Speech and Language Interaction in a (Virtual) Cultural Theatre

Anton Nijholt, Arjan van Hessen & J. Hulstijn
University of Twente, Centre of Telematics and Information Technology
PO Box 217, 7500 AE Enschede, the Netherlands

ABSTRACT

In this paper we survey research on interaction modalities in a theatre information and booking system. The original system, developed in the context of the Parlevink Research project, allowed natural language interaction using the keyboard. However, current research has made it possible to allow other modalities. Rather than presenting questions and answers on the screen, current research allows output using speech synthesis in combination with screen-based information. Moreover, since the system has been embedded in a virtual reality environment, which models a theatre, it has become interesting to investigate possible interactions between users and theatre agents. Presently we distinguish between an information and transaction agent which provides the user with information about performances and a navigation agent which helps the user to navigate in the virtual world. Possible interactions between these agents have not yet been investigated.

1. Introduction

Traditional human-computer interaction is mostly concerned with interfaces for professional users in professional situations. We sense a certain reluctance of the traditional research community to pay attention to non-professional users in non-professional environments and with aims that are not necessarily directed towards efficient interaction or increasing productivity. They warn against 3-D visualisation or the use of agents. Even well known Web design guru's like Jakob Nielsen warn us against advanced Web design since, as he supposes, Web users are or become, conservative in their use of the Web.[1] We don't really disagree with his observations about the majority of users, but we also believe this majority behaviour will change. This change can be predicted by looking at the exploration behaviour of children and students, but also by looking at the behaviour of potential users in realistic situations: shop for fun, leaf a brochure, look around, etc.

In this paper we present our research on developing an environment in which users can display different behaviours and have goals that emerge during the interaction with our environment. Users who, for example, decide they want to spend an evening outside their home and, while having certain preferences, cannot say in advance where exactly they want to go, whether they first want to have a diner, whether they want to go to a movie, theatre, or to opera, what time they want to go, etc. During the interaction with a system like ours, both goals, possibilities and the way they influence each other become clear, both to the user as to, hopefully, the system. One way to support such users is the use of virtual reality.

The first research issues in Virtual Reality dealt with simulation, training and design. Since then, we have seen the development of virtual classrooms, virtual meeting places and virtual environments that support group working (cf. Rogers, 1995 or the current European I3 projects). The introduction of VRML (Virtual Reality Modelling Language) has made it possible to have virtual

[1] Cf. Jakob Nielsen in his Alertbox for March 22, 1998: "Web users are conservative: they don't want inconsistent site designs or fancy pages filled with graphic gimmicks and animations. And they frequently don't have the latest client software available. As a result, Web designers have to be conservative in what they show to users: page design must be conservative and minimalist."

reality in Web pages and, as a consequence, in Web interfaces.

WorldWideWeb allows interactions and transactions through Web pages using speech and language, either by artificial or by live agents, image interpretation and generation, and, obviously, the more traditional ways of presenting explicitly pre-defined information by allowing users access to text, tables, figures, pictures, audio, animation and video. In a task- or domain-restricted way of interaction current technology allows the recognition and interpretation of rather natural speech and language in dialogues. However, using VRML, rather than the current two-dimensional web-pages, many interesting parts of the Web will become three-dimensional. This allows for the building of virtual worlds inhabited by interacting user and task agents with which the user can interact using different types of modalities, including speech, language and image interpretation and generation. In these worlds agents can work on behalf of users, hence, human computer interaction will make use of 'indirect management', rather than interacting through direct manipulation of data by users.

In this paper we present research on interaction in a virtual theatre, a realistic model of an existing theatre in the Netherlands. This so-called 'Muziekcentrum' offers its potential visitors information about performances (music, cabaret, theatre, opera) by means of a brochure that is published once a year. In addition to this yearly brochure it is possible to get information at an information desk in the theatre (during office hours), to get (more recent and updated) information by phone (either by talking to a theatre employee or by using Interactive Voice Response technology) and to get information from local daily and weekly papers and monthly announcements issued by the theatre. The central database of the theatre holds the information that is available at the beginning of the 'theatre season'. Our aim is to make this information about theatre and performances much more accessible to the general audience by using multi-modal accessible multimedia web pages.

