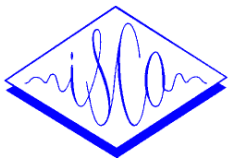

ITRW on Multi-Modal Dialogue
in Mobile Environments
Kloster Irsee, Germany
June 17-19, 2002

TURN-TAKING GESTURES AND HOURGLASSES IN A MULTI-MODAL DIALOGUE SYSTEM

Jens Edlund
CTT (Centre for Speech Technology)
KTH, Sweden
edlund@speech.kth.se

Magnus Nordstrand
CTT (Centre for Speech Technology)
KTH, Sweden
magnusn@speech.kth.se

Abstract An experiment with 24 subjects was performed. The subjects were split in three groups, and asked to extract information from the AdApt dialogue system for somewhat over 30 minutes per subject. The system configuration varied in that one group had turn-taking gestures from an animated talking head, another had an hourglass symbol to signal when the system was busy, and the third had no turn-taking feedback at all. The results show that although the hourglass setup showed no decrease in efficiency compared to the facial gestures, it made the subjects less satisfied. The lack of turn-taking feedback was noticed and mentioned by half of the subjects in that group.

Keywords: multi-modal, dialogue system, turn-taking, animated face

Introduction

Letting the system provide feedback about what it is doing helps to maintain a good dialogue flow in a conversational dialogue system. In the present dialogue system, such feedback is realised as facial gestures in an animated talking head. The system generates output using the GESOM model (Edlund et al., 2002), which lets us change the realisation of such feedback without changing the dialogue system as such.

It is quite feasible to run a large conversational system in a PDA or even a mobile phone, provided that the bulk of the processing is done on a central server and the PDA/phone only deals with input/output.

The talking head used in the present system could possibly run in a PDA, but in a standard mobile phone of today it would not. There are, however, other ways of providing multi-modal feedback that may work. For example, the well-known hourglass symbol could be used to indicate that the system is busy.

A test was performed where users were asked to try one of three system setups: facial turn-taking gestures, an hourglass symbol, and a configuration with no visual turn-taking feedback at all. The results show no significant difference in efficiency between the hourglass and the gestures, but a tendency for users to get their turn wrong more often when no feedback is provided. Several users mentioned feeling uncertain about when to speak as a problem in the no-feedback configuration, and the results of a series of evaluation questions based on the PARADISE subjective measure show that the users who tested the hourglass configuration were significantly less happy with the system.

1. Related work

People take turns speaking in human to human dialogues, and each party begin and end smoothly. Conversations rarely break down due to simultaneous speech or interruption, even though the pauses between turns are short (Torres et al., 1997). In all likelihood there is some non-verbal exchange between the speakers. This information can be provided by hand gestures, body posture, speech, gaze, or any combination thereof (Goodwin, 1981; McNeill, 1992). Evidently, the use of such information in a dialogue system would be of great help in the communication between the user and the agent.

2. The AdApt system

AdApt (Gustafson et al., 2000) is a multi-modal dialogue system that was developed at CTT with Telia Research as an industrial partner. It provides information about apartments for sale in downtown Stockholm. The system uses speech and mouse clicks as input and a 3D-animated agent (Beskow, 1997) that produces lip-synchronized synthetic speech and gestures as output. The system also displays the location of apartments on a map.

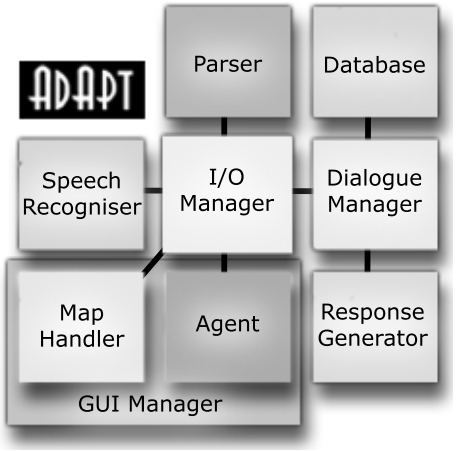

Figure 1. AdApt system overview

2.1 Dialogue flow

AdApt is a distributed system (see fig. 1) with a central module controlling input/output, the IO Manager. In a normal dialogue turn, the system gets input from the speech recogniser and/or the clickable map. The input is merged in the IO Manager and sent to the parser, which returns a semantic representation to the IO Manager. The semantic representation is sent to the dialogue manager, which takes the necessary action to create a response. The response, again, is sent to the IO Manager, which passes it on to the GUI Manager, which in turn realises the response as speech, map objects and gestures.

