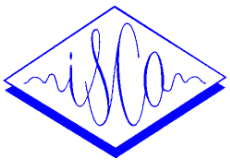
ISCA Archive
http://www.isca-speech.org/archive



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

AdApt – a multimodal conversational dialogue system in an apartment domain

Joakim Gustafson, Linda Bell, Jonas Beskow, Johan Boye, Rolf Carlson, Jens Edlund, Björn Granström, David House and Mats Wirén**

Centre for Speech Technology, Royal Institute of Technology, S_100 44 Stockholm, Sweden
*also Telia Research, S-123 86 Farsta, Sweden

ABSTRACT

A general overview of the AdApt project and the research that is performed within the project is presented. In this project various aspects of human-computer interaction in a multimodal conversational dialogue systems are investigated. The project will also include studies on the integration of user/system/dialogue dependent speech recognition and multimodal speech synthesis. A domain in which multimodal interaction is highly useful has been chosen, namely, finding available apartments in Stockholm. A Wizard-of-Oz data collection within this domain is also described.

1. INTRODUCTION

This paper presents a general overview of the AdApt project and the research that is performed within the project. Some of the research topics are also described in more detail. The AdApt project has been running at CTT (Centre for Speech Technology) for a little more than one year. One of the aims of the project is to investigate various aspects of human-computer interaction in a multimodal conversational dialogue systems. We also intend to focus on the integration of user/system/dialogue dependent speech recognition and audiovisual synthesis. 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. 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 AdApt project builds on previous research at KTH and Telia Research, and is part of the co-operative effort between KTH and industrial partners within CTT. In the first multimodal spoken dialogue system at KTH, Waxholm [5], a user was able to get tourist information about the Stockholm Archipelago. An animated agent provided users with both auditive and visual feedback. In the August project, we wanted to examine how members of the general public would interact with a multimodal spoken dialogue system with an animated talking head. The August system covered several domains, and was available daily to any visitor at the Stockholm Cultural Centre as part of the Cultural Capital of Europe '98 program. The August database, which contains more than 10,000 spontaneous computer-directed utterances, has been examined in detail [10]. The travel-planning system at Telia Research was developed to study spoken dialogue in a domain with a possibly high degree of user initiatives. The system included an agenda-driven dialogue manager and a parallel model with both shallow and deep language-processing, [7].

2. HUMAN-MACHINE INTERACTION

Several research issues are approached in the AdApt project. Human-machine interaction in multimodal dialogue systems, together with further development of the base technologies, especially dialogue dependent speech recognition [15] and multimodal speech synthesis are areas in focus, [3,4].

In the areas of human-computer interaction in multimodal dialogue systems, we will be addressing the following long-term research issues: 1) How do people make use of the different modalities and what are the implications of their choices in terms of system architecture? 2) How should the system interpret references, not only to objects previously mentioned in the dialogue, but also to objects currently visible on the screen? 3) Can multimodality be used to increase the robustness of a spoken dialogue system? Users often change from one modality to another when the human-computer interaction becomes difficult. 4) How does the multimodal setting with a mouse, an interactive map and an animated speaking agent influence how people speak? 5) Can the system influence users in their choice of modality?

3. THE ADAPT DOMAIN

The real-estate domain explored in the Adapt project provides a challenging environment for research in the field of multimodal dialogue systems. (see also [9, 13] for descriptions of different types of systems using the real-estate domain). An apartment is a complex object that has properties which can be presented graphically (e.g. its location in the city), as well as properties that can be presented verbally (price, description of interior details, etc). Furthermore, the real-estate domain attracts interest from a wide range of people, regardless of whether they are seriously thinking of getting into the market to purchase an object or not. Since this domain is one that many people want to keep up to date with, and look around to see what is offered and at what price, the resulting interactions are likely to be characterized by browsing rather than pure information seeking.