From a more global point of view our research topics are:

- Modelling effective interactions between humans and computers, with an emphasis on the use of speech and language
- Commercial transactions, (local, regional) governmental information services, education and entertainment in virtual environments
- Web-based information and transaction services, in particular interactions in virtual environments

We believe that the case study we have chosen allows several illustrations of these topics. Clearly, our domain of application is much more general than current systems concerned with air-travel, public transport, telephone directory systems, etc. To give a simple example, when a user asks if there is a performance this evening of a particular artist or a particular genre the system can consult the database and conclude that the answer is no and then generate this answer. In current train travel information systems a negative answer might be considered as acceptable. In our domain, independently whether we want to anticipate users' expectations as much as possible or whether we want to increase the selling of tickets, it seems to be useful and natural to present information about other performances at the same day (in case the user really wants to go that particular evening) or other performances during that week (if the user really wants to go to a particular genre), or later that year (if the user really wants to go to a particular performer). In addition, our dialogues involve transactions. Such dialogues display a more complex structure than mere inquiry or advisory dialogues. Two tasks are executed in parallel: obtaining information and ordering products (Jönsson, 1993). In our corpus similarly complex behaviour can be found. Users browse, inquire and retract previous choices, for instance when tickets are too expensive.

Hence, we allow interactions that are much less goal-directed than in usual spoken dialogue systems. During the dialogue the user will determine and update his or her goals. Of course, it is also the environment in which the dialogues are embedded and the possibility to explore environment and interaction modalities that invites users to browse through the available information just like leaf through a brochure.

2. Building a Virtual Theatre

Our virtual theatre has been built according to the design drawings made by the architects of the building. Part of the building has been realised by converting AutoCAD drawings to VRML97. Video recordings and photographs have been used to add 'textures' to walls, floors, etc. Sensor nodes in the virtual environment activate animations (opening doors) or start events (entering a dialogue mode, playing music, moving spotlights, etc.).[2] Visitors can explore the environment of the building, hear the carillon of a nearby church, look at a neighbouring pub and movie theatre, etc. They can enter the theatre and walk around, visit the hall, admire the paintings on the walls, enter the main performance hall, go to the balconies and, take a seat in order to get a view of the stage from that particular location. Information about today's performances is available on a blackboard that is automatically updated using information from the database with performances. In addition, as may be expected, visitors may go to the information desk in the theatre, see previews and start a dialogue with an information and transaction agent called 'Karin'. The first version of Karin looked like other standard avatars available on World Wide Web. The second version, available in a prototype of the system, has a more human-like appearance such that visitors feel happy to talk with her and ask about performances and make reservations.

One may argue the necessity of this realistic modelling of the theatre, its environment and its information and transaction services. We have taken the point of view that (potential) visitors are interested in or are already familiar with the physical appearance of this theatre. Inside the virtual building there should be a mix of reality (entrance, walls, paintings, desks, stages, rooms, etc.) and new, non-traditional, possibilities for virtual visitors to make use of interaction, information, transaction and navigation services that extend the present services of the theatre. User evaluation studies (to be performed in the beginning of 1999) will probably make clear how much need there is to have a reasonable realistic representation of the theatre information and transaction service interactions that are offered at this moment.

[2] Presently the world has been made for Cosmo Player 2.0. Although any WWW user can visit the theatre, in order to have a reasonable performance a fast PC with 233 Mhz II Pentium processor, 64 MB RAM and a 3D accelerator video card is essential for an acceptable frame rate.

3. Interactions in the Virtual Theatre

With the coming of age of language and speech technology and the development of tools to build applications in Information and Communication Technology, new ways for human-computer interaction have been developed. Current computer technology, and, to a lesser extent, progress in the performance of algorithms for Automatic Speech Recognition (ASR), Text-to-Speech (TTS) synthesising, and Natural Language Processing (NLP), has made it possible to build complex, reasonably well-performing dialogue systems with which people can communicate to get information, make reservations, and order products. One may expect that in the near future speech and language interaction will become generally applied and accepted, not because of trail-blazing research results but rather of careful application of current technology and results obtained from research on dialogue management and multi-modal interaction.

The research reported here concerns the development of an environment (the virtual theatre) in which we can experiment with different modalities (and their combinations) for interaction. That is, our aim is twofold: to develop models for multi-modal interaction and, by involving potential users in our experiments, match user characteristics, interaction modalities and functional properties. In previous years we have done natural language processing research devoted to robustness, parsing, dialogue modelling and dialogue management. Rather than doing research on speech technology we are able, by embedding commercially available systems in multi-modal environments, to use speech recognition and speech synthesis systems that have been developed elsewhere.