Each time data is passed through the IO Manager, the system has an opportunity to provide feedback to the user. An unsuccessful recognition may result in the agent asking the user to repeat the utterance, and an unsuccessful parse may cause it to inform the user that it does not understand the utterance. In the system set-up used here, non-verbal feedback is used to show the user whether or not the agent has understood.

2.2 Fragment parsing

Before implementing the final version of AdApt, a Wizard-of-Oz study was performed (Bell et al., 2000) The simulations showed that people tended to break up their utterances into fragments. Many user queries would look something like this:

"*I would like a* /pause/ *three-room apartment in this area*"

The pauses were often long - much longer than the half second or so response time of the endpoint detection used in AdApt's speech recogniser. This meant that the system would receive a lot of utterance fragments (e.g. "I would like a"). In order to come to terms with this, a parser that can distinguish full utterances (i.e. closing utterances) from fragments (i.e. non-closing utterances) was implemented (Bell et al., 2001).

Given that we know whether we are presented with a full utterance or with a fragment, the system can take the appropriate action - respond or wait for more input, respectively. Unfortunately, the system sometimes takes a little while to generate a response, and the user may well decide to say something more during that time. This presents a problem, especially since the present set-up of AdApt does not support barge-in. However, in a dialogue system that aims at natural language communication, barge-ins may be difficult to handle even if the speech recogniser could deal with them flawlessly. In a situation where the system is just about to reply, for example, the best course of action may well be to go ahead and do so, even if the user barges in. Fig. 2 show an example from out test data where this would be the case.

```
USER:    vad kostar den
         what does it cost
         [this utterance was correctly recognised and
         triggered a database search]
USER:    hur mycket kostar den
         how much does it cost
         [this utterance never reached, but would have
         caused a barge-in]
SYSTEM:  lägenheten kostar 2.150 miljoner kronor
         the apartment costs 2.150 million crowns
```

Figure 2. A barge-in from the test data. The barge-in was captured from a sound recording of the entirety of the session, not from the recogniser's log files.

In cases like this, we would get smoother dialogues if the users would wait for their turn before speaking. There are various ways to influence the user in this direction: one might let the system say "Just a moment, I'll see", or show a well-known computer symbol suggesting that the system is busy (e.g. an hourglass). One could also let the animated agent show that it is thinking through the use of facial gestures.

Figure 3. AdApt GUI

2.3 Gestures

Facial gestures can be shown shown in the AdApt GUI (see fig. 3). Turn-taking gestures are defined as states, e.g. *busy*, *continued attention*, which the talking head can be placed in. Placing the talking head in the busy state makes it perform one of a number of predefined gestures (Beskow et al., 2002). This is done when the system receives a full utterance. The continued attention state also triggers a gesture, but the gesture is picked from a different set. This happens when the system receives an utterance fragment. The present system never verbally prompts the user to continue when it receives an utterance fragment. In many cases, the rest of the fragment comes very shortly after, in which case verbal prompting would leave us with a barge-in again.

A number of gestures in the *busy* and *continued attention* states were created. A group user test, adopted from Granström et al., 2002, was then performed to see whether they were perceived as intended when put against each other in the same dialogue context (Nordstrand, 2002). The busy state basically contains gestures where the animated head looks away from the user, and the continued attention state holds two types of gesture: one where the animated head tilts forward whilst keeping its gaze on the user and raising the eye-brows and one where the head tilts to the side slightly. The group test contained a large number of gestures, most of which were clearly perceived as intended. The four most clearly perceived gestures out of each group were picked for the present test, and several slightly different versions of them were added to prevent the system from being too repetitive.

