

Towards multi-modal interactions in virtual environments: a case study²

ABSTRACT: We discuss interaction modalities in web-based virtual environments. We argue that the language behaviour of users in virtual environments differs essentially from language behaviour in the context of other spoken dialogue systems. Allowing different modalities will also ask for a model that allows user and system to expect that knowledge obtained from previous interactions is available to whatever 'agents' are addressed by the user. We survey our research on interaction modalities in a theatre information and booking system. A first version of the system allowed natural language interaction using keyboard only. Current research has made it possible to allow other modalities. Rather than presenting questions and answers on the screen, we allow output using speech synthesis in combination with screen-based information. Since the system has been embedded in a virtual reality environment, which models a theatre, it becomes possible to investigate interactions between users and theatre agents.

1. Introduction

Traditional human-computer interaction is 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, animation or the use of agents. Web design guru's warn us against advanced Web design since, as they suppose, Web users are or become, conservative in their use of the Web. It is not that we disagree with their observations about the majority of users, but we 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 dinner, whether they want to go to a movie, theatre, or to opera, what time they want to go, etc. During the interaction, 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.

¹With contributions by Joris Hulstijn, Arjan van Hessen and Mathieu van de Berk (Parlevink Research Group).

²Magister conference in this Symposium.

In this paper we present research on interaction in a virtual theatre, a realistic model of an existing theatre. The real 'Muziekcentrum' offers potential visitors information about performances (music, cabaret, theatre, opera) by means of a brochure that is published once a year. 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 much more accessible by using multi-modal accessible multi-media 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 and global) information services, education and entertainment in virtual environments
- Web-based information and transaction services, in particular interactions in virtual environments

The case study we have chosen allows the illustration 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. For 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 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 is 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 dialogues. Two tasks are executed in parallel: obtaining information and ordering. In our corpus complex behaviour related to these tasks can be found. Users browse, inquire and retract previous choices, for instance when tickets are too expensive. Hence, we allow interactions that are less goal-directed than in existing spoken dialogue systems. During the dialogue the user will determine and update goals. 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.). Visitors can explore the surroundings of the building, hear the carillon of a nearby church, look at neighbouring pubs and a 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 daily updated using information from the database. Clearly, visitors may go to an information desk, see previews and start a dialogue with an agent called 'Karin'. The first version of Karin looked like other standard avatars available on World Wide Web. The second version, now available in a prototype of the system, has a more human-like appearance making visitors happy to talk with her and ask about performances and make reservations.

One may argue the necessity of this realistic modelling of the theatre and its 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.

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

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. 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

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 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.

Our current version of the dialogue system of which Karin is the face is called THIS v1.0 (Theatre Information System). The approach used can be summarised as *rewrite* and *understand*. User utterances are simplified using 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. 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 found in the utterance or context data structure 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 this approach 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. Based on the most recent user utterance, on the context and on the database, the system decides on the 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 parameters. Among them are *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. 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 useful when the user has a number of alternative options to choose from. 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 different modalities, 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 our present scope of research, we need to model and use interaction knowledge available from 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 pages has the possibility to explore the environment and to start interactions with all kinds of agents. Agents that help to navigate, agents that allow access to particular information, agents that give information about other visitors and their opinions about performances, agents that want to sell tickets. Etc.

In such an environment users will display both chatting, browsing and goal-directed behaviour. This has consequences for the way the system manages the dialogue with the user. From dialogues that will be collected and from user evaluation studies we will try to determine guidelines for improved design.

4.2. A Talking Face for the Information Agent

The visual part of our information agent is presented as a talking face. It has become clear from several studies that people engage in social behavior 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 not 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, 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.

We developed a virtual face in a 3D-design environment. The face consists of various three-dimensional coordinates and is connected through faces. These faces are shaded to visualize a three-dimensional virtual face. The 3D data is converted to VRML-data that can be used for real-time viewing of the virtual face. A picture of a real human face can be mapped onto the virtual face. We are researching various kinds of faces to determine which can be best used for this application. Some are rather realistic and some are more in a cartoon-style. This face is the interface between the users of the virtual theatre and the theatre information system. A dialogue window is shown when users approach the information-desk while they are navigating in the virtual theatre. The face is capable of visualizing the speech synchronously to the speech output. This involves lip-movements according to a couple of visemes. The face also visualizes facial expressions according to user's input or the system's output.