3.1 A Navigational Agent

Clearly, the WWW-based virtual theatre we are developing allows navigation input through keyboard and mouse. Such input allows the user to move and to rotate, to jump from one location to an other, to interact with objects and to trigger

them. In addition, a navigation agent has been developed that allows the user to explore the environment and to interact with objects in this environment by means of speech commands. Obviously, we do not want completely separated modalities. It should be left to the user to choose between the interacting means or to use both. A smooth integration of the pointing devices and speech in a virtual environment requires means to resolve deictic references that occur in the interaction. The navigation agent should be able to reason about the geometry of the virtual world in which it moves.

The current version of the navigational agent is not really conversational. Straightforward speech commands make it possible for the user to explore the virtual environment. That is, visit certain locations, turn left, enter a room, go back to a previous location, walk around an object, trigger actions, etc. Apart from such navigation commands speech input allows deictic clarifications and motion modifications. Navigation also requires that names have to be associated with the different parts of the building, the objects and the agents, which can be found inside of it. Clearly, users may use different words to designate them, including implicit references that have to be resolved in a reasoning process.

3.2 An Information and Transaction Agent

As mentioned before, a second agent called Karin allows a natural language dialogue with the system about performances, artists, dates, prices, etc. Karin wants to give information and to sell tickets. Karin is fed from a database that contains all the information about performances in our local theatre. Developing skills for Karin, in this particular environment, is one of the main aims of our research project. This research fits in a context of much more general 'intelligent' (web-based) information and transaction services. Therefore we distinguish two approaches in our research.

The first approach aims at doing long-term research on interaction modelling and management. The emphasis is on speech and language interaction. Research is performed on syntax, robustness, unification, use of statistics, semantics, pragmatics, (multi-modal) dialogue modelling and management, NLP architectures, object-oriented design and implementation issues, etc. In addition, there are general computer science issues of design, implementation and evaluation of human-computer interaction systems

The second approach deals with domain and task dependent issues: how can we make use of existing information sources, how can we present this information, what kinds of modalities are useful (for specific tasks, domains, users), how do we integrate different modalities, how do we model the knowledge about domain and users and the knowledge of the users of our system, how can we evaluate our prototypes, how do we involve potential users in the design of their systems and what can be said about societal implications of the systems we are developing. In general we try to solve these problems in our particular context by using theory and ideas developed elsewhere.

In the next paragraphs we introduce our present natural language dialogue system and the way it has been embedded in a virtual environment. In a next section we return to (future) research on user modelling in our domain of application.

Our current version of the dialogue system of which Karin is the face is called THIS v1.0 (Theatre Information System). The approach used in THIS can be summarised as rewrite and understand. User utterances are simplified using a great number of rewrite rules. The resulting simple sentences are parsed. The output can be interpreted as a request of a certain type. System response actions are coded as procedures that need certain arguments. Missing arguments are subsequently asked for. The system is modular, where each 'module' corresponds to a topic in the task domain. There are also modules for each step in the understanding process: the rewriter, the recogniser and the dialogue manager. The rewrite step can be broken down into a number of consecutive steps that each deal with particular types of information, such as names, dates and titles. The dialogue manager initiates the first system utterance and goes on to call the rewriter and recogniser process on the user's response. Also, it provides an interface with the database management system (DBMS). Queries to the database are represented using a standard query language like SQL. Results of queries are represented as bindings to variables which are stored in the global data-structure called context. The arguments for

the action are dug out by the dedicated parser, associated with the category. All arguments that are not to be found in the utterance are asked for explicitly.

The task model is implemented as a set of actions or procedures. Parameters are implemented as arguments of a certain type to the procedures. When a user request is recognized, the corresponding procedure is called. Actions are normally specific to a certain topic or category. Finally, what we need is a set of system responses for each situation that might arise. Again, responses are organized around category. So, given a corpus of dialogues it is possible to identify categories of utterances with the same topic. Such categories can be used in subsequent stages of the design process to guide the construction of resources. More information about the approach in this natural language dialogue system can be found in Lie et al. (1998).