3. Experiment

3.1 Subjects

The user study included three groups of eight paid subjects between 20 and 40 years of age. Half of the subjects were men. The groups were balanced for gender, but otherwise randomly distributed. None of the subjects had professional knowledge of speech technology, although the majority claimed to have used some speech interface at some point, to have had some experience with computers, and to have used apartment search tools on the Web. All subjects claimed a reasonable knowledge of the geography of downtown Stockholm, which is the area the AdApt system is concerned with. None of the subjects had used or seen the AdApt system before. Table 1 shows a compacted version of some of the pre-test query results.

Table 1. Previous experience

	Speech	**Programming**	**Apartment search tools**
Never	2	5	8
Tried	22	17	16
Regularly	0	2	N/a

3.2 Test setup

One goal of the test was to attempt to answer the following questions:

1 How well do the busy gestures and the symbol perform in preventing users from speaking to the system when the system is preparing a reply?

2 Which of the two techniques is more efficient?

3 Does either of the two techniques appeal more to users?

We hoped to show that both the gestures and the symbol have a significant effect, and expected the hourglass approach to be more efficient, but that it might make the users feel less satisfied with the system or perceive it as slower. Three system configurations were used. One of them used the busy gestures described in section 2.3, one turned the mouse pointer into an hourglass, Windows-style, and the third configuration did not give any turn-taking feedback at all. The system ran on

three machines: speech recognition on one, surveillance on another, and everything else on a third.

The subjects received the following information about the system before the test commenced:

- The system takes mouse and voice input.
- The system responds with voice synthesis from an animated talking head and markings on a map.
- It is quite possible to get stuck in some situations. If this happens, say "Urban, börja om" ("Urban, start over", Urban is the name given to the talking head used in the system).
- The system has information about apartments for sale in downtown Stockholm.
- The task is to find information about apartments that you might want to buy or live in.

In addition, the subjects were informed that with this little guidance, at least the first minutes would be difficult at best.

Each subject talked to the system for a little over 30 minutes. They were left alone in an undisturbed room with the system, although they were being recorded with an open microphone, and were naturally aware of this. No further assistance were given to any of them (except for someone entering the room silently to restart a crashed module on a couple of occations), and after the test, conversation with them was kept to a minimum until they had filled out the evaluation forms.

3.3 Analysis

The subjects dialogues were logged, as were the sound files from the speech recognition. In addition to this, an open microphone was used to record the entire conversations. The resulting tapes have been transcribed, and the start of utterances that did not reach the recogniser timed. Unfortunately, the timestamps in the log files have turned out to be unreliable - a negative effect of the distributed system setup used. As a result, much more manual work than expected is needed in order to make any serious attempt at measuring the efficiency of the turn-taking feedback using times. A rough measure of how many times the users made additional utterances when the system was preparing yields table 2. The subjects got on average about 60 system responses each, and the responses were quite evenly distributed across the groups, making a total of somewhat less than 500 responses per group.

Table 2. % of system responses where the subject started saying something

		All	Last 2/3
Test configuration	Gest	9.9%	8.7%
	Symbol	8.9%	6.9%
	None	9.9%	12.5%

The column marked "All" shows the percentage of all utterances in all sessions where the subjects started to speak as the system was preparing a response. The numbers are misleading, however, since the subjects spent their first five to ten minutes aquainting themselves with the system. In the case of the no-feedback group, the subjects would wait for very long times before repeating anything early on in the sessions, regardless of whether the system needed more input or not. After some time, they would realise that waiting for a response would often not get them anywhere, and start repeating their requests. Unsurprisingly, they then spoke quite often as the system was preparing responses. The column marked "Last" shows the same figure for the last two thirds of the sessions. The difference between the symbol and the gesture group is not significant, and the differences between the no-feedback group and the other two are barely so. This measure, however, is too crude to be fully reliable, but we will claim that it is a workable tendency. There is an attempt being made at CTT to make a PARADISE evaluation on the material, which will perhaps help matters.

The users where asked to fill in two forms after the tests. The first one was a user satisfaction form based on the method described in PARADISE (Walker et al., 1997), and contained a number of questions about various aspects of the system. Table 3 show which configuration got the best rating when compared pair-wise, with "S" marking a significant difference.

