-FO are used to perfom the transformations. The servlets have access to the information pool and the relational database.

5. Relational database: in order to improve the system performance, the XML-annotated corpus data is stored in a relational database. The database basically replaces a standard file system. An XSL-T-program translates the XML-annotated corpus data into a suitable format for the DBMS.

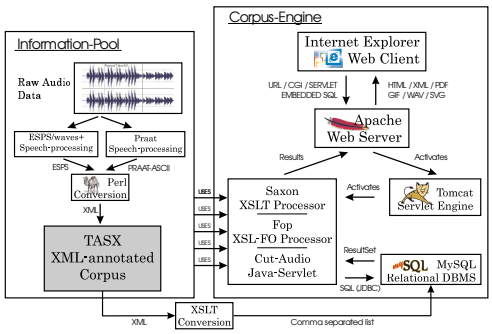


Figure 2: The system architecture of the corpus system. The corpus system is split into two subsystems: the information pool (left) storing the TASX-annotated data and the corpus engine (right) distributing the data over the internet.

The implementation of the corpus system is based on open source software. The TASX-annotator is a pure Java application; all other tools are smaller XSL-T and perl scripts. As a result, the complete TASX-environment runs on Windows and Unix platforms. The software will be distributed under GPL and can be downloaded from our website[2].

2.6 Statistical analysis

In the inital design phase of the TASX system we planned to implement the statistical analysis in XSL-T and Java. Indeed, a number of smaller programs have been realized in this

[2] http://coli.lili.uni-bielefeld.de/~milde/tasx/

technique. Unfortuneatly it quickly became evident, that XSL-T is not suited to perform such calculations on larger sets of data. It lacks high precision arithmetic functions and consumes to much memory. When using external Java functions, a large number of data conversions have to take place. Also the resulting code is very hard to read and debug.

Instead we have chosen to use the R system, an open source implementation of the `S-Plus` statistics language (Ihaka and Gentleman, 1996), (Venables and Ripley, 1999). R implements all major statistical tests and calculations and is equipped with a large number of high level graphic routines to generate visually informative presentations of the results. Even more important it includes efficient input/output routines to load and save semistructured data (either XML-annotated or plain ascii text).

3 Conclusions

Despite the early stage of the research the TASX-based approach has already proved to be highly efficient and reliable. The time consuming task of segementing speech data is partially substituted by automatic analysis. In the automatic transformation process from non-XML to XML-annotated data a number of errors in the human annotations can be detected. Furthermore, due to the highly structured format of the TASX-converted data more complex research questions can be investigated in a systematic way.

The very good availabilty of XML aware software and tools enabled us to develop a powerful linguistic environment in a very short time. Even more important, the TASX-annotated data can be transformed into large number of different formats. The will hopefully lead to the creation of linguistic resources which can be used over a long period of time by different researchers with a wide range of scientific goals.

References

S. Bird and M. Liberman. 1999. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania.

Steven Bird, Kazuaki Maeda, and Xiaoyi Ma. 2001. Agtk: the annotation graph toolkit. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Hennie Brugman and Peter Wittenburg. 2001. Mpi tools for linguistic annotation. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

J. Carletta, D. McKelvie, and Isard A. 2002. Supporting linguistic annotation using xml and stylesheets. In G. Sampson and D. McCarthy, editors, *Readings in Corpus Linguistic, Continuum International*.

L. Dybkaer, M. B. Moeller, N. O. Bernsen, J. Carletta, A. Isard, M. Klein, D. McKelvie, and A. Mengel. June 1999. The mate workbench. In David Traum, editor, *Proceedings of ACL'99, Demonstration Abstracts. University of Maryland*, pages 12 – 13.

J.E. Garcia, U. B. Gut, and A. Galves. 2002. Vocale - a semi-automatic annotation tool for prosodic research. In B. Bel and I. Marlien, editors, *Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage*, pages 327 – 330.

D. Gibbon and T. Trippel. 2001. Pax - an annotation based concordancing toolkit. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

R. Ihaka and R. Gentleman. 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Michael Kipp. 2001. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the Eurospeech 2001, Aalborg*, pages 1367 – 1370.

J.-T. Milde and U. B. Gut. 2001. The TASX-engine: an XML-based corpus database for time aligned language data. In Peter Buneman

Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA.*

J.-T. Milde and U. B. Gut. 2002. The tasx-environment: an xml-based toolset for time aligned speech corpora. In *Proceedings of the third international conference on language resources and evaluation (LREC 2002, Gran Canaria.*

T. Schmidt. 2001. Gesprächstranskription auf dem Computer - das System EXMARaLDA. *Gesprächsforschung, http://www.gespraechsforschung-ozs.de*, 2.

W. N. Venables and B. D. Ripley. 1999. *Modern Applied Statistics with S-Plus. Third Edition.* Springer. ISBN 0-387-98825-4.

Task-based multimodal dialogs

Dave Raggett, W3C/Openwave

Abstract

A model is presented for representing web-based multimodal dialogs as sets of prioritized tasks. This is motivated by an analysis of VoiceXML and requirements for richer natural language interaction. The model facilitates mixed initiative across a set of narrow application focussed domains.

Introduction

Setting the scene - my role in the web - the restricted nature of current voice-based human-machine dialogs - examples of a richer interaction style - the need for humility in the face of human intelligence - the opportunity for a modest extension in dialog capabilities.

I have been involved in the Web for many years, helping to drive the development of standards for HTML, HTTP and more recently work on voice browsing and multimodal interaction. HTML has enabled people to access content and services right across the world at the click of a button. HTML has been used to create a rich visual experience, but is not well suited for aural interaction. Work on aural style sheets has made it possible to style HTML when rendered to speech in combination with keyboard input, but the prevalence of table-based visual markup has made it difficult for people with visual impairments to easily browse visual web content. A better solution would help all of us when there is a need for hands and eyes free operation, or when we don't have access to a computer. At the time of writing there are well over a billion phones world-wide, could these be adapted to provide an effective means to access Web services? An affirmative answer would have a dramatic impact on the Web.

Speech Interaction

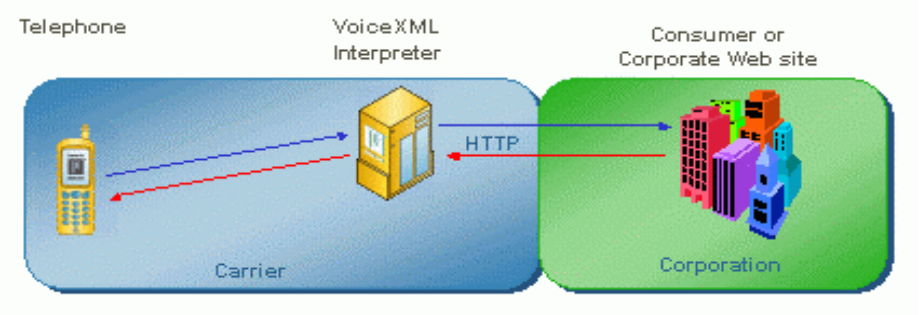
Speaker dependent speech recognition has been used for several years in dictation products, e.g. Scansoft's Dragon Dictate and IBM's ViaVoice. These products require the user to train the system to their voice to attain an adequate level of accuracy. More recently, speaker independent continuous speech recognition software has become available. This is made possible by using context free grammars to dramatically constrain the recognition task. The user is conditioned to respond within the scope of the grammar via carefully chosen prompts. This can be combined with word or phrase spotting techniques.

The need to write speech applications as complex programs is a powerful inhibitor for would be developers. As a result, a number of companies began to explore the use of markup as a means to reduce the effort needed from application developers. Some examples include, PML from AT&T and Lucent, SpeechML from IBM, VoxML from Motorola, and my own work at HP Labs on TalkML. These have focussed on menuing and form filling as metaphors for user interaction. AT&T, Avaya, Lucent and Motorola subsequently pooled their efforts to merge their experience into a joint design for a new language called VoiceXML, This work was later picked up by W3C's Voice Browser working group and supplemented by additional work on markup specifications for speech grammars and speech synthesis, drawing upon work by Sun Microsystems.

Learning from VoiceXML

The successful features, e.g. navigation links (main menu), form filling metaphor, tapered prompts, barge-in, traffic-lights model for confirmations. Flexibility through a judiscious mix of declarative and procedural elements. Mixed initiative in VoiceXML.

VoiceXML is being successfully deployed by wireless and wireline telephone network operators, and by companies for various kinds of call centers. A tutorial on VoiceXML is available on the W3C site. Users dial up to connect to a voice browser running a VoiceXML interpreter. This in turn contacts a web server to request the corresponding VoiceXML document. An application may extend across several VoiceXML documents. Developers are comfortable with markup and exploit their skills at dynamically generating markup on the fly, and providing for a division of labor between web servers and backend application servers.



VoiceXML supports global navigation links and form filling via the <link/> and <form>...</form> elements. VoiceXML supports the use of grammars for both speech recognition and DTMF (touch tone) input. For forms you can set form-level and field-level grammars. The results of speech recognition are treated either as activating a link or as setting the values of one or more named variables. There is no explicit model of dialog history. VoiceXML offers a judiscious mix of declarative and procedural features, with the ability to use ECMAScript for dynamically computed attribute values, and the ability to define event handlers in various scopes.

Different styles of interaction

VoiceXML applications are generally based upon a system directed dialog where the application does most of the talking and the user responds with short simple utterances. As an example, here is a fictious application for ordering pizza: [play it]

```
Computer: Welcome to Joe's Pizza ordering service
Computer: Select pizza size from large, medium or small?
User:  large
Computer: what number of these pizzas do you want?
User: two
Computer: Select first topping from mozzarella, pepperoni and anchovies?
User: mozzarella
Computer: Do you want another topping, yes or no?
User: yes
Computer: Select second topping from mozzarella, pepperoni and anchovies?
User: pepperoni
Computer: Do you want any other pizzas, yes or no?
```

```
...
```

The prompts are designed to elicit very simple responses, thereby avoiding the difficulties of dealing with all the possible variations in responses such as "yeah sure, I er would like large pizzas". If the user doesn't answer in a reasonable time, the application repeats the prompt, perhaps rewording it. If the answer doesn't match the grammar, the application provides guidance, for example:

```
Computer: what number of these pizzas do you want?
User: I reckon two would do the job
Computer: please say the number on its own
User: two
Computer: Select first topping from mozarella, pepperoni and anchovies?
...
```

The dialog gets the job done, but is very rigid. With larger grammars, a more natural interaction style becomes possible, for example: [play it]

```
Computer: Welcome to Joe's Pizza
Computer: What would you like?
User: I would like two large pizzas with mozzarella and one small pizza
with tomatoes and anchovies
Computer: would you like any drinks with that?
User: Sure, 3 large diet cokes, oh and add pepperoni to the large pizzas
Computer: Is that all?
User: yes
Computer: Okay, that will be ready for you in 5 minutes
User: thanks
```

In this example, the application starts with an open ended prompt. The context should be sufficient to guide the user to respond within the domain defined by the application. If the user's response can't be understood, the application provides guidance. Word spotting can be used as part of this process, where the presence of particular words triggers particular behaviors.

The example involves a structured data model going beyond the limits of flat lists of name/value pairs. The user's second response modifies information provided in the first response, necessitating some kind of query against the current state of the application data. This is something that would be hard to do with VoiceXML.

Multimodal dialogs

Visual interfaces based upon HTML are event driven and controlled by the user. This is very different from the system directed dialogs prevalent with VoiceXML. Microsoft's SALT proposal extends HTML to trigger speech prompts and activate speech grammars via HTML events, such as onload, onfocus, onmouseover and onclick etc. The results of speech recognition are handled in two steps. The first is for the recognizer to apply the speech grammar to the spoken utterance to create an annotated XML representation of the parse tree. The second step is to use an XPath expression to extract data from this tree and to insert it into a named variable.