In the AdApt project, we aim to develop a fully functional multimodal dialogue system. Apart from spoken input, users have the possibility of providing the system with additional information by clicking on an apartment icon (see Figure 1) or marking areas on an interactive map of Stockholm. Considerable efforts to study multimodal input in map task domains have been reported by the SRI group, see for example [11]. We will therefore address the question of how to integrate the results from a speech recognizer with mouse input such as the selection of an icon or the specification of an area on the map. The system requires a dialogue manager capable of integrating input/output modalities [14], resolving anaphora and contextual references, understanding the pragmatic functions of the input utterances, and handling misunderstandings and out-of-domain utterances in a robust and graceful manner. The implementation also demands solutions concerning extraction and structuring of relevant information from web pages, designing a flexible lexicon manager, specifying a semantic formalism for representing the input utterances, and automatic generation of output utterances with correct prosody from semantic representations. The coordination of audiovisual synthesis with graphical output (tables and maps) is also an important issue. We are in the process of developing a strategy for automatically supporting the user's decision of selecting objects depending on underlying but not expressed preferences. Such models have previously been tried in the real-estate domain, [8].

4. AUDIO-VISUAL SYNTHESIS

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 parameterized, deformable 3D facial models, controlled by rules within a text-to-speech framework. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account, [4]. We employ a generalized parameterization technique to adapt a static 3D-wireframe of a face for visual speech animation, [5]. The parameters are designed to allow for intuitive interactive or rule-based control and include both articulatory parameters, such as lip rounding as well as non-articulatory, such as eyebrow raising.

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 [9, 14]. 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, [12]. 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. It is important that there is a one-to-many relation between the symbols and the actual gesture implementation to avoid stereotypic agent behavior. Currently a weighted random selection between different realizations is used.

5. DATABASE FROM THE WEB

An advantage of the apartment-seeking domain is that it is easy to continuously access genuine data about apartments for sale through the real-estate web sites that are available on the net. These sites typically feature an apartment index, where some structured data is available: address, size, price. The indices also contain hyper-links to free text descriptions of the objects. These descriptions are written by various real-estate agents or the present owner, and so vary considerably in terms of style, verbosity and degree of syntactic well-formedness.

Within Adapt, the objective information in these free texts is highly useful, whereas the many subjective statements are not. The objective statements in about 200 texts were extracted manually in order to get an understanding of what one could expect to find. Based on this data, simple pattern matching techniques were developed to parse the free texts automatically. Initial tests show that the automated parsing picks up around 60% of the information that was produced by manual extraction, which is sufficient for our immediate purposes.

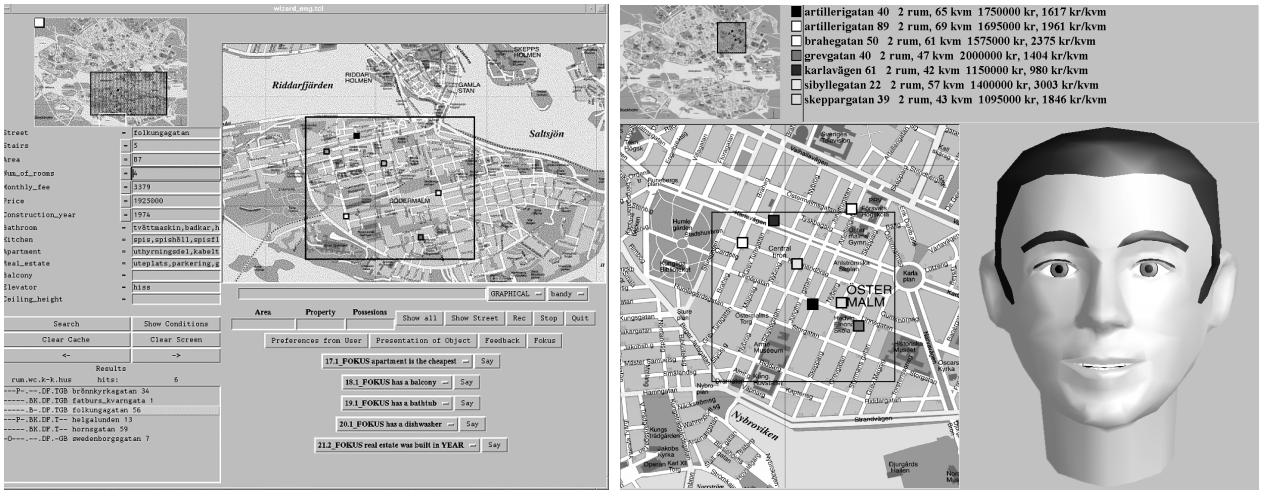
After downloading, HTML stripping and parsing, the structured data and the results from the free text parsing are merged and stored in a freely searchable database. The information is divided in facts about the surroundings and the building (shared facilities such as sauna, year of construction), and facts about the apartment in itself (contents of the kitchen, monthly cost, presence of bathtub). In addition to what is said explicitly on the real-estate web sites, the database facilitates the use of derived information: the average price level of an area can be computed, and restaurants and communications in the vicinity can be looked up in the yellow pages.