The second form, presented to the users after they had completed the first, contained nothing but the question "Do you have any additional comments?". The majority of the subjects wrote half a page or more in response to this, and to our surprise, almost half of them mentioned turn-taking. These judgements where easily divided into three categories. The first is made up of explicitly positive statements, like "the gestures he made when he was thinking were nice". The second is made up of explicitly negative statements, like "I never knew if I was supposed to talk or wait for a reply". The third category is made up of statements showing that the subject has understood that the system tried to signal

Table 3. Evaluation results

	gest	**symbol**	**none**
gest	-	-	-
symbol	gest(s)	-	-
none	gest	none(s)	-

that it was busy, but felt that the signal was not clear enough: "The thinking gestures were a bit mild" and "I thought the hourglass signalled that he was thinking, but I felt unsure". Note that the no-feedback system could not possibly get any comments on turntaking signals, since there were none. The results of this categorisation are presented in table 4.

Table 4. Number of subjects making comments on turn-taking issues

Setup	**Positive**	**Negative**	**Noticed but somewhat critical**	**Total comments**
Gest	1	-	2	3
Symbol	1	2	1	4
None	n/a	4	n/a	4

4. Conclusion

The user test described in this paper has in part verified our intuitions. The PARADISE style user satisfaction evaluation suggests that the hourglass made the subjects feel less happy about the system, even though we have not been able to show any significant decrease in efficiency or success rate in the hourglass group. The subjects' comments show that many subjects *did* take notice of the turn-taking signals (or lack thereof). Unfortunately, we have not been able to show that the our turn-taking signals make for a more efficient dialogue in a convincing manner. Preliminary time measurements suggest that the time it took for the user to give more input when the system needed it was significally longer when no turn-taking signals were present, but the these results come from a small part of the data. We hope that these figures will turn out to be correct, and that the PARADISE evaluation shows differences between the different setups.

Acknowledgments

Many heartfelt thanks to Anna Hjalmarsson for her help, both with the tests and with the data analysis. Thanks also to Rolf Carlson and the people at Telia Research for their support. This research 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 organisations.

References

Bell, L., Boye, J., and Gustafson, J. (2001). Real-time Handling of Fragmented Utterances. In *Proceedings of the NAACL Workshop on Adaption in Dialogue Systems*, Pittsburgh, PA.

Bell, L., Boye, J., Gustafson, J., and Wiren, M. (2000). Modality Convergence in a Multimodal Dialogue System. In *Proceedings of Götalog 2000*, pages 29–34. Fourth Workshop on the Semantics and Pragmatics of Dialogue.

Beskow, J. (1997). Animation of Talking Agents. In *Proceedings of AVSP'97*, pages 149–152, Rhodes, Greece.

Beskow, J., Edlund, J., and Nordstrand, M. (2002). Description and Realisation of Multimodal Output Dialogue Systems. submitted to ICSLP'02.

Edlund, J., Beskow, J., and Nordstrand, M. (2002). GESOM - (GEneric System Output Model) - a model for describing and generating multi-modal output. (this volume).

Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, New York, NY.

Granström, B., House, D., and Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. In *Proceedings of the Speech Prosody 2002 Congference*, pages 347–350, Laboratoire Parole et Langage, Aix-en-Provence.

Gustafson et al., J. (2000). Adapt - a Multimodal Conversational Dialogue System in an Apartment Domain. In *Proceedings of ICSLP 2000*, pages 134–137, Beijing, China.

McNeill, D. (1992). *Hand and Mind: What gestures reveal about thought*. University of Chicago Press, Chicago.

Nordstrand, M. (2002). A gesture library for an animated talking agent. Master's thesis, TMH (Speech, Mucis & Hearing), KTH.

Torres, O., Cassell, J., and Prevost, S. (1997). Modeling gaze beahavior as a function of discourse structure. First International Workshop on Human-Computer Conversation. Bellagio, Italy.

Walker et al., M. A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In Cohen, P. R. and Wahlster, W., editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Somerset, New Jersey. ACL.