We use Cosmo Player, which is a plug-in for an HTML-Browser, for viewing VRML-files. These files are specifications of a three dimensional virtual environment. The whole virtual theatre is a collection of VRML files, which can be viewed by the browser. As mentioned earlier, the user will see a virtual face when the information desk is approached. A dialogue window also pops up at this time. This is called the JAVA Schisma applet. In this window, the user can formulate questions or give answers to the system's questions. The user types the questions on a keyboard in Dutch sentences. The answers to the questions are to be determined on the server side: the Schisma server. Answers or responding questions are passed to the JAVA Visual Speech Server Application on the server side.

This application filters the textual output of the dialogue system in parts that are to be shown in a table or a dialogue window and parts that have to be converted to speech. The parts that are to be shown in the dialogue window or a table, like lengthy descriptions of particular shows or lists of plays are sent to the Schisma Client Applet where they are showed on the screen. The parts of the Schisma output that are to be spoken by the virtual face are converted to speech with the Text-to-Speech Server. The input is the raw text and the output is the audio file of this spoken text and information about the phonemes in the text and their duration.

For example, the Dutch word for "speech generation" is "spraakgeneratie". This word contains the following phonemes: S p r *a k x e n @ r a t s l. When the resulting audio file is played, each phoneme has it's own duration. This information is gathered from the TTS-server:

s 79 p 71 r 38 a 106 50 127 k 53 x 90 e 113 20 102 n 60 @ 38 r 53 a 101 t 23 s 113 l 119 20 75

The characters are the phonemes and the first number after the characters are durations of the corresponding phonemes in milliseconds. If more numbers follow then the first number is a percentage of the whole duration in which the pitch of the voice changes to the following number. So the first 'a' is spoken for 106 milliseconds and on 50% of this 106 milliseconds the pitch changes to 127 Hz. The audio file, which the TTS-server produced, will be compressed to a Real-Audio file for a fast transfer-rate. The previously described information from TTS-server will be sent to the JAVA Visual Speech Client Applet together with the converted audio file. The Visual Speech Client Applet uses the phoneme information to map the phonemes onto different mouth states or visemes. All the phonemes are categorized in five visemes. When the audio file is loaded on the client side, the mouth states and their durations are passed to the External Authoring Interface (EAI). This is an interface between JAVA and the VRML browser. This interface triggers animations in the virtual environment. It starts the sound playback and all the corresponding animations. Only the mouth states are specified in the VRML-file. The animation is done by interpolating between mouth states in the given amount of time. This results in smooth lip-movements.

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 or customer centric systems and that aims at improving relations with clients, that aims at supporting customer service and that aims at customising offerings to suit individual interests.

6. References

K. van Deemter et al. Generation of spoken monologues by means of templates. In: *Speech and Language Engineering*, TWLT8, Universiteit Twente, Nederland, 1994, 87-96.

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

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

J. Hulstijn, R. Steetskamp, H. ter Doest, S.P. van de Burgt & A. Nijholt. Dialogues in a Theatre Information and Booking System. In: *Proceedings 5th International Symposium on Social Communication*, Santiago de Cuba, September 1997, 87-99.

J. Hulstijn & A. van Hessen. Utterance generation for transaction dialogues. *Proceedings International Conference on Spoken Language Processing*, Sydney, Australia, November 1998, to appear.

D.H. Lie, J. Hulstijn, R. op den Akker & A. Nijholt. A transformational approach to natural language understanding in dialogue systems. *Proceedings NLP and Industrial Applications*, Moncton, New Brunswick, Augustus 1998, 163-168.

A. Nijholt, A. van Hessen & J. Hulstijn. Speech and language interactions in a (virtual) cultural theatre. *Proceedings NLP and Industrial Applications*, Moncton, New Brunswick, Augustus 1998, 176-182.

A. Nijholt, M. van den Berk & A. van Hessen. A natural language web-based dialogue system with a talking face. *Proceedings Text, Speech & Dialogue*, Brno, Czech Republic, September 1998, to appear.

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