In this way, the AdApt database can be said to reflect the flux of the real-estate market, where one area quickly can become more popular (and more high-priced) than another. The database module is used in the Wizard-of-Oz interface (by the wizard), where the database is frozen to make comparisons more reliable. The dynamic version, which is updated six times daily, will be used by the Adapt system

6. WIZARD-OF-OZ DATA COLLECTION

To be able to collect the data necessary for developing the AdApt system, a simulated prototype version of the system was constructed. This simulation tool included an animated talking head, an interactive map of Stockholm showing names of streets, major neighborhoods, parks, etc., an overview map allowing the user to scroll the detailed map, and a table for displaying graphical information. The location of individual apartments was shown as colored icons on the map. Figure 1 shows the AdApt graphical user interface.

The user's input was sent to the wizard interface where a human operator controlled the system's response. Much care was



The Wizard Interface **The User Interface**

Figure 1. The AdApt wizard and user interface.

devoted to design the wizard interface to allow rapid system response times (typically between one and two seconds), thus giving users the impression of a fully functional system. The wizard chose his answer from a number of pop-up menus, where information about specific apartments from the database was included in one of a number of possible answer templates. Apart from the spoken input to the system, the wizard also had a display of the subject's map, where his or her graphical operations were visible. The wizard interface can be seen in Figure 1. In addition, there was an interface to the database in which the wizard entered the user's verbally expressed preferences. The result of each query was displayed in a field where each individual apartment was given a list of letters that summarized the most important features. To exemplify: "..F...D..C. Hornsgatan 16", would mean that the apartment on Hornsgatan 16 had a fireplace (F), a dishwasher (D) and cable-tv (C). To obtain all available information about an apartment in the list, the wizard could select it with the mouse. This extensive information was displayed in the query form that also was used to do the actual search. The information about a selected apartment was also inserted in all answer templates mentioned above. If the apartment in the above example had been selected, the following answer would have been generated (for example): "*The apartment on Hornsgatan has a dishwasher*". It was also possible to tell the system how to refer to the apartments. If the deictic reference mode was selected the answer would instead have been: "*This one has a dishwasher*" at the same time as the icon for the apartment was highlighted.

The animated agent indicated with facial expressions that it was 'listening' while the subjects spoke and then turned to a 'thinking' gesture as soon as silence has been detected. This made the system appear more reactive, since the wizard could select what to say from the menus of utterance patterns while the system automatically displayed the thinking gesture. In most turns the answer from the system came within a second. However, in the turns where the wizard had to perform a database search the waiting time was slightly longer. To hold the floor, the wizard used utterances like "*Could you please wait while I search my database*?".

The experimental task used in the Wizard-of-Oz collections involved the construction of deictic references to specific apartments on a map. To avoid verbal biasing, pictorial scenarios were used. The scenarios were deliberately rather unspecific, and the users were encouraged to take their time and browse available apartments until they found the most appropriate one. Subjects who referred to apartments had the option of using either graphical or verbal means, or both. For each displayed icon, limited information about the corresponding individual apartment was provided in the row of a table. Here, the apartment's address, size and listed price were displayed. Icons on the map that represented apartments at adjacent or identical positions were only allowed to overlap to a limited extent in order to keep them simultaneously visible to the user. The Wizard-of-Oz collections resulted in a database of 33 users performing 50 dialogues. The total number of utterances is 1845. A section of one of these dialogues is presented in Table 1. The database was manually transcribed and annotated, and the graphical input to the system was timed against the spoken input.