Presently the input to Karin is keyboard-driven natural language and the output is both screen and speech based. In development is an utterance generation module for information and transaction. Based on the most recent user utterance, on the context and on the database, the system has to decide on a response action, consisting of database manipulation and dialogue acts. Long term actions are planned. A reservation for instance, involves subactions for performance selection, discussion of price and number of tickets and confirmation of the transaction. Each dialogue act is put into words by the utterance generation module. It determines the utterance-structure, wording, and prosody of each system utterance. We make use of the *Fluency* text-to-speech package for Dutch. The utterance generation makes use of a list of prosodically annotated utterance templates. Templates contain gaps to be filled with attribute-value pairs that are annotated with syntactic and lexical features. The module respects observations from topic and focus theories on the discourse effect of Dutch word-order and intonation (Van Deemter et al., 1994; Dirksen, 1992).

Each template is determined by three parameters: *utterance type*, *given* information and *wanted* or *new* information. There are basic dialogue control acts, like *bye* or *yes*. There are templates that combined with the given and new parameters produce assertions with normal word-order and standard intonation. Given information is usually de-accented, expressed by a pronoun or even left out (cf. Rats 1996). New information is accented and generally appears at the end of the utterance. Often we produce only short answers to the user's query. There are also templates for yes/no questions and alternative questions. These usually have the inverted word-order and a raising intonation, although some clarification questions are more naturally asked with a declarative word-order. Templates for wh-questions combine with the information items in the 'wanted' parameter. It is possible to generate utterances that express *contrastive* intonation. This is especially useful when the user has a number of alternative options to choose from. Legally, transactions involve obligations of both parties: the system should provide all relevant information for the user to make a fair choice and the user is bound by the transaction once it has been concluded. Our task is to make users feel committed, even though it is 'only a computer' they have been talking to. Utterance generation plays a major role in this. In the near future we will experiment with spoken natural language input and output for the information and transaction service functions of the system. Again, we hope that the environment and the other modalities we allow, will make it possible to use current imperfect speech recognition and synthesis technology.

3.3 Improving the Skills of the Agents

We slowly extend and improve the interaction and navigation intelligence of our navigation agent. For example, if the agent knows it can not do what the user wants or it does not understand the user, it can explain its shortcomings, or the (im)possibilities of exploring the theatre in the way the user wishes, and it can suggest alternatives that come close to what the user wants. Clearly, the navigation agent should not only be able to interpret speech commands and questions in the context of the part of the virtual theatre that is displayed on the screen, but also in the context of the 'browsing history' to which speech commands can refer. Obviously, a browsing history leads to expectations and users may have implicit references to these expectations. Rather than having an agent that understands commands

and references in these commands that concern only navigation, it becomes necessary to allow the conversion of a command-driven agent to an agent that can maintain a (goal-oriented) dialogue or an agent that can maintain a useful conversation. Moreover, but outside the scope of our research at this moment, we need to model and use interaction knowledge available from Karin, our information and transaction agent, and interaction knowledge available from other visitors or their agents in the virtual environment. A next step might be that due to the interactions (with users or other agents) the appearance of the virtual world and the knowledge distributed in this world changes. That is, interactions lead to updates of knowledge available in agents and (other) objects which in turn may lead to feedback to the user or visitor of the virtual theatre.

4. Finding Information in a Virtual Environment

4.1 Chatting and Browsing

It is useful to distinguish between implicit and explicit presentation of information. To give an example, in the virtual world we are presenting, all kinds of information become available by browsing this world. If this world is transparent and accessible, rather than asking explicit questions to information agents people will 'walk around' and see whether they themselves can find answers to their questions. To avoid misunderstanding, we do not necessarily assume that when people enter our world they have questions. Questions can emerge because visitors get interested in our world.

We think this an important property of the environments we are developing. In general, telephone-based spoken dialogue systems are goal-directed. A caller wants to know when his train leaves or arrives. He wants a connection with the chairman of the speech & music department because he has certain questions about their research program. In our point of view a visitor of our web-based cultural pages has the possibility to explore the environment and to start interactions with all kinds of agents. Agents that help to navigate in this environment, agents that allow access to information offered in this environment. Agents that give information about other visitors and their opinions about theatre and performances. Commercial agents that want to sell tickets. User agents that try to find out whether something interesting is going on about which their creator has to be informed. Etc.