SALT doesn't provide much in the way of declarative support for dialogs. As a result SALT applications tend to involve plenty of scripting. By contrast, VoiceXML is reasonably good for representing dialogs, but poor when it comes to event driven behavior. What is needed is a dialog model that supports the best of both approaches.

W3C's work on multimodal interaction aims to support synchronization across multiple modalities and devices with a wide range of capabilities. The vision of a multimodal interface to the Web in every pocket calls for an architecture suitable for low end devices. This necessitates a distributed approach with network based servers taking on tasks which are intensive in either computation, memory or bandwidth. Examples include speech recognition, pre-recorded prompts, speech grammars, concatenative speech synthesis, rich dialogs and natural language understanding.

W3C's vision of multimodal also includes the use of electronic ink as produced by a stylus, brush or other tool. IBM, Intel and Motorola have proposed an XML format for transferring ink across the network. This would enable the use of ink for text input, for gestures used as a means of control, for specialized notations such as mathematics, music and chemistry, and for diagrams and artwork. Ink is not restricted to flat two dimensional surfces, and in principle can be applied to curved surfaces or three dimensional spaces. It is thus a goal for multimodal dialog frameworks to address the use of ink.

Mixed domains: Personal Assistants

Commercial offerings like General Magic's "Portico" and Orange's "Wildfire" provide users with personal assistants that allow you to browse mail boxes, listen to messages, compose and send messages, dial by name from your contact list, request and review appointments, listen to selected news channels and so forth.

This notion of a personal assistant can be considered as a group of intersecting application subdomains. In current systems, users are required to remember a set of navigation commands that move you from one subdomain to another. In some systems you have to say "main menu" to return to the top-level before issuing the command to move to the next subdomain of interest. A richer dialog model should allow you to move naturally between different subdomains without such restrictions.

VoiceXML supports the dialog model where you have permanently active navigation commands, together with task specific form filling dialogs, only one of which is active at any given time. It seems natural to consider a more flexible model whereby many tasks can be active at the same time, and waiting for the user to say something relevant to that task. Perhaps we can define a task based architecture as an evolutionary step beyond VoiceXML?

A task based architecture for multimodal dialogs

Navigation links and form fields in VoiceXML can be seen as examples of a more general notion of tasks, and suggests an approach involving a dialog interpreter that supports sets of active and pending tasks, where each task has a name and a priority ...

The previous sections have established the motivation for studying a more elaborate model for multimodal dialogs. Such a model doesn't spring fully formed out of the blue, so what follows should be considered as a preliminary sketch. Let's start with some ideas about tasks:

- tasks triggered by voice commands where the corresponding grammar is active for long periods
- tasks triggered by graphical user interface events, such as moving the pointer over some field, clicks on links, key presses, or recognized gestures based upon stylus movements
- tasks triggered by a timer, based upon specified offsets from other events, using the model established in W3C's SMIL specification
- tasks related to the current dialog focus, for instance, collecting information needed to fill out a form, this generally involves a turn-taking model, as in VoiceXML's fields
- tasks that ask the user to confirm or repeat something that wasn't heard reliably — you would normally ask the user to say it differently to increase the chances of success
- tasks that follow links, change the dialog focus, change the application state, or other actions, for instance handling a request for a prompt to be repeated, or a request for help
- tasks that create new tasks, terminate current tasks, or which change the priority of other tasks
- hierarchically structured tasks, where one task delegates work to subsidiary tasks, it creates for that purpose
- re-usable tasks involving a well defined interface and information hiding (VoiceXML subdialogs)

To make it easier for application developers, tasks should be represented declaratively. In the context of the Web this suggests markup. For instance, you could specify a task that is triggered by a mouse click, but which is only active between specified start and stop conditions. The corresponding markup could be derived from W3C's SMIL and XML Events specifications. The means to express actions will be discussed below following a consideration of how to approach natural language understanding.

To allow for richer voice interaction, a reasonable premise is for multiple grammars to be active at any time, and corresponding to different tasks. When the user says something that matches an active grammar, the utterance is handled by the task associated with that grammar. What if the utterance matches several grammars? This could happen because more than one task has activated the same grammar, or more likely, because the recognizer isn't quite sure what the user said. The solution is to prioritize tasks. The priorities can then be taken into account as part of the recognition process and combined with the recognition uncertainties to determine the most likely interpretation.

Natural language understanding

This is perhaps the most tricky area to deal with due to our very incomplete understanding of how the human brain operates. Language carries information at multiple levels and assumes a huge amount of knowledge about the world. Common sense is easy for people but intractable for machines, at least at the current state of technology. To get anywhere, it is critical to dramatically constrain natural language understanding to a narrow area that is amenable to a mechanical treatment, and within the scope of application developers.

The output from recognizers

Speech grammars define the set of expected utterances and are used to guide the recognizer. The output from the recognizer can be defined as an annotated natural language parse tree represented in XML. By defining the ouput of the recognizer as the most likely parse tree, there is a considerable loss of information compared with that available to the recognizer

itself. This is a trade-off. A simplified representation makes it easier to apply subsequent stages of natural language processing, as compared with a richer representation giving the estimated likelihoods of a plurality of interpretations (for instance, a lattice of possible phoneme sequences).

Speech technology vendors have worked long and hard to improve the robustness of speech recognition for things like numbers, currency values, dates, times, phone numbers and credit card details. It therefore makes sense to incorporate the results of such processing into the output from the recognizer. The output is the most likely natural language parse tree, annotated with recognition confidence scores and the results of semantic preprocessing by recognizers. W3C has been working on an XML representation for this, called NLSML or natural language semantics markup language. This work is still at an early stage and may well change name by the time it is done.

Natural language understanding rules

The next step is to apply natural language understanding rules to interpret the utterance in the context of the current task and application state. The result is a sequence of actions to be performed. The actions cover such things as changing the application state, starting and stopping other tasks, following links, changing the dialog focus and so on. See the earlier section on tasks for other ideas. How should these natural language understanding rules be represented and what do they need to be capable of?

One posibility is support a sequence of if-then rules where the "if" part (the *antecedent*) operates on the output of the recognizer, the current application state, task specific data, and the dialog history. The "then" part (the *consequent*) specifies actions, but also can access information passed to it from the antecedent, and from the same sources as are available to the antecedent. These rules could be directly associated with grammar rules or could be bound to grammars at the task level. The rules could in turn invoke additional rule sets (modules).

The detailed representation of these rules is likely to be a contentious issue. XML experts will probably place a premium on consistency with existing XML specifications, for instance XPath and XSLT. Others who place a premium on simplicity for end-users may prefer a more consise and easier to learn syntax that is closer to conventional programming languages. For added flexibility it would be advisable to allow for breaking out to a general purpose scripting language such as ECMAScript, or a rule oriented language such as Prolog.

Task specific data

Tasks may provide locally scoped data. This corresponds to locally scoped variables in subroutines in common programming languages. This information is hidden from other tasks, unless exposed through defined methods. This assumes that tasks can be treated as *objects* with *methods*. An object-oriented approach blends declarative and procedural styles, and makes it straightforward for tasks to provide appropriate behaviors in response to a variety of events.

Application state

For many applications there will be a need for richly structured application information, whether this is for ordering pizza or for a personal assistant with access to mail boxes, contact lists and appointment calendars. Application developers will need a consistent interface to this data, and it is not unreasonable to do so via XML. This doesn't mean that data is expressed internally as XML files, but rather that the interface to the data can be handled via operations on XML structures.

In some cases, this may involve a time consuming transaction with a back-end system, e.g. a database on another server. Application developers need to be aware of such delays when designing the interaction with the end-user. For delays of about two seconds or longer, it is necessary to let the user know that some time consuming task is underway. A tick-tock sound effect is sometimes used as the aural equivalent of an hour glass. For longer delays, it is worth considering how to involve the user in some other activity until the task has been completed.

Dialog history

Sometimes the user might refer back to something mentionned earlier in the dialog. It may be possible to handle this in terms of a reference to the current application state, otherwise, it is necessary to maintain a representation of the sequence of prompts and responses. Observations of human short term memory suggest that only a small number turns need to be available. The *dialog history* can be represented at several levels, for instance:

- the text of the utterances as spoken by the user and by the application
- the parse trees as output by the recognizer
- semantically meaningful information placed in the dialog history by the natural language understanding rules or directly by active tasks (e.g. handlers for mouse clicks)

The dialog history is accessible by the antecedents and consequents of the natural language understanding (NLU) rules. Linguistic phenomena such as anaphora, deixis, and ellipsis can be treated in terms of operations by the NLU rules on the current or preceding utterances. Anaphoric references include pronouns and definite noun phrases that refer to something that was mentioned in the preceding linguistic context, by contrast, deictic references refer to something that is present in the non-linguistic context. Ellipsis is where some words have been left out when the context makes it "obvious" what is missing. If the NLU rules aren't able to make sense of the utterance then application developers should provide some fall back behavior.

Application developers may want to allow the user to make responses that combine multiple modalities. One example is where the user is shown a street map centered around his/her current position. The user might ask how long it would take to walk to "here" while clicking on the map with a stylus. The NLU rules in this case would have to search the dialog history for positional information as recorded by the handler for the click event.

A distributed model of events and actions

The need to support a mass market of low-end devices makes it imperative to provide a distributed architecture. The Web already has a model of events, as introduced into HTML, the next step will be to extend this across the network.

The events are divided into actions and notifications. Actions are events that cause a change of state, while notifications are events that are thrown as a result of such changes. Here are some examples:

Changing the input focus in an XHTML page
: A notification event is thrown by a field when it acquires or loses the focus. The corresponding message includes the name of the event and an identifier for the field involved. The corresponding action event targets the field that will as a result acquire the focus.

Changing the value of an XHTML field
: An event to change the value can be sent as a result of user action via one or more modes of input, for instance, the keypad, stylus or speech. The action event includes the new value and targets the field to be updated. As a consequence of the update, a notification event is thrown to all observers interested in learning about changes to that field.

Changing to a new XHTML page
: The action event to change to a new page can be triggered in several ways, for instance, by tapping on a link, selecting a link with the keypad or saying the appropriate command. The corresponding notification events signal the unloading of the current page, and the loading of the new page.

Changing the page structure and content
: The results of a spoken utterance could lead to changes to the visual page's structure and content. In a conventional, web page, this would be achieved through scripting and calls that manipulate the document object modal.

Events can effect user interface specific features or modality independent abtractions. For example, when the user says a command to follow a link, this could be targeted at a button in the visual interface, resulting in this button appearing to depress momentarily. If the action is targeted at the page, the button won't be effected.

The XML Events specification describes markup for use in binding handlers to events following the model defined in the W3C DOM2 Recommendation. The framework needs to be extended to support the notion of action events, and to describe the representation of events as XML messages. This can be kept separate from the underlying transport protocols. In 2.5G and 3G mobile networks, the IETF SIP events specification looks like a natural fit.

In an asynchronous system, care needs to be taken to avoid inconsistencies arising. In one example, the user says something to select a choice from a menu, but then uses the stylus to tap on different choice on the same menu. In the time taken to recognize the speech and send the corresponding action, the visual interface will have already changed the value, based upon the stylus tap.

The simplest policy is to apply actions in the order they are received. An alternative would be to include a time stamp and to ignore an action that occurred before the latest action that was

applied. If a more sophistocated approach is needed, it may be feasible to define script handlers that intercept the actions before they are applied.

Dialog models involving explicit turn taking provide a further basis for synchronization. The events are tagged with the turn, and this can be used to identify events that arrive out of turn. Further work is needed to understand how turn taking relates to the user interface model in XHTML.

One idea is to use an identifier corresponding to the web page. If an event is delivered after the page has changed, the event can be easily discarded or directed to an appropriate handler. For applications that last over multiple web pages, a session context seems appropriate, and fits with existing ideas for WML and VoiceXML.

When it comes to actions that change the structure and content of a document, then it would be interesting to compare and contrast approaches based upon transferring small scripts (scriplets) and more declarative approaches based upon markup. In both cases, it may be necessary to consider security mechanisms to avoid problems with hostile third parties intervening in the dialog between devices and servers.

Next Steps

This paper has presented an analysis of the requirements for multimodal dialogs and proposed a sketch of a task-based architecture using events for synchronization across modalities and devices. It is to be hoped that this paper will help to stimulate further discussion bridging the academic and commercial communities. Experience has shown that it takes several years to create Web standards. Now is the time to ensure that the next generation of Web user interfaces are grounded on solid review by both communities.

References

General Magic
http://www.generalmagic.com/
HTML
http://www.w3.org/MarkUp/
HTTP
http://www.w3.org/Protocols/
SALT Forum
http://www.saltforum.org/
Synchronized Multimedia Interaction Language (SMIL)
http://www.w3.org/AudioVideo/
TalkML
http://www.w3.org/Voice/TalkML/
VoiceXML Forum
http://www.voicexml.org/
VoiceXML tutorial
http://www.w3.org/Voice/Guide/
Wildfire
http://www.wildfire.com/
W3C NLSML specification

http://www.w3.org/TR/nl-spec/
W3C Speech Grammar specification
http://www.w3.org/TR/speech-grammar/
W3C Speech Synthesis specification
http://www.w3.org/TR/speech-synthesis
W3C VoiceXML 2.0 specification
http://www.w3.org/TR/voicexml20/
W3C Voice Browser activity
http://www.w3.org/Voice/
W3C XML Events specification
http://www.w3.org/TR/xml-events/
W3C XPath specification
http://www.w3.org/TR/xpath
W3C XSLT specification
http://www.w3.org/TR/xslt

MIAMM - Multidimensional Information Access using Multiple Modalities

Norbert Reithinger, Christoph Lauer
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken, Germany
{bert,clauer}@dfki.de

Laurent Romary
CNRS, INRIA & Universités de Nancy
Campus Scientifique - BP 239
F-54506 Vandoeuvre Lès Nancy, France
laurent.romary@loria.fr

Abstract

Haptic interactions add new challenges to multi-modal systems. With the MIAMM[1] project we develop new concepts and techniques to allow natural access to multimedia databases. In this paper we give an overview of our approach, and discuss the architecture and the MMIL interface language. A short example provides a general feeling of the possible interactions.

Keywords: Architectures, interface languages, haptics

1 Introduction

The main objective of the MIAMM project (www.miamm.org) is to develop new concepts and techniques in the field of multi-modal interaction to allow fast and natural access to multimedia databases. This will imply both the integration of available technologies in the domain of speech interaction (German, French, and English) and multimedia information access, and the design of novel technology for haptic designation and manipulation coupled with an adequate graphical presentation.
A design study for the envisioned handheld appliance is show in figure 1. The user interacts with the device using speech and/or the haptic buttons to search, select, and play tunes from an underlying database. In the example, the user has loaded her list of favourites. She can change the speed of rotation pressing the buttons.

Figure 1: Design study of the MIAMM device

Haptic feedback can also provide e.g. the rhythm of the tune currently in focus through tactile feedback on the button. If the user wants to have the list rotate upward, she presses the topmost button on the left and has to apply a stronger force to accelerate the tape more quickly.
The experimental prototype will use multiple PHANToM devices (www.sensable.com), see figure 3 (Michelitsch et.al. 2002), simulating the haptic buttons. The graphic-haptic interface is based on the GHOST software development kit provided by the manufacturer. The other modules of the system will be contributed by the project partners.
In the remainder of the article we will shortly present the software architecture of MIAMM, the basic principles for the design of a unified interface language within the architecture, and finally a short example dialog that is the basis for the ongoing implementation.

[1] Multidimensional Information Access using Multiple Modalities, EU/IST project n°2000-29487

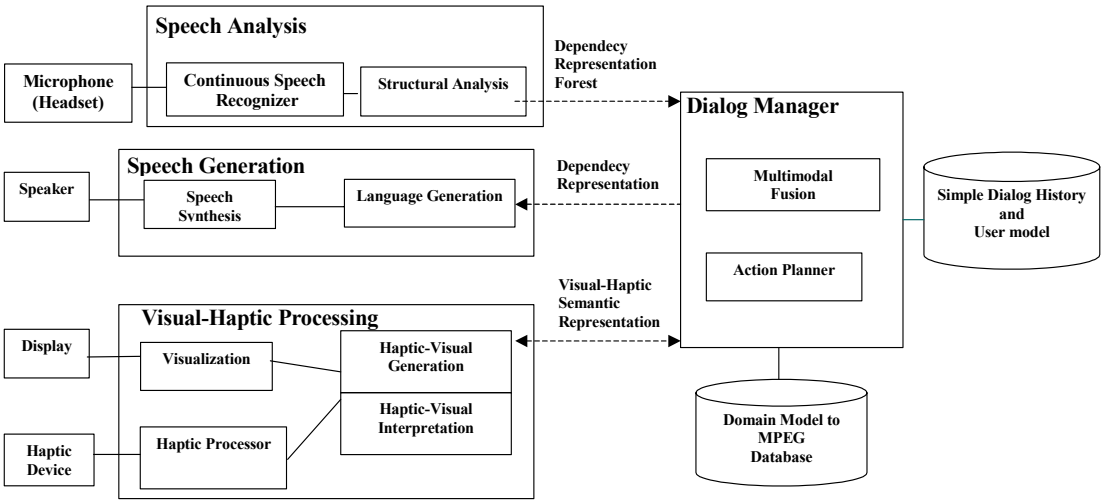


Figure 2: Miamm general architecture

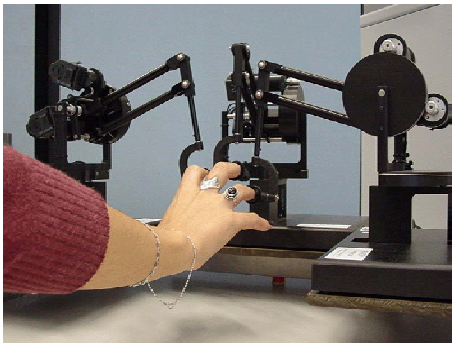


Figure 3: The simulation of the buttons using PHANToM devices

2 The Architecture

The participants of the Schloss Dagstuhl workshop *"Coordination and Fusion in Multimodal Interaction"* (see http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/ for the presentations) discussed in one working group architectures for multi-modal systems (WG 3). The final architecture proposal follows in major parts the "standard" architecture of interactive systems, with the consecutive steps mode analysis, mode coordination, interaction management, presentation planning, and mode/media design For MIAMM we discussed this reference architecture and checked its feasibility for a multi-modal interaction system using haptics. We came to the conclusion that a more or less pipelined architecture does not suit the haptic modality. For modalities like speech no immediate feedback is necessary: you can use deep reasoning and react in the time span of 1 second or more.

Consider however the physiology of the senso-motoric system: the receptors for pressure and vibration of the hand have a stimulus threshold of 1µm, and an update frequency of 100 to 300 Hz (Beyer&Weiss 2001). Therefore, the feedback at the buttons must not be delayed by any time-consuming reasoning processes to provide a realistic interaction: if the reaction of the system after depressing one button is delayed beyond the physiologically acceptable limits, it will be an unnatural interaction experience.

As a consequence, our architecture (see figure 2) considers the modality specific processes as agents which may have an internal life of their own: only important events must be sent to the other agents, and other agents can ask about the internal state of agents.

The system consists of two agents for natural language processing, one for the analysis side, and one for the generation and synthesis. The visual-haptic agent is responsible for the visualization, the assignment of haptic features to the force-feedback buttons, and for the interpretation of the force imposed by the user. The dialog manager consists of two main blocks, namely the multi-modal fusion which is responsible for the resolution of multi-modal references and of the action planner. A simple dialog history and user model provide contextual information. The action planner is connected via a domain model to the multi-media database. All accesses to the database are facilitated by the domain-model inference engine.

In the case of the language modules, where reaction time is important, but not vital for the true experience of the interaction, every result, e.g. an analysis from the speech interpretation, is forwarded directly to the consuming agent. The visual-haptic agent with its real-time requirements is different. The dialog manager passes the information to be presented to the agent, which determines the visualization. It also assigns the haptic features to the buttons. The user can then use the buttons to operate on the presented objects. As long as no dialog intention is assigned to a haptic gesture, all processing will take place in the visual-haptic agent, with no data being passed back to the dialog manager. Only if one of this actions is e.g. a selection, it passes back the information to the dialog manager autonomously. If the multi-modal fusion needs information about objects currently in the visual focus, it can ask the visual-haptic agent.

4 The interface language MMIL

The implementation of the MIAMM demonstrator should be based upon the definition of a unified representation format that will act as a lingua franca between the various modules identified in the architecture of the system. This representation format (called MMIL, Multi-Modal Interface Language) must be able to accumulate the various results yielded by each of these modules in a coherent way so that on one hand, any other module can base its own activity upon the information which it precisely requires and, on the other hand, it is possible to log the activity within the MIAMM demonstrator on the sole basis of the information which is transited within the components of the system. This last functionality is particularly important in the context of the experimentation of innovative interaction scenarios combining spoken, graphical and haptic modalities, for which we will have to evaluate the exact contribution of each single mode to the general understanding and generation process.

One of the underlying objectives behind the definition of the MMIL language is to account for the incremental integration of multi-modal data to achieve, on one hand, a full understanding of the user's multi-modal act (possibly made of a spoken utterance and a gestural activity), and, on the other hand, provide all the necessary information to generate multi-modal feedback (spoken output combined with a graphical representation and/or haptic feedback) to the user. The integration (fusion) or design (fission) of multi-modal information should obviously be based on the same representation framework, as these two activities could be seen as two dual activities in any communication scenario. In this context, one of the complexities of the design of the MMIL language will be to ensure that such a multi-modal coordination can both occur at a low level of the architecture (e.g. synchronous combination of graphics and haptics), up to high-level dialog processes (e.g. multi-modal interpretation of a deictic NP in combination with a haptic event). One question that can be raised here is the decoupling of real time synchronization processes (haptic-graphics) from understanding processes, which occur at a lower temporal rate[2].

One other important issue is to make sure that MMIL is kept independent from any specific theoretical framework, so that it can cope for instance with the various parsing technologies adopted for the different languages in MIAMM (template based vs. TAG based parsing). This in turn may provide to MMIL some degree of genericity, which could make it reusable in other contexts.

Given this, we can identify the following three basic requirements for the MMIL language:

- The MMIL language should be flexible enough to take into account the various types of information identified in the preceding section and be extensible, so that further developments in the MIAMM project can be incorporated;
- Whenever it is possible, it should be compatible with existing standardization initiatives (see below), or designed in such a way (in particular from the point of view of documentation) that it can be the source of future standardizing activities in the field;
- It should obviously be based on the XML recommendation, but should adopt a schema

[2] Even if we consider it useful to deal with haptic synchronization at the dialog manager level, the performance of such a dialog manager might not be sufficient to keep up with the update rate required by haptic devices.

definition language that is powerful enough to account for the definition of both generic structures and level specific constraints.

One major challenge for MIAMM appears to be the creation of a new ISO committee (TC37/SC4) on language resources which should comprise, among other things some specific activities on multi-modal content representation (see Bunt & Romary 2002). Such a format is likely to be close to what is needed within MIAMM and our goal is to keep as close as possible to this international initiative.

5 A short example interaction

To our knowledge, the envisioned interaction techniques have not been investigated yet. Therefore, task and human factors analysis plays an important part (see also Michelitsch et al. 2002). From the task analysis, we have first sample dialogs, which we use as starting point for implementation.

With the interactions below we demonstrate, how the internal processing will proceed in the realised prototype. We assume that the user has listened in the morning to some songs and stored the list in the memory of MIAMM. First the user says

"show me the songs I listened to this morning"

The utterance is analysed, resulting in an intention based MMIL representation. The multi-modal fusion resolves the time and retrieves the list of tunes from the persistent dialog history. The action planner selects as the next system goal to display the list and passes the goal, together with the list to the visual-haptic agent. A possible presentation can be like the one shown in figure 1. The user now manipulates the tape, uses force to accelerate the tape, or revert the presentation direction. A marker highlights the interpret's name that is currently in focus. All this activities are encapsulated in the visual-haptic agent.

The user next selects one singer by uttering

"select this one"

while pressing the selection button on the right. Both agents, speech analysis and visual-haptic processing, send time-stamped MMIL representations to the dialog manager. The visual-haptic agent does not send graphical information, but rather the identifier of the selected object and the intention, e.g. marked. The multi-modal fusion gets both structures, checks time and type constraints, and fills the selection intention with the proper object. The action planner then asks the database via the domain model to retrieve all information for this singer and again dispatches a display order to the visual-haptic agent.

4 Conclusion

We presented the main objectives and first specifications of the MIAMM project. The first experiments as well as precise specification of both the basic user scenarios and the architecture show that incorporating a haptic device does not necessarily make the design of a multi-modal dialog system more complex but forces the designer to be aware of the requirements of the modalities to provide a coherent view of their various roles in the interaction. The first prototype will be operational at the end of 2002.

Acknowledgements

This paper is a quick overview of a team work conducted by the MIAMM crew: Charles BEISS, Georg MICHELITSCH, Anita CREMERS, Norbert REITHINGER, Ralf ENGEL, Laurent ROMARY, Dirk FEDELER, Andreas RUF, Silke GORONZY, Susanne SALMON-ALT, Uwe JOST, Eric MATHIEU, Eric KOW, Amalia TODIRASCU, Ralph KOMPE, Marta TOLOS RIGUEIRO, Frédéric LANDRAGIN, Myra VAN ESCH, Christoph LAUER, Henrik-Jan VAN VEEN, Markus LÖCKELT, Ashwani KUMAR, Jason WILLIAMS, Elsa PECOURT.

References

Lothar Beyer and Thomas Weiss (2001) *Elementareinheiten des somatosensorischen Systems als physiologische Basis der taktil-haptischen Wahrnehmung*. In "Der bewegte Sinn", Martin Grunewald and Lothar Beyer, eds., Birkhäuser Verlag, Basel, pp. 25-38.

Harry Bunt and Laurent Romary (2002) Towards Multimodal Content Representation, LREC2002 Workshop on Standardization of Terminology and Language Resource, Las Palmas May 2002.

Georg Michelitsch, Hendrik A.H.C. van Veen, and Jan. B.F. van Erp (2002) Multi-Finger Haptic Interaction within the MIAMM Project. In Eurohaptics 2002, Univ. Edinburgh.

Engagement between Humans and Robots for Hosting Activities

Candace L. Sidner

Mitsubishi Electric Res. Labs
201 Broadway
Cambridge,MA 02139
Sidner@merl.com

Abstract

To participate in conversations with people, robots must not only see and talk with people but make use of the conventions of conversation and of how to be connected to their human counterparts. This paper reports on research on engagement in human-human interaction and applications to (non-autonomous) robots interacting with humans in hosting activities.

Keywords: Human-robot interaction, hosting activities, engagement, conversation, collaborative interface agents, embodied agents.

1. INTRODUCTION

As a result of ongoing research on collaborative interface agents, including 3D robotic ones, I have begun exploring the problem of engagement in human interaction. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in interaction, evaluating staying involved, and deciding when to end connection.

To understand the engagement process I am studying human to human engagement interaction. Study of human to human engagement provides essential capabilities for human - robot interaction, which I view as a valid means to test theories about engagement as well as to produce useful technology results. My group has been experimenting with programming a (non-autonomous) robot with engagement abilities.

2. HOSTING ACTIVITIES

My study of engagement centers on the activity of hosting. Hosting activities are a class of collaborative activity in which an agent provides guidance in the form of information, entertainment, education or other services in the user's environment (which may be an artificial or the natural world) and may also request that the human user undertake actions to support the fulfillment of those services. Hosting activities are situated or embedded activities, because they depend on the surrounding environment as well as the participants involved. They are social activities because, when undertaken by humans, they depend upon the social roles of humans to determine next actions, timing of actions, and negotiation among the choice of actions. Agents, 2D animated or physical robots, who serve as guides, are the hosts of the environment. This work hypothesizes that by creating computer agents that can function more like human hosts, the human participants will focus on the hosting activity and be less distracted by the agent interface. Tutoring applications require hosting activities; I have experimented with a robot host in tutoring, which is discussed in the next section.

Another hosting activity, which I am currently exploring, is hosting a user in a room with a collection of artifacts. In such an environment, the ability of the host to interact with the physical world becomes essential, and justifies the creation of physical agents. Other activities include hosting as part of their mission: sales activities of all sorts include hosting in order to make customers aware of types of products and features, locations, personnel, and the like. In these activities, hosting may be intermingled with selling or instructional tasks. Activities such as tour guiding or serving as a museum docent are primarily hosting activities (see [1] for a robot that can perform tour guide hosting).

Hosting activities are collaborative because neither party determines completely the goals to be undertaken. While the user's interests in the room are paramount in determining shared goals, the host's (private) knowledge of the environment also constrains the goals that can be achieved. Typically the goals undertaken will need to be negotiated between user and host. Tutoring offers a counterpart to room exploration because the host has a rather detailed private tutoring agenda that includes the user attaining skills. Hence the host must not only negotiate based on the user's interest but also based on its own (private) educational goals. Accordingly the host's assessment of the interaction is rather different in these two example activities.

3. WHAT'S ENGAGEMENT ABOUT?

Engagement is fundamentally a collaborative process (see [2], [3]), although it also requires significant private planning on the part of each participant in the engagement. Engagement, like other collaborations, consists of rounds of establishing the collaborative goal (the goal to be connected), which is not always taken up by a potential collaborator, maintaining the connection by various means, and then ending the engagement or opting out of it. The collaboration process may include negotiation of the goal or the means to achieve it [4], [5]. Described this way, engagement is similar to other collaborative activities.

Engagement is an activity that contributes centrally to collaboration on activities in the world and the conversations that support them. In fact conversation is impossible without engagement. This claim does not imply that engagement is just a part of conversation. Rather engagement is a collaborative process that occurs in its own right, simply to establish connection between people, a natural social phenomenon of human existence. It is entirely possible to engage another without a single word being said and to maintain the engagement process with no conversation. That is not to say that engagement is possible without any communication; it is not. A person who engages another without language must rely effectively on gestural language to establish the engagement joint goal and to maintain the engagement. Gesture is also a significant feature of face-to-face interaction where conversations are present [6].

It is also possible to use language and just a few words to create and maintain connection with another, with no other intended goals. An exchange of hellos, a brief exchange of eye contact and a set of good-byes can accomplish a collaboration to be in connection to another, that is, to accomplish engagement. These are conversations for which one can reasonably claim that the only purpose is simply to be connected. The current work focuses on interactions, ones including conversations, where the participants wish to accomplish action in the world rather than just the relational connection that engagement can provide.

4. FIRST EXPERIMENT IN HOSTING: A POINTING ROBOT

In order to explore hosting activities and the nature of engagement, the work began with a well-delimited problem: appropriate pointing and beat gestures for a (non-autonomous) robot, called Mel, while conducting a conversation. Mel's behavior is a direct product of extensive research on animated pedagogical agents [7]. It shares with those agents concerns about conversational signals and pointing as well. Unlike these efforts, Mel has greater dialogue capability, and its conversational signaling, including deixis, comes from combining the Collagen™ and Rea architectures [8]. Furthermore, while 2D embodied agents [9] can point to things in a 2D environment, 2D agents do not effectively do 3D pointing.

Building a robot host relied significantly on the Paco agent [10] built using Collagen™ [11,12] for tutoring a user on the operation of a gas turbine engine. Thus Mel took on the task of speaking all the output of the Paco system, a 2D application normally done with an on-screen agent, and pointing to the portions of the display, as done by the Paco agent. The user's operation of the display through a combination of speech input and mouse clicks remains unchanged. The speech understanding is accomplished with IBM ViaVoice™'s speech recognizer, the IBM JSAPI (see the ViaVoice SDK, at www4.ibm.com/software/ speech/dev/sdk_java.html) to parse utterances, and the Collagen middleware to provide interpretation of the conversation, to manage the tutoring goals and to provide a student model for tutoring.

The Paco 2D screen for gas turbine engine tutoring is shown in figure 1. Note that the agent is represented by a small window, where text, a cursor hand and a smiling face appear (the cursor hand, however, is pointing at a button at the bottom of the screen in the figure). The face changes to indicate six states: the agent is speaking, is listening to the user, is waiting for the user to reply, is thinking, is acting on the interface, and has failed due to a system crash.

Our robotic agent is a homegrown non-mobile robot created at Mitsubishi Electric Research Labs [Paul Dietz, personal communication], consisting of 5 servomotors to control the movement of the robot's head, mouth and two appendages. The robot takes the appearance of a penguin (called Mel). Mel can open and close his beak, move his head in up-down, and left-right combinations, and flap his "wings" up and down. He also has a laser light on his beak, and a speaker provides audio output for him. See Figure 2 for Mel pointing to a button on the gas turbine control panel.

While Mel's motor operations are extremely limited, they offer enough movement to undertake beat gestures, which indicate new and old information in utterances [13], and a means to point deictically at objects with its beak. For gas turbine tutoring, Mel sits in front of a large (2 foot x 3 foot) horizontal flat-screen display on which the gas turbine display panel is projected. All speech activities normally done by the on-screen agent, as well as pointing to screen objects, are instead performed by Mel. With his wings, Mel can convey beat gestures, which the on-screen agent does not. Mel does not however change his face as the onscreen agent does. Mel points with his beak and turns his head towards the user to conduct the conversation when he is not pointing.

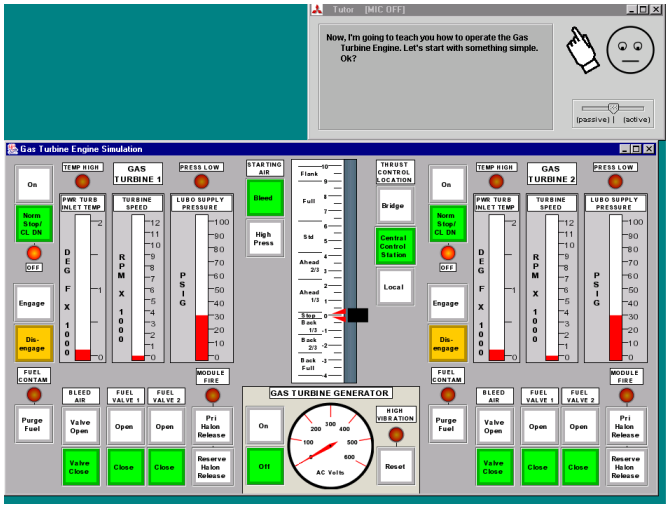


Figure 1:The Paco agent for gas turbine engine tutoring

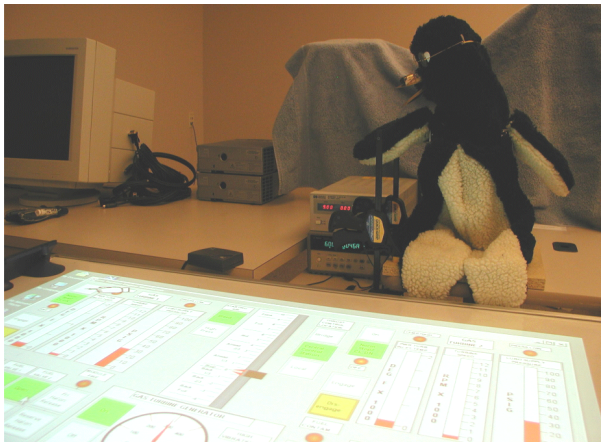


Figure 2: Mel pointing to the gas turbine control panel

The architecture of a Collagen agent and an application using Mel is shown in figure 3. Specifics of Collagen internal organization and the way it is generally connected to the applications are beyond the scope of this paper; see [11] for more information. Basically, he application is connected to the Collagen system through the application adapter. The adapter translates between the semantic events Collagen understands and the events/function calls understood by the application. The agent controls the application by sending events to perform to the application, and the adapter sends performed events to Collagen when a user performs actions on the application. Collagen is notified of the propositions uttered by the agent via uttered events. They also go to the AgentHome window, which is a graphical component responsible in Collagen for showing the agent's words on screen as well as generating speech in a speech-enabled system. The shaded area highlights the components and events that were added to the basic Collagen middleware. With these additions, utterance events go through the Mel annotator and BEAT system [13] in order to generate gestures as well as the utterances that Collagen already produces. More details on the architecture and Mel's function with it can be found in [14].

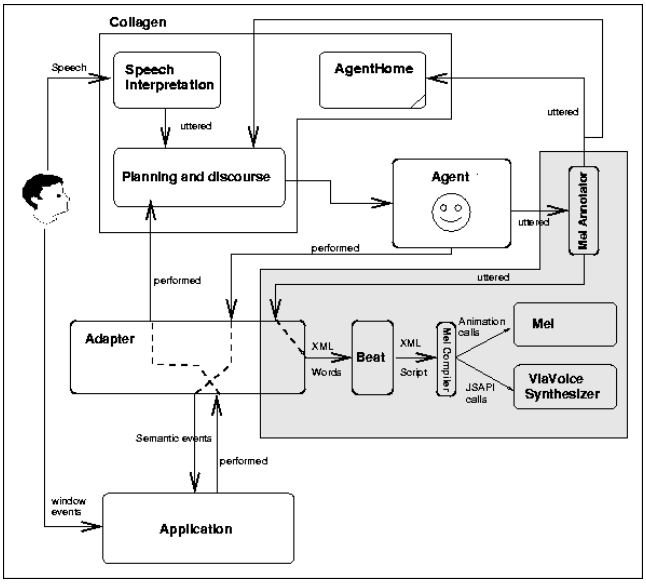


Figure 3: Architecture of Mel

5. MAKING PROGRESS ON HOSTING BEHAVIORS

Mel is quite effective at pointing in a display and producing a gesture that can be readily followed by humans. Mel's beak is a large enough pointer to operate in the way that a finger does. Pointing within a very small margin of error (which is assured by careful calibration before Mel begins working) locates the appropriate buttons and dials on the screen. However, the means by which one begins a conversation with Mel and ends it are unsatisfactory. Furthermore, Mel has only two weak means of checking on engagement during the conversation: to ask "okay?" and await a response from the user after every explanation it offers, and to await (including indefinitely) a user response (utterance or action) after each time it instructs the user to act.

To expand these capabilities I am studying human-human scenarios to determine what types of engagement strategies humans use effectively in hosting situations.

Figure 4 provides a constructed engagement scenario that illustrates a number of features of the engagement process for room hosting. These include: failed negotiations of engagement goals, successful rounds of collaboration, conversational capabilities such as turn taking, change of initiative and negotiation of differences in engagement goals, individual assessing and planning, and execution of end-of-engagement activities. There are also collaborative behaviors that support the action in the world activities (called the domain task) of the participants, in this case touring a room. In a more detailed discussion of this example below, these different collaborations will be distinguished. Significant to the interaction are the use of intentionally communicative gestures such as pointing and movement, as well as use of eye gaze and recognition of eye gaze to convey engagement or disengagement in the interaction.

In this scenario in part 1 the visitor in the room hosting activity does not immediately engage with the host, who uses a greeting and an offer to provide a tour as means of (1) engaging the visitor and (2) proposing a joint activity in the world. Both the engagement and the joint activity are not accepted by the visitor. The visitor accomplishes this non-acceptance by ignoring the uptake of the engagement activity, which also quashes the tour offer.

However, the visitor at the next turn finally chooses to engage the host in several rounds of questioning, a simple form of collaboration for touring. Questioning also maintains the engagement by its very nature, but also because the visitor performs such activities as going where the host requests in part 2. While the scenario does not stipulate gaze and tracking, in real interactions, much of parts 2 through 6 would include various uses of hands, head turns and eye gaze to maintain engagement as well as to indicate that each participant understood what the other said.

In part 4, the host takes over the initiative in the conversation and offers to demonstrate a device in the room; this is another offer to collaborate. The visitor's response is not linguistically complex, but its intent is more challenging to interpret because it conveys that the visitor has not accepted the host's offer and is beginning to negotiate a different outcome. The host, a sophisticated negotiator, provides a solution to the visitor's objection, and the demonstration is undertaken. Here, negotiation of collaboration on the domain task keeps the engagement happening.

However, in part 6, the host's next offer is not accepted, not by conversational means, but by lack of response, an indication of disengagement. The host, who could have chosen to re-state his offer (with some persuasive comments), instead

takes a simpler negotiation tack and asks what the visitor would like to see. This aspect of the interaction illustrates the private assessment and planning which individual participants undertake in engagement. Essentially, it addresses the private question: what will keep us engaged? With the question directed to the visitor, the host also intends to re-engage the visitor in the interaction, which is minimally successful. The visitor responds but uses the response to indicate that the interaction is drawing to a close. The closing ritual [14], a disengagement event, is, in fact, odd given the overall interaction that has preceded it because the visitor does not follow the American cultural convention of expressing appreciation or at least offering a simple thanks for the activities performed by the host.

Part 0

<Visitor enters and is looking around the room when host notices visitor.>

Host: Hello, I'm the room host. Would you like me to show you around?

Part 1

Visitor: <Visitor ignores host and continues to look around>

What is this? <Visitor looks at and points to an object>

Host: That's a camera that allows a computer to see as well as a person to track people as they move around a room.

Part 2

Visitor: <looks at host> What does it see?

Host: Come over here <Host moves to the direction of the object of interest> and look at this monitor <points>. It will show you what the camera is seeing and what it identifies at each moment.

Part 3

Visitor: <follows host and then looks at monitor> Uh-huh. What are the boxes around the heads?

Host: The program identifies the most interesting things in the room--faces. That shows it is finding a face.

Visitor: oh, I see. Well, what else is there?

Part 4

Host: I can show you how to record a photo of yourself as the machine sees you.

Visitor: well, I don't know. Photos usually look bad.

Host: You can try it and throw away the results.

Part 5

Visitor: ok. What do I do?

Host: Stand before the camera.

Visitor: ok.

Host: When you are ready, say "photo now."

Visitor: ok. Photo now.

Host: Your picture has been taken. It will print on the printer outside this room.

Visitor: ok.

Part 6

Host: Let's take a look at the multi-level screen over there <points><then moves toward the screen>.

Visitor: <the visitor does not follow pointing and instead looks in a different direction for an extended period of time>

Host: <host notices and decides to see what the visitor is looking at.> Is there something else you want to see?

Visitor: No I think I've seen enough. Bye.

Host: ok. Bye.

FIGURE 4: Scenario for Room Hosting

While informal constructed scenarios can provide us with some features of engagement, a more solid basis of study of human hosting is needed. To that end I am currently collecting several videotaped interactions between human hosts and visitors in a natural hosting situation. In each session, the host is a lab researcher, while the visitor is a guest invited by the author to come and see the work going on in the lab. The host demonstrates new technology in a research lab to the visitor for between 28 and 50 minutes, with variation determined by the host and the equipment available.

6. ENGAGEMENT AMONG HUMAN HOSTS AND VISITORS

This section discusses engagement among people in hosting settings and draws on videotaped interactions collected at MERL.

Engagement is a collaboration that largely happens together with collaboration on a domain task. In effect, at every moment in the hosting interactions, there are two collaborations happening, one to tour a lab and the other to stay engaged with each other. While the first collaboration provides evidence for ongoing process of the second, it is not enough. Engagement appears to depend on many gestural actions as well as conversational comments. Furthermore, the initiation of engagement generally takes place before the domain task is explored, and engagement happens when there are not domain tasks being undertaken. Filling out this story is one of my ongoing research tasks.

In the hosting situations I have observed, engagement begins with two groups of actions. The first is the approach of the two participants accompanied by gaze at the other. Each notices the other. Then, the second group of actions takes place, namely those for opening ritual greetings [15], name introductions and hand shakes. Introductions and hand shakes are customary American rituals that follow greetings between strangers. For people, who are familiar with one another, engagement can begin with an approach, gaze at the potential partner and optionally a mere "hi." These brief descriptions of approach and opening rituals only begin to describe some of the variety in these activities. The salient point approach is that it is a collaboration because the two participants must achieve mutual notice. The critical point about openings is that an opening ritual is necessary to establish connection and hence is part of the engagement process.

All collaboration initiations can be thwarted, and the same is true of the collaboration for engagement, as is illustrated in the constructed scenario in Figure 4 in parts 0 and 1. However, in the videotaped sessions, no such failures occur, in large part, I surmise, due to the circumstances of the pre-agreement to the videotaped encounter.

Once connected, collaborators must find ways to stay connected. In relational only encounters, eye gaze, smiles and other gestures may suffice. However, for domain tasks, the collaborators begin the collaboration on the domain task. Collaborations always have a beginning phase where the goal is established, and proposing the domain task goal is a typical way to begin a domain collaboration. In the videotaped hosting activities, the participants have been set up in advance (as part of the arrangement to videotape them) to participate in hosting, so they do not need to establish this goal. They instead check that the hosting is still their goal and then proceed. The host performs his part by showing several demos of prototype systems. In three of the videotaped sessions, the host (who is the same person in all the sessions) utters some variant of "Let's go see some demos." This check on starting hosting is accompanied by looking at the visitor, smiles and in some cases, a sweep of the hand and arm, which appears to indicate either conveying a direction to go in or offering a presentation.

How do participants in a domain collaboration know that the engagement process is succeeding, that the participants are continuing to engage each other? When participants follow the shared recipes for a domain collaboration, they have evidence that the engagement is ongoing by virtue of the domain collaboration. However, many additional behaviors provide signals between the participants that they are still engaged. These signals are not necessary, but without them, the collaboration is a slow and inefficient enterprise and likely to breakdown because their actions can be interpreted as not continuing to be engaged or to participating in the domain task. Some of these signals are also essential to conversation for the same reason. The signals include:

- talking about the task,
- turn taking,
- timing of uptake of a turn,
- use of gaze at the speaker, gaze away for taking turns[17],
- use of gaze at speaker to track speaker gestures with objects,
- use of gaze by speaker or non-speaker to check on attention of other,
- hand gestures for pointing, iconic description, beat gestures, (see [19], [7]), and in the hosting setting, gestures associated with domain objects,
- head gestures (nods, shakes, sideways turns)
- body stance (facing at other, turning away, standing up when previously sitting and sitting down),
- facial gestures (not explored in this work but see [20]),
- non-linguistic auditory responses (snorts, laughs),
- social relational activities (telling jokes, role playing, supportive rejoinders).

Several of these signals have been investigated by other researchers, and hence only a few are noteworthy here. The timing of uptake of a turn concerns the delay between the end of one speaker's utterances and the next speaker's start at

speaking. It appears that participants have expectations about next speech occurring at an expected interval. They take variations to mean something. In particular, delays in uptake can be signals of disengagement or at least of conversational difficulties. Uptake delay may only be a signal of disengagement when other cues also indicate disengagement: looking away, walking away, or body stance away from the other participant.

In hosting situations, among many other circumstances, domain activities can require the use of hands (and other parts of the body) to operate equipment or display objects. In the videotaped sessions, the host often turns to a piece of equipment to operate it so that he can proceed with a demo. The visitors interpret these extended turns of attention to something as part of the domain collaboration, and hence do not take their existence as evidence that the performer is distracted from the task and the engagement. The important point here is that gestures related to operating equipment and object display when relevant to the domain task indicate that the collaboration is happening and no disengagement is occurring. When they are not relevant to the domain task, they could be indicators that the performer is no longer engaged, but further study is needed to gauge this circumstance.

Hosting activities seem to bring out what will be called *social relational activities*, that is, activities that are not essential for the domain task, but seem social in nature, and yet occur during it with some thread of relevance to the task. The hosts and visitors in the videotaped sessions tell humorous stories, offer rejoinders or replies that go beyond conveying that the information just offered was understood, and even take on role playing with the host and the objects being exhibited. Figure 5 contains a transcript of one hosting session in which the visitor and the host spontaneously play the part of two children using the special restaurant table that the host was demonstrating. The reader should note that their play is highly coordinated and interactive and is not discussed before it occurs. Role playing begins at 00 in the figure and ends at 17. [The host P has shown the visitor C how restaurant customers order food in an imaginary restaurant using an actual electronic table, and is just finishing an explanation of how wait staff might use the new electronic table to assist customers.] Note that utterances by P and C are labeled with their letter and a colon, while other material describes their body actions.

52: P left hand under table, right hand working table, head and eyes to table, bent over

C watching P.

P: so that way they can have special privileges to make different things happen

C nods at "privileges" and at "happen"

54: P turns head/eyes to C, raises hands up

C's had down, eyes on table

55: P moves away from C and table, raises hands and shakes them; moves totally away full upright

56: P: Uh and show you how the system all works

C: looks at P and nods

58: P sits down

P: ah

00: P: ah another aspect that we're

P rotates each hand in coordination

C looks at P

01: P: worried about

P shakes hands

02: P: you know

C nods

04: P: sort of a you know this would fit very nicely in a sort of theme restaurant

P looks at C; looks down

05: C: MM-hm

C looks at P, Nods at "MM-hm"

P: where you have lots of

06: P draws hands back to chest while looking at C

C: MM-hm

P: kids

C nods, looking at P

07: P: I have kids. If you brought them to a

P has hands out and open, looks down then at C

C still nods, looking at P

09: P: restaurant like this

P brings hands back to chest

C smiles and looks at P

10: P looks down; at "oh oh" lunges out with arm and (together points to table and looks at table)

P: they would go oh oh

11: C: one of these, one of these, one of these

C point at each phrase and looks at table

P laughs; does 3 pointings while looking at table

13: P: I want ice cream <point>, I want cake <point>

C: yes yes <simultaneous with "cake">

C points at “cake” looks at P, then brushes hair back

P looking at table

15: P: pizza <points>

P looking at table

C: Yes yes French fries <point>

C: looks at table as starts to point

16: P: one of everything

P pulls hands back and looks at C

C: yes

C: looks at P

17: P: and if the system just ordered {stuff} right then and there

P looks at C, hands out and {shakes}, shakes again after "there"

C looking at P; brushes hair

C: Right right (said after “there”)

20: P: you'd be in big trouble || <laughs>

P looking at C and shakes hands again in same way as before

C looking at P, nods at ||

23: C: But your kids would be ecstatic

C looking at P

P looking at C and puts hands in lap

Figure 5 Playtime example

One might argue that social relational activities occur to support other relational goals between participants in the engagement and domain task. In particular, in addition to achieving some task domain goals, many researchers claim that participants are managing their social encounters, their "social face," or their trust [21,22] in each other. Social relational activities may occur in support these concerns. This claim seems quite likely to this author. However, one need not take a stand the details of the social model for face management, or other interpersonal issues such as trust, in order to note that either indirectly as part of social management, or directly for engagement, the activities observed in the videotaped sessions contribute to maintaining the connection between the participants. Social relational activities such as the role playing one in Figure 5 allow participants to demonstrate they are socially connected to one another in a strong way. They are more than just paying attention to one another, especially to accomplish their domain goals. They actively seek ways to indicate to the other that they have some relation to each other. Telling jokes to amuse and entertain, conveying empathy in rejoinders or replies to stories, and playing roles are all means to indicate relational connection.

The challenge for participants in collaborations on domain tasks is to weave the relational connection into the domain collaboration. Alternatively participants can mark a break in the collaboration to tell stories or jokes. In the hosting

events I am studying, my subjects seem very facile at accomplishing the integration of relational connection and the domain collaboration.

All collaborations have an end condition either because the participants give up on the goal (c.f. [23]), or because the collaboration succeeds in achieving the desired goals. When collaboration on a domain task ends, participants can elect to negotiate an additional collaboration or refrain from doing so. When they so refrain, they then undertake to close the engagement. Their means to do so is presumably as varied as the rituals to begin engagement, but I observe the common patterns of pre-closing, expressing appreciation, saying goodbye, with an optional handshake, and then moving away from one another. Preclosings [24] convey that the end is coming. Expressing appreciation is part of a socially determined custom in the US (and many other cultures) when someone has performed a service for an individual. In my data, the visitor expresses appreciation, with acknowledgement of the host. Where the host has had some role in persuading the visitor to participate, the host may express appreciation as part of the preclosing. Moving away is a strong cue that the disengagement has taken place.

Collaboration on engagement transpires before, during and after collaboration on a domain task. One might want to argue that if that is the case, then more complex machinery is needed than that so far suggested in conversational models of collaboration (cf. [2],[3],[25]). I believe this is not the case because much of the collaboration on engagement is non-verbal behavior that simply conveys that collaboration is happening. For much of the collaboration to be engaged, no complex recipes are needed. The portions of engagement that require complex recipes are those of beginning and ending the engagement. Once some domain collaboration begins, engagement is maintained by the engagement signals discussed above, and while these signals must be planned for by the individual participants and recognized by each counterpart, they do not require much computational mechanism to keep going. In particular, no separate stack is needed to compute the effects of engagement because the engagement itself is not discussed as such once a domain task collaboration begins.

How does one account for the social relational behaviors discussed above in this way? While social relational behaviors also tell participants that their counterparts are engaged, they are enacted in the context of the domain task collaboration, and hence can function with the mechanisms for that purpose. Intermixing relational connection and domain collaboration are feasible in collaboration theory models. In particular, the goal of making a relational connection can be accomplished via actions that *contribute* to the goal of the domain collaboration. However, each collaborator must ascertain through presumably complex reasoning that the actions (and associated recipes) will serve their social goals as well as contribute to the domain goals. Hence they must choose actions that contribute to the ongoing engagement collaboration as well as the domain collaboration. Furthermore, they must undertake these goals jointly. The remarkable aspect of the playtime example is that the participants do not explicitly agree to demonstrate how kids will act in the restaurant. Rather the host, who has previously demonstrated other aspects of eating in the electronic restaurant, relates the problem of children in a restaurant and begins to demonstrate the matter when the visitor jumps in and participants jointly. The host accepts this participation by simply continuing his part in it. It appears on the surface that they are just jointly participating in the hosting goal, but at the same time they are also participating jointly in a social interaction. Working out the details of how hosting agents and visitors accomplish this second collaboration remains to be done.

Presumably not all social behaviors cannot be interpreted in the context of the domain task. Sometimes participants interrupt their collaboration to tell a story that is either not pertinent to the collaboration or while pertinent, somehow out of order. These stories are interruptions of the current collaboration and are understood as having some other conversational purpose. As interruptions, they also signal that engagement is happening as expected as long as the conversational details of the interruption operate to signal engagement. It is not interruptions in general that signal disengagement or a desire to move to disengage; it is failure of uptake of the interruption that signals disengagement possibilities. Thus, failure to uptake the interruption is clearly one means to signal a start towards disengagement.

Open Questions

The discussion above raises a number of questions that must be addressed in my ongoing work. First, in my data, the host and visitor often look away from each other at non-turn taking times, especially when they are displaying or using demo objects. They also look up or towards the other's face in the midst of demo activities. The SharedPlans collaboration model does not account for the kind of fine detail required to explain gaze changes, and nothing in the standard models of turn taking does either. How are we to account for these gaze changes as part of engagement? What drives collaborators to gaze away and back when undertaking actions with objects so that they and their collaborators remain engaged?

Second, in my data, participants do not always acknowledge or accept what another participant has said via linguistic expressions. Sometimes they use laughs or expressions of surprise (such as "wow") to indicate that they have heard and understood and even confirm what another has said. These verbal expressions are appropriate because they express appreciation of a joke, a humorous story or outcome of a demo. I am interested in the range and character of these phenomena as well as how they are generated and interpreted.

Third, this paper argues that much of engagement can be modeled as part of domain collaboration. However, a fuller computational picture is needed to explain how participants decide to signal engagement as continuing and how to recognize these signals.

7. A NEXT GENERATION MEL

While I pursue theory of human-human engagement, I am also interested in building new capabilities for Mel that are founded on human communication. To accomplish that, I will be combining hosting conversations with other research at MERL on face tracking and face recognition. These will make it possible to greet visitors in ways similar to human experience and may also allow us to make use of nodding and gaze change (though not what a human gazes at), which are important indicators of conversation for turn taking as well as expressions of disinterest. Building a robot that can detect faces and track them and notice when the face disengages for a brief or extended period of time provides a piece of the interactive behavior.

Another challenge for a robot host is to experiment with techniques for dealing with unexpected speech input. People, it is said, say that darndest things. Over time I plan to be able to collect data for what people say to a robot host and use it to train speech recognition engines. However, at the beginning, and every time the robot's abilities improve dramatically, I do not have reliable data for conversational purposes. To operate in these conditions, I will make some rough predictions of what people say and then need to use techniques for behaving when the interpretation of the user's utterances falls below a threshold of reliability. Techniques I have used in spoken-language systems in onscreen application [16] are not appropriate for 3D agents because they cannot be effectively presented to the human visitor. Instead I expect to use techniques that (1) border on Eliza-like behavior, and (2) use the conversational models in Collagen [12] to recover when the agent is not sure what has been said.

8. SUMMARY

Hosting activities are a natural and common interaction among humans and one that can be accommodated by human-robot interaction. Making the human-machine experience natural requires attention to engagement activities in conversation. Engagement is a collaborative activity that is accomplished in part through gestural means. Previous experiments with a non-autonomous robot that can converse and point provide a first level example of an engaged conversationalist. Through study of human-human hosting activities, new models of engagement for human-robot hosting interaction will provide us with a more detailed means of interacting between humans and robots.

9. ACKNOWLEDGMENTS

The authors wish to acknowledge the work of Myroslava Dzikovska and Paul Dietz on Mel, Neal Lesh, Charles Rich, and Jeff Rickel on Collagen and PACO.

10. REFERENCES

1. W. Burgard and A. B. Cremes, "The Interactive Museum Tour Guide Robot," *Proceedings of AAAI-98*, 11-18, AAAI Press, Menlo Park, CA, 1998.
2. B.J. Grosz and C. L. Sidner. "Plans for discourse," in *Intentions and Plans in Communication and Discourse*. P. Cohen, J. Morgan, and M.Pollack (eds.), MIT Press, 1990.
3. B. J. Grosz and S. Kraus. "Collaborative Plans for Complex Group Action," *Artificial Intelligence*, 86(2): 269-357, 1996.
4. C. L. Sidner. "An Artificial Discourse Language for Collaborative Negotiation," in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, MIT Press, Cambridge, MA, Vol.1: 814-819, 1994.
5. C. L. Sidner. "Negotiation in Collaborative Activity: A Discourse Analysis," *Knowledge-Based Systems*, Vol. 7, No. 4, 1994.
6. D. McNeill. *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press, Chicago, 1992.
7. W. L. Johnson, J. W. Rickel 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.
8. J. Cassell, Y. I. Nakano, T. W. Bickmore ,C. L. Sidner and C. Rich. "Non-Verbal Cues for Discourse Structure," *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July 2001.

9. J. Cassell, J. Sullivan, S. Prevost and E. Churchill, *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.
10. J. Rickel, N. Lesh, C. Rich, C. L. Sidner and A. Gertner, "Collaborative Discourse Theory as a Foundation for Tutorial Dialogue," To appear in the *Proceedings of Intelligent Tutorial Systems 2002*, July 2002.
11. 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.
12. C. Rich and C. L. Sidner, "COLLAGEN: A Collaboration Manager for Software Interface Agents," *User Modeling and User-Adapted Interaction*, Vol. 8, No. 3/4, 1998, pp. 315-350.
13. J. Cassell, H. Vilhjálmsson, and T. W. Bickmore, "BEAT: the Behavior Expression Animation Toolkit*" Proceedings of SIGGRAPH 2001*,pp. 477-486, ACM Press, New York, 2001.
14. C. L. Sidner and M. Dzikovska, "Hosting Activities: Experience with and Future Directions for a Robot Agent Host," in *Proceedings of the 2002 Conference on Intelligent User Interfaces*, ACM Press, New York, pp. 143-150, 2002.
15. H.H. Luger, "Some Aspects of Ritual Communication," *Journal of Pragmatics.* Vol. 7: 695-711, 1983.
16. C. L. Sidner and C. Forlines, "Subset Languages For Conversing With Collaborative Interface Agents," submitted to the *2002 International Conference on Spoken Language Systems*.
17. S. Duncan, "Some Signals and Rules for Taking Speaking Turns in Conversation," in *Nonverbal Communication*, S. Weitz (ed.), Oxford University Press, New York, 1974.
18. J. Cassell, T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan, "Human Conversation as a System Framework: Designing Embodied Conversational Agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), MIT Press, Cambridge, MA, 2000.
19. J. Cassell, , "Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.), MIT Press, Cambridge, MA, 2000.
20. C. Pelachaud, N. Badler, and M. Steedman, "Generating facial expressions for speech," *Cognitive Science*, 20(1):1-46, 1996.
21. Bickmore, T. and Cassell, J. "Relational Agents: A Model and Implementation of Building User Trust". *Proceedings of CHI-2001*, pp. 396-403, ACM Press, New York, 2001.
22. Katagiri, Y., Takahashi, T. and Takeuchi, Y. Social Persuasion in Human-Agent Interaction, *Second IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, IJCAI-2001*, Seattle, pp. 64-69, August, 2001.
23. P.Cohen and H. Levesque, "Persistence, Commitment and Intention," in *Intentions in Communication*, P. Cohen, J. Morgan and M.E. Pollack (eds.), MIT Press, Cambridge, MA, 1990.
24. E. Schegeloff and H. Sacks, "Opening up closings," *Semiotica*, 7:4, pp 289-327, 1973.
25. K. E. Lochbaum, , "A Collaborative Planning Model of Intentional Structure," *Computational Linguistics*, 24(4): 525-572, 1998.

Intelligent Interactive Information Presentation for Cultural Tourism

Oliviero Stock
ITC-irst
Via Sommarive, 18
38050 Povo TN, Italy,
stock@itc.it

Massimo Zancanaro
ITC-irst
Via Sommarive, 18
38050 Povo TN, Italy,
zancana@itc.it

Abstract

Cultural heritage appreciation is a privileged area of application for innovative, natural–language centred applications. In this paper we discuss some of the opportunities and challenges with a specific view of intelligent information presentation, that takes into account the user characteristics and behaviour and the context of the interaction. We make reference to the new PEACH project, aimed at exploring various technologies for enhancing the visitors' experience during their actual visit to a museum.

Introduction

Since the second half of the Eighties, we have considered cultural heritage appreciation a privileged area of application for innovative, natural–language centred applications. From the application point of view, we believe this is an area of high interest, as a) the "users" of cultural heritage increase in number at a fast pace; b) there is a natural request for a quality shift: from presentation of cultural heritage as a standard mass product, similar to supermarket goods, to a way to provide the single person with the possibility of acquiring information and understanding on things that interests him most, and to assist his cultural development; c) the way in which the cultural experience is carried on has not changed much for centuries; and especially the young seem to require novel modes of being exposed to the cultural material, so that they would engage and entertain them; d) for Italy and Mediterranean countries cultural heritage can be a natural resource that fuels economy (Minghetti *et al*, 2002); e) human-computer interface technology can have a decisive role in providing solutions for the individual.

From the research point of view in the first phase we have considered this as an opportunity for exploring ideas related to multimodal interfaces. The AlFresco System was a system that integrated language, pointing in input and language and images in output (Stock *et al*, 1997). But the main aspect is that it integrated in a coherent way different interaction attitudes: the goal-oriented language based modality and the navigation-oriented hypermedia modality. Well before the web era the AlFresco generalized communication act management approach was perhaps anticipating some of the present challenges of web interaction. Subsequently we have begun working on information presentation in the physical environment. This brought in a number of new issues and some constraints (see Stock, 2001). Ideas were experimented in two projects, Hyperaudio (Not *et al*, 1998) and the European project HIPS (Benelli *et al*, 1999).

We shall present here some new lines of research that we are now carrying on.

1. The PEACH Project

The PEACH (Personal Experience with Active Cultural Heritage) project objective is that of studying and experimenting with various advanced technologies that can enhance cultural heritage appreciation. The project, sponsored by the Trento Autonomous Province, is mainly based on IRST research, with important

contributions by the other two partners: DFKI and Giunti Multimedia.

The research activity focuses on two technology mainstreams, natural interactivity (encompassing natural language processing, perception, image understanding, intelligent systems etc.) and micro-sensory systems. Throughout the project, synergy and integration of different research sectors will be emphasized. Two general areas of research are highlighted:

- The study of techniques for individual-oriented information presentation: (i) use of formalisms and technologies derived from the field of natural language generation in order to build contextual presentations; (ii) use of speech and gestures as input and audio and animated characters as output; (iii) use of multi-agent architectures to provide suggestions and propose new topics.
- The study of techniques for multi-sensorial analysis and modeling of physical spaces—that is, the use of visual sensors such as video cameras, laser telemetry, infrared sensors, and audio sensors such as an array of microphones and ultrasonic signals for monitoring a dynamic environment, and collecting information about objects and about the environment for accurate virtual reconstruction.

The scope of the project is to significantly increase the quality of cultural heritage appreciation, in such a way as to transform passive objects into active ones that can be manipulated by the observer, and thus aiding to bridge the gap between our past, which they represent, and our future, of which they are the seeds. *Extended Appreciation* and *(Inter)active Objects* are facets of an underlying unifying vision called *Active Cultural Heritage.*

1.1 Extended Appreciation

The traditional modes of cultural heritage appreciation impose numerous limitations that are not always obvious. For instance, in observing a large statue, notwithstanding physical proximity, the observer most likely will be unable to capture details from every angle, as these may be too far from his/her viewpoint. In these cases, direct observation creates limitations that can be overcome with *augmented reality*, such as by using a palm computer to observe the details of the statue, taken from cameras or reconstructed in a virtual environment. Moreover, access to some objects can be difficult or even impossible for some visitors, such as disabled or elderly people. Creating an accurate virtual representation of the objects would extend fruition of the exhibit to these visitors as well.

In general, remote appreciation opens interesting possibilities, also for the study of an artefact that due to its fragile nature must be kept under restricted conditions and is thus not accessible to everyone. The possibility of interacting with an accurate virtual representation, allows the non-invasive access to a work of art in the manner, time, and place most appropriate for the visitor. Objects can be manipulated in an innovative, didactic, and fun way such as by modifying a work of art, partially or in its entirety.

1.2 (Inter)Active Objects

It is particularly important for the individual to be able to "navigate" an independent information course based on individually and dynamically created presentations. One of the scopes of the project is transcending a museum's restrictive environment by transforming a passive object observed by the visitor into an active subject capable of providing new information in a context-sensitive manner, a kind of *hyperlink* for accessing additional situation-specific information to be presented coherently.

Much of the technology for accessing information in the Internet today (for example, *adaptive user profiling, information promotion, database browsing, query by example*) has a natural place of application in this environment.

2. The Museum as a Smart Environment

A system that generates presentations of artworks in a museum must mould to the behaviour of a person visiting the museum. On the one hand, the system must facilitate movements within the space by (i) aiding the orientation of the user using appropriate linguistic support such as *"to your right you will see…"*; (ii) proposing suggestions about the best route for continuing the visit, such as with *"…along the same lines, the next room contains*

an interesting…". On the other hand, the system must be able to interpret the implicit intentions of the person's movements. For example, the prolonged observation of one object may be interpreted as a sign of interest.

A system of this type will be able to take into consideration the constrains posed by the environment in accessing information (e.g. objects in an adjacent room may be far if the two rooms are not connected) emphasising the emotional impact of seeing the "real" work of art. Such a system will also be able to affect the visitor's perception within the environment by attracting his/her attention to a particular work or detail; for instance, taking advantage of new technology such as the ability to superimpose computer-generated images to the real scene (via special transparent visors) or by generating verbal presentations based on rhetorical and persuasion-oriented strategies.

In this way, the museum visit is a full-fledged interaction between the visitor and the museum itself. In order to render possible this interaction, it is necessary that the museum - in fact the underlying information system - (i) knows the physical position of the visitor (and, as much as possible, his focus of visual attention); (ii) communicates individual information on the objects under exhibition—for instance through a portable device, 3D audio, or using a special wearable device that automatically superimposes generated images to the real scene; and (iii) receives requests from the visitor—verbally and/or through gestures.

A museum of this type will not be simply reactive, limiting itself to satisfy the questions of the visitors, but will also be proactive, explicitly providing unasked information; for instance, suggesting the visit to particularly interesting or famous objects, or allowing access to a "window" (e.g. a flat screen on the wall) that can deepen the study of the object under observation. Such suggestions can be made based on the observations of the person's behaviour, for example, the route chosen by a visitor or how much time is spent in front of a work, information noted about the user, such as age and culture, or considerations relative to the environment like rooms that are too crowded or that are temporarily closed. The system should be able to overhear the visitor's interaction (Busetta *et al*, 2001) and provide further suggestions on the basis of an internal model of priorities (for example, satisfying visitor's interests, fulfilling educational goals, or, perhaps, increasing museum bookshop's sales).

Another important dimension is that of attracting the young and keep them hooked to the cultural experience. With children the playful attitude is essential. We are conceiving new technology-based environments, with spoken interaction (see also the NICE project with a similar theme[1]), where as a side effect children will be motivated to look with attention and learn about the cultural heritage. One of the central aspect is the communication attitude. A humorous interaction is a key resource with children. The role of humor to keep attention, memorizing names and help creative thinking is well known. We are now beginning to see some concrete results in modeling some processes of humour production. To this end our initial work in computational humor will find a useful terrain of experimentation here (see Stock and Strapparava, 2002).

3. The Role of Information Presentation

According to (Bordegoni *et al*, 1997), a medium is a physical space in which perceptible entities are realized. Indeed, in a museum (as well as in a cultural city, an archaeological site, etc.) the most prominent medium is the environment itself. The main requirement for the presentation of information task is that of *integrating* the 'physical' experience, *without competing* with the original exhibit items for the visitor's attention.

From a multimedia point of view, this means that additional uses of the visual channel have to be carefully weighed. In this context, audio channel should play the major role in particular for language-based presentations, although the role of non-speech audio (e.g., music or ambient sounds) should also be investigated. Yet when a visual display is available (for example a PDA or a wall-size flat screen) images on the can be used support the visitor in the orientation task (3D or 2D images can used to support linguistic reference to physical objects). In this latter case, the visual channel is shared between the display

[1] http://www.niceproject.com

and the environment but the goal is still to provide support to environment-related tasks.
From a multimodal point of view, different modalities can be employed to focus the visitor's attention on specific objects or to stimulate interest in other exhibits. For example, the linguistic part of the presentation (through speech audio) can make large use of deictic and cross-modal expressions both with respect to space (such as *"here"*, *"in front of you"*, *"on the other side of the wall"*, etc.) and to time (*"as you have seen before"*, etc.) (Not and Zancanaro, 2000).

The peculiarity of the environment as a medium is its staticity: the system cannot directly intervene on this medium (i.e. the system cannot move or hide exhibits nor change the architecture of a room, as in virtual settings, at least without considering technology-based futuristic extensions.) Therefore, it may appear that a main limitation of the presentation system is the need to adapt the other media in service of the environment. Yet a multimodal approach can get round the staticity constraint in two ways at least:
a) Dynamically changing the user's perception of the environment: by exploiting augmented reality techniques (for example as described in Feiner, 1997) it is possible to overlay labels or other images on what the visitor is actually seeing. In this way, for example, the system can plan to highlight some relevant exhibits in the environment or shadow other less relevant ones. 3D audio effects or the selection of characteristic voices or sounds for audio messages can stimulate user's curiosity and attention (Marti et al., 2001). Yet a similar effect can be obtained by exploiting the power of language, as we did: language-based presentations can be carefully planned to attract the visitor's attention to more important exhibits and shadow less relevant ones. The simplest example: when in a visitor enters a room for the first time, she usually receives a general room presentation followed by one that directs the visitor's attention to the exhibit the system hypothesizes most interesting for her.
b) Changing the user's physical position: the system can induce the user to change her physical position either by a direct suggestion (e.g. "go to the other side of the room, the big fresco you'll see on the wall is La Maestà") or indirectly, for instance by introducing a new topic (e.g., "La Maestà, one of the absolute masterpieces of European Gothic painting, is located on the wall behind you").

Ultimately, the goal of such system is to support visitors in making their visiting experience meet their own interests; but in some cases a visitor should be encouraged not to miss some particular exhibits (for example, you cannot visit the Louvre for the first time and miss the Mona Lisa). Sometimes this task can be accomplished by direction giving, but there are other ways to promote exhibits: for example, by providing at the beginning of the visit a list of hotspots, or by planning a presentation that, in a coherent way, links the exhibit in sight to other ones through reference to the visitor's interest. More generally, further research is needed towards implementing pedagogically-motivated systems with meta-goals to pursue, educational strategies to follow and intentions to satisfy. In this respect, the interaction between the visitor and the system must evolve from simple interaction to full-fledged collaboration (for a discussion on this topic applied to cultural tourism see Stock, 2001).

4. Seven Challenges

The theme discussed here constitute a terrain where several areas of research can yield important contributions. We shall briefly review some challenges, relevant for language-oriented presentations.

Visitor Tracking. In our own experience after various investigations we have ended sticking to our initial choice - infrared emitters at fixed positions, sensors on mobile devices. This choice was also combined with a compass, but we are sure shortly there will be more interesting solutions available (e.g. ultrasounds). For the outdoor scenario, we need a combination of GPS and finer localization devices. Other techniques can be envisioned and should be further investigated, for example, beyond the physical position, it would be useful to know the direction of sight of the visitor. For the moment it requires head-mounted displays and complex

vision recognition hardware, but one can foresee a future where gaze detection may be possible with less obtrusive hardware, at least in structured and internally represented domains. Representations and reasoning on what is at sight need obviously to develop substantially.

Novel devices. Acoustic output has been shown to be here preferable over written language which can be usefully exploited for highlights and follow-ups. Yet improvements on high quality graphics on a small device would be highly appreciated since pictures have been shown to be very helpful in signalling references to objects in the physical space.
Wearable devices and head-mounted displays can play a role in specific settings. In particular, head-mounted displays can be very useful if coupled with a technology that can overlay computer generated graphics to real scenes (see Feiner *et al*, 1997). This technique is particularly interesting for archaeological sites where the visitor would be able to "see" the buildings as they were originally.
But often the best device for these kind of applications is no device at all. Speech recognition in the environment coupled with "spatialized" audio would allow the visitors to experience multisensory and unobtrusive interaction with the environment. The "narration" must develop with individual-oriented characteristics and at personal times, so it cannot just be produced once for all by physically dislocated sound sources.

Expressing Space and time reference. We need our systems to be able to reason about where things are, what kind of spatial entity they are, how they look like from a given position, how best the visitor can reach them (Baus *et al*, 2002), when they will appear. For example, the system should be able to instruct the visitor to "reach the room at the end of this corridor" rather than "go forward 10 meters and then turn left". There is a substantial tradition in AI dealing with qualitative temporal reasoning and a somehow less extended one dealing with spatial reasoning (Stock, 1997). Representations must provide us with material at the right level of detail so that we can properly express it in words. Of course we need also that the language we produce is sophisticated in the proper use of word characteristics, for time taking into account concepts like word aspect and tense, or, for space, for example, being able to choose specific spatial prepositions. A newer important element in research is qualitative modelling of movement (see Galton, 1997), particularly relevant here, as we have seen that movement is the most relevant input modality, strictly coupled with our suggested medium - the environment.

Beyond descriptive texts. Perhaps the biggest challenge is concerned with keeping the attention of the user high and granting a long term memory effect. We need to be able to device techniques of material presentation that hook the visitor, that continuously build the necessary anticipation and release tension. The "story" (we mean the multimodal story that includes language, graphics and the visible physical environment) must be entertaining, and it should include mechanisms of surprise. The expectation sometimes must be contradicted and this contrast will help in keeping the attention and memorizing the situation. A typical mechanism of this kind is at the basis of various forms of computational humor (see Stock, 1996). Especially with children humor (and play) can be a powerful means for keeping them interested. Another aspect where much more research is needed is concerned with mechanisms of persuasion: i.e. how we can build rhetorical mechanisms aimed at the goal that the hearer/experiencer adopts desired beliefs and goals. It is not only a matter of rational argumentation, a field a bit more developed, but an integration of various aspects, including some modelling of affect. At the end our philosophy is that the user is responsible for what she does and hence for the material that is presented to her, but yet through the presentation some specific goals of the museum curator can be submitted for adoption.

New visit modalities. The advent of technology opens the way to new modalities of visit, particularly important with children. A treasure hunt is an obvious example, where the external goals cause the innocent visitor to look for details and come across many different exhibits with "artificially" induced attention.
An easier development is that at the end of a visit, a report of the visit is produced

electronically, available for successive elaboration. For instance it will allow the user to re-follow on a virtual environment what she has seen and to explore related material at a deeper level through added hyperlinks.

Support for group visits. A relevant percentage of visitors come to the museum together with other people. For natural science museums, the typical case is a parent with children, for art museums it is the group of friends. The group dimension is largely unexplored: how best can a family (or other group) be exposed in individually different manners to the material in the environment so that they discuss what they have seen and have a conversation that adds to their individual experience, bringing in new interests and curiosity?
Only limited research is devoted to group visit (see for example, Woodruff *et al*, 2001) and most issues are still open. Of course, we can envisage a big difference in the parent-child case with respect to the friends scenario. Another interesting issue can be the study of dynamic grouping, for example when grouping extends over time (see for example, Rahlff, 1999) or is dynamically created during the visit.

Experimental evaluation. The most enthusiastic comments of users of these kind of systems (Marti and Lanzi, 2001) regard the possibility to move freely during the visit while being assisted by the dynamic guide. The visitors felt comfortable in listening at descriptions without interacting too much with the PDA interface, mainly used in case of poor performance by the system (delay in loading a presentation, lack of information etc.). A feature that was especially appreciated was how information came tailored to the context. The visitors recognized the capability of the tourist guide to follow their movements offering appropriate and overall coherent information at the right moment.
However, our community has not become sophisticated enough in evaluating mobile systems for a cultural task. What we really need are techniques as powerful as the Wizard of Oz (simulation by hidden humans of systems that at least in part do not exist yet, and observation of user behaviour with the new means) so that the results will really help decide on the specific design choice. Equally important, as in any educational environment, is to evaluate retention of concepts and vividness of memory after time (hours, weeks, years).

The PEACH project, started recently, will deliver its results in a three year period with experimentation at the Castello del Buonconsiglio in Trento, with focus on the famous frescoes of Torre Aquila. DFKI in particular will also experiment at the Voelklinger Huette, a world cultural heritage site dedicated to iron and steel industry near Saarbruecken.

References

Stock. O. (2001) *Language-Based Interfaces and Their Application for Cultural Tourism. AI Magazine*, Vol.22, n.1, 2001, pp. 85-97, American Association for Artificial Intelligence, Menlo Park, CA.

Minghetti, V., Moretti, A., Micelli, S. (2002) *Reengineering the Museum's Role in the Tourism Value Chain: Towards an IT Business Model*. In Werthner, H. (Ed.), Information Technology & Tourism. New York: Cognitant Communication Corporation. 4(2), 131-143.

Stock O., Strapparava C., Zancanaro M. (1997) *Explorations in an Environment for Natural Language Multimodal Information Access* in M. Maybury (ed.) Intelligent Multimodal Information Retrieval. AAAI Press, Menlo Park, Ca./MIT Press, Cambridge, Mass.

Not E., Petrelli D., Sarini M., Stock O., Strapparava C., Zancanaro M. (1998) *Hypernavigation in the Physical Space: Adapting Presentations to the User and the Situational Context*. In New Review of Hypermedia and Multimedia, vol. 4.

Benelli, G., Bianchi, A., Marti, P., Not, E., Sennati D. (1999) *HIPS: Hyper-Interaction within the Physical Space*. In Proceedings of IEEE Multimedia System '99, International Conference on Multimedia Computing and Systems, Firenze

Busetta P., Serafini L., Singh D., Zini F. (2001) *Extending Multi-Agent Cooperation by Overhearing*. In Proceedings of the 9th International Conference on Cooperative Information Systems - CoopIS2001. Lectures in Computer Sciences vol. 2172, Trento, September.

Stock O. and Strapparava C. (2002) *Humorous Agent for Humorous Acronyms: The HAHAcronym Project*. Proceedings of the Fools' Day Workshop on Computational Humor, TLWT-20, Trento

Bordegoni M., Faconti G. Maybury M.T. Rist T. Ruggeri S. Trahanias, P. Wilson, M. (1997) *A Standard Reference Model for Intelligent Multimedia Presentation Systems*. Computer Standards and Interfaces, 18, pp. 477--496, 1997

Not E. and Zancanaro M. (2000) *The MacroNode Approach: mediating between adaptive and dynamic hypermedia*. In Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems, *AH'2000*, Trento, August.

Feiner S., MacIntyre B., Hollerer T., Webster A. (1997) *A Touring Machine: Prototyping 3D Mobile Augmented Reality Systems for Exploring the Urban Environment*. In Proceedings of ISWC 97 (Int. Symp. on Wearable Computing). Cambridge, MA. October.

Marti P., Gabrielli L., Pucci F. (2001) *Situated Interaction in Art*. Personal Technologies, 5:71-74.

Baus J., Krüger A. and Wahlster W. (2002) *A resource-adaptive mobile navigation system.* Proceedings of IUI2002: International Conference on Intelligent User Interfaces 2002,ACM Press.

Stock O. ed. (1997) Spatial and Temporal Reasoning. Kluwer Academic Publishers, Dordrecht.

Galton A. (1997) *Space, Time and Movement*. In O. Stock (ed.) Spatial and Temporal Reasoning. Kluwer Academic Publishers, Dordrecht.

Rahlff O.-W. (1999) *Tracing Father*. In Proceedings of i3 Annual Conference, Siena, October .

Marti P. and Lanzi P. (2001) *I enjoyed that this much!* A technique for measuring usability in leisure-oriented applications, In Joanna Bawa & Pat Dorazio, The Usability Business: Making the Web Work.

Stock. O. (1996) *Password Swordfish: Humour in the Interface*. In Proceedings of the International Workshop on Computational Humour, TWLT-12 Enschede.

Woodruff A.., Aoki P.M., Hurst A., Szymanski M. (2001) *The Guidebook, the Friend, and the Room: Visitor Experience in a Historic House*. Extended Abstract, ACM SIGCHI Conf. on Human Factors in Computing Systems, Seattle, WA, March.