7. CORPUS OBSERVATIONS

In the analysis of the data collected so far, several interesting patterns emerged. For one thing, the number of incomplete or fragmented utterances was very large. Typically, an initial fragment would be followed by a silent pause and then by one or several additional fragments. This also explains the, relatively speaking, large number of topicalized phrases in the corpus. Instead of saying "Hur mycket kostar lägenheten på Nybrogatan?" (*How much is the apartment on Nybrogatan*?) some subjects would say "Lägenheten på Nybrogatan /paus/ hur mycket kostar den"? (*The apartment on Nybrogatan* /pause/ *how much is that*?).

Corresponding topicalized phrases were also observed in the multimodal user utterances. A subject would click on an apartment icon and then say something like "*how much is it*?" In very few cases did the click occur during or after the spoken input had come into the system. Instead, subjects typically selected an apartment icon graphically between one and four seconds before speaking. Analyses of the distribution of silent pauses in the corpus indicate that the system's previous turn influenced the length of the pauses between fragments. Pause lengths were measured for a subsection of the corpus where subjects used an initial feedback cue to acknowledge the system's previous turn [3]. Here, the pause between fragments was on average 0.5 seconds after a direct question from the system while it was as long as 3 seconds after the system had presented a number of apartments. As can be seen in the example dialogue below, the subjects' utterances to the system varied in length and complexity.

System	Subject
Jag heter Urban och jag kan berätta om bostäder i Stockholm *My name is Urban and I can tell you about apartments in Stockholm*	Visa lägenheter i Vasastaden och Östermalm *Show apartments in Vasastaden and Östermalm*
Hur många rum vill du ha? *How many rooms do you want?*	Ett till två *One to two*
Jag hittar för många lägenheter för att visa . I vilket område ska jag leta i första hand? *I find too many apartments to show. In which area should I begin by looking?*	Vasastan *Vasastan*
Ett område på kartan är markerat . där finns det just nu 8 lägenheter till salu *An area on the map has been marked. There you can find 8 apartments for sale*	Visa lägenheter byggda mellan 1900 och 1950 *Show apartments built between 1900 and 1950*
I det markerade området hittar jag 6 lägenheter *In the marked area there are 6 apartments*	Eh berätta om frejgatan 51 *Er tell me about Frejgatan 51*
Den svarta huset är byggt 1926 *The black building was constructed in 1926*	Vad finns i lägenheten? *What is in the apartment?*
Den svarta fastigheten har kabeltv *The black building has cable tv*	Trevligt finns det balkong *Nice is there a balcony*
Nej *No*	Visa lägenheter ur samma urval med balkong *Show apartments from the same selection with a balcony*
Hittar inga sådana *Cannot find any*	Okej utöka sökning till ett större område *Okay widen the search to a larger area*

Table 1. An example dialogue from the AdApt Wizard-of-Oz collection.

The average utterance in the AdApt database was 7.3 words, but the number of words per utterance ranged from a single word to 47 words. While some utterances were simple and almost telegraphic in their style, others were quite long and complex. Utterances that included both deictic references and discourse references were observed in the data. For example, a user would graphically select an apartment and at the same time say: "Har **den här** också **det**?" (*"Does **this one** also have **that**?")* : In a specific dialogue context, it might be uncertain exactly which discourse referent the subject is trying to pick out. This makes them rather difficult for the dialogue system to handle.

Quite a large number of disfluencies were found in the AdApt corpus. Detailed analyses of disfluency distribution in the corpus revealed that filled pauses and prolongations of segments were particularly frequent, and repetitions and truncated words were also relatively common, [2]. Individual variations were substantial, so that while the speech of some subjects was highly disfluent , a few speakers in the corpus were not disfluent at all. The average disfluency rate per word (excluding unfilled pauses) was 6%, which although high is in the range of figures in the literature.

8. DISCUSSION

We are in the process of developing the modules that were simulated in the Wizard-of-Oz set-up. The detailed analyses of the collected dialogues have provided valuable information that will be used when we design these modules. An input handler is necessary to handle the multimodal input from the user. The large number of topicalized utterances, both uni- and multimodal, require that the input handler is adaptable to the current dialogue state. In some cases, a reference to a specific apartment is in itself a complete utterance. For example, a user may select an apartment icon or say :"*The apartment on King's Street*" This should be interpreted as a complete utterance after the system question: "*Which apartment do you mean?*". However, if the system had instead presented a number of apartments such an utterance from the user should be regarded as incomplete. Here, the system should wait for two to three seconds and anticipate a continuation of the query. If nothing more is said within this time limit a clarification subdialogue can be initiated: "*What do you want to know about that apartment?*"

The system's dialogue manager will use a referent tracker to represent the current apartment, apartment attributes, and for comparing apartments and their respective attributes. Referent tracking in the system is also closely related to focus prediction and assignment in the generation of output utterances, for both visual and verbal prosody. Here, the interaction between apartment description, comparison and dialogue state is discussed.

9. CONCLUDING REMARKS

We have in our paper described the AdApt project and some of the research issues involved. The collected database has been of great value for our continued research in the area of multimodal dialogue systems. The wizard environment is currently being transformed into a complete multimodal dialogue system.

10. REFERENCES

1. Bell, L., Boye, J, Gustafson, J and Wirén, M. 2000. Modality Convergence in a Multimodal Dialogue System. *Proceedings of Götalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*, pages 29-34.
2. Bell, L., Eklund R. and Gustafson, J. 2000. A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Speech Interface. *Proc. ICSLP '00*, [These proceedings.]
3. Bell, L. and Gustafson, J. Positive and Negative User Feedback in a Spoken Dialogue Corpus. *Proc. ICSLP '00*, [These proceedings.]
4. Beskow J (1995). Rule-based Visual Speech Synthesis, In *Proceedings of Eurospeech '95*, 299-302, Madrid, Spain.
5. Beskow J (1997). Animation of Talking Agents, In *Proceedings of AVSP '97*, 149-152. Rhodes, Greece.
6. Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993): An experimental dialogue system: WAXHOLM. *Proceedings of Eurospeech '93*, Berlin, 1993, Vol. 3, pp. 1867-1870.
7. Boye, J., Wirén, M., Rayner, M., Lewin, I., Carter, D., and Becket, R. Language-Processing Strategies for Mixed-Initiative Dialogues. *Proceedings of IJCAI-99 Workshop On Knowledge And Reasoning In Practical Dialogue Systems*, pp. 17-24.
8. Carenini G., Moore J (1998), Multimedia Explanations in IDEA Decision Support Systems. *Working Notes of AAAI Spring Symposium on Interactive and Mixed-Initiative Decision Theoretic Systems*. Stanford, California (USA), 1998.
9. Cassell J, Bickmore T, Campbell L, Hannes V, and Yan H (2000). Human Conversation as a System Framework: Designing Embodied Conversational Agents, In Cassell J, Sullivan J, Prevost S and Churchill E (eds) *Embodied Conversational Agents*, 29-63, Cambridge MA: The MIT Press.
10. Gustafson, J. and Bell, L. (2000). Speech Technology on Trial – Experiences from the August System. In *Natural Language Engineering* (1) 2000, in press.
11. Moran D, Cheyer A, Julia L, Martin D and Park S (1998) Multimodal User Interfaces in the Open Agent Architecture. *Journal of Knowledge Based Systems*, 10:295-303.
12. Nass C and Gong L (1999). Maximized modality or constrained consistency? *Proceedings of AVSP'99*, 1-5, Santa Cruz, USA.
13. Oviatt, S. L., DeAngeli, A. & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *Proceedings of Conference on Human Factors in Computing Systems: CHI '97*. New York, ACM Press. 415-422.
14. Pelachaud C, Badler N I and Steedman M (1996). Generating Facial Expressions for Speech *Cognitive Science* **28**, 1-46.
15. Seward, A. (2000) A Tree-Trellis N-best Search Algorithm for Real-Time Continuous Recognition using Stochastic Context-Free Grammars, *Proc. ICSLP '00*, Beijing, China. [These proceedings.]