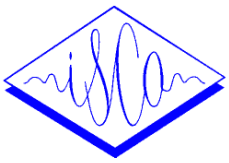

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

FEATURES OF NAIVE CALLERS' DIALOGUES WITH A SIMULATED SPEECH UNDERSTANDING AND DIALOGUE SYSTEM

C MacDermid

Social and Computer Sciences Research Group
Dept. of Sociology, University of Surrey
Guildford, Surrey, UK

ABSTRACT

During the development of speech-based database enquiry systems for dialogue over the telephone with members of the general public, Wizard of Oz simulations (in which an accomplice plays the role of the system) were conducted. The simulations provided evidence of naive callers adapting their speech to the system's presumed capabilities. They have also shown that callers tolerate speech recognition errors where there is graceful error recovery. However, the data have raised questions about the need for constraints to be imposed on callers' initial utterances if dialogues are to be successful.

Keywords: *Simulation, Wizard of Oz, spoken dialogue systems.*

1. INTRODUCTION

The SUNDIAL (Speech UNderstanding and DIALogue) project in the ESPRIT research programme has developed four national demonstrators of a speech-based database enquiry system capable of dialogue over the telephone with members of the general public.

The system is designed for real-time, intelligent, co-operative dialogue about specific topics (flight arrivals or train timetable information in these demonstrators), supporting a speaker-independent vocabulary of around 1000 words in English, French, German or Italian, with continuous-speech input.

All four demonstrators run on Sun 4 hardware. The UK demonstrator, for example, runs on a Sun Sparc 10 linked by Ethernet to a Sun 386i. (The front-end processor, the parser and the dialogue manager run on the Sparc 10 and the speech digitisation, parameterisation and synthesis run on the 386i.) The Italian system uses a PC for the front-end processor and a Sun 4/260 for the parser, dialogue manager and synthesis. The German demonstrator runs on a single Sun Sparc II.

Existing database access systems available to the general public via the telephone are mostly isolated-word recognition systems whose recognition vocabulary is largely restricted to yes/no and the ten digits. By necessity, a dialogue with such a system must be highly constrained and system-led. The SUNDIAL project explored the possibility of much more sophisticated dialogue, not only through introducing a large recognition vocabulary for continuous speech but also through the development of smooth recovery strategies to prevent dialogue failure following communication breakdown. The project drew on research into human-human dialogues in the domain of flight enquiries, which illustrated both the types of enquiries commonly made and the sophisticated strategies used for recovery from breakdown [1]. It was accepted that the recognition accuracy of a large-vocabulary spontaneous-speech understanding system would be imperfect in comparison to human sophistication (particularly when recognising speech over the telephone) and thus the ability to recover from breakdown should be an integral part of the SUNDIAL system. This would allow the system to offer the caller some initiative in the dialogue, unlike the existing, rigid menu-driven systems which make few allowances for recognition failures.

2. SIMULATING SUNDIAL

Having studied in detail how humans make enquires to human flight enquiry agents, we needed to know how they would formulate their enquiries to a computer system. It was important not to rely on the introspections of the system designers alone, because they may adapt their speech to fit

the capabilities and limitations of the system, whereas naive callers might not be so considerate.

Rather than build the system on hunches, only to find them not borne out in the final evaluation of the system, we conducted a series of Wizard of Oz simulations (in which an accomplice plays the role of the system) at each stage of system development [2]. From these we have obtained data on callers' conversations with what they believe to be the computer system, to provide design recommendations and, more recently, to evaluate system components.

3. THE BIONIC WIZARD

In the original UK simulation, the wizard (or accomplice) spoke to the caller via a vocoder to make the voice sound more machine-like, but otherwise followed dialogue behaviour observed in human flight enquiry calls. Recognition errors were introduced randomly, but the recognition accuracy and sophistication of the dialogue were high.

In the more recent 'bionic' Wizard of Oz simulations, the wizard used the Infovox text-to-speech synthesiser incorporated in the SUNDIAL system, thus increasing the realism of the voice quality and allowing evaluation of this system component in parallel with the simulation. In addition, a Prolog software tool was developed to enable