One may expect that in such an environment many users will display both a chatting & browsing behaviour and a goal-directed behaviour. Obviously, this has consequences for the way the system has to manage the dialogue with the user. Presently we are building the environment that invites such dialogues. From dialogues that will be collected and from user evaluation studies we will try to determine guidelines for an improved design.

4.2 Talking Faces

In our experimental system Karin's spoken dialogue contribution is presented by visual speech, that is, a 'talking face' on the screen, embedded in the virtual world, mouths the systems questions and (not too complex) responses. If necessary, part of the information is given in the form of a table on the screen.

The virtual face that has been designed allows animation of lip and face movements and animation of some simple face expressions. In order to have speech-image synchronisation 3D images of (a limited number of) visemes are called when corresponding phonemes have to be spoken. It should be noted that we are talking about a web-based application. This requires the solution of several technical problems dealing with sending (Real Audio) sound files and commands and synchronising them.

It has become clear from several studies that people engage in social behaviour toward machines. It is also well known that users respond differently to different 'computer personalities'. It is possible to influence the user's willingness to continue working even if the system's performance is far from perfect. They can be made to enjoy the interaction, they can be made to perform better, etc., all depending on the way the interface and the interaction strategy has been designed. It also makes a difference to interact with a talking face display or with a text display. Finally, the facial appearance and the expression of the face matters. From all these observations (see

Friedman, 1997 for details) we conclude that introducing a talking face can help to make interactions more natural and shortcomings of the technology more acceptable to users.

5. Future work on User Modelling

Obviously, we can learn about the user during his or her interaction with our system. In order to have a satisfactory dialogue it may indeed be necessary to keep track of the user's beliefs and beliefs revisions during the dialogue and to design and maintain a model of the user embedded in the dialogue discourse. We think that it is also important to have user profiles and to have sometimes detailed knowledge of a particular user in order to provide the appropriate interaction and information service. Interaction identification, task identification, user profile identification and individual user identification are among the tasks that have to be performed in our virtual environment. Hence, our starting point is (again) a little different from what is usual in designing interfaces from a computational linguistics or artificial intelligence point of view. We would like to look at technology that is becoming available in the context of call centre technology, electronic profiles (e.g., Firefly's passport technology) or customer centric systems and that aims at improving (personalised) relations with clients, that aims at supporting customer service and that aims at customising offerings to suit individual interests.

6. References

S.P. van de Burgt, T. Andernach, H. Kloosterman, R. Bos & A. Nijholt. Building dialogue systems that sell. Proceedings *NLP and Industrial Applications*, New Brunswick, June 1996, 41-46.

C. van Deemter et al. Generation of spoken monologues by means of templates. In *Speech and Language Engineering*, TWLT8, University of Twente, 1994.

A. Dirksen. Accenting and deaccenting, a declarative approach. In *COLING'92*, Nantes, 1992.

H. ter Doest. A corpus-based probabilistic unification grammar for the SCHISMA task domain. To appear in: Proceedings *Esslli Workshop*, Saarbrücken, 1998.

B. Friedman (ed.). *Human Values and the Design of Computer Technology*. CSLI Publications, Cambridge University Press, 1997.

J. Hulstijn et al. Utterance Generation for Transaction Dialogues. Proceedings *International Conference on Spoken Language Processing*, Sydney, Australia, 1998, to appear.

A. Jönsson. *Dialogue Management for Natural Language Interfaces*. PhD thesis, Linköping University, 1993.

B. Laurel. *The Computer as Theatre*. Academic Press, New York, 1990.

D.H. Lie, J. Hulstijn, R. op den Akker & A. Nijholt. *Rewrite and Understand:* a transformational approach to natural language understanding in a dialogue system. These proceedings, 1998.

D.W. Massaro. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. A Bradford Book, The MIT Press, 1997.

P. Nugues, C. Godéreaux, P.-O. El Guedj & F. Revolta. A Conversational Agent to Navigate in Virtual Worlds. Proceedings *Dialogue Management in Natural Language Systems.* Twente Workshop on Language Technology 11, June 1996, 23-33.

M. Rats. *Topic Management in Information Dialogues*. PhD thesis, ITK, Tilburg, 1996.

A. Rogers. Virtual reality – the new media? *British Telecom Technology Journal*, Volume 13, No 4, October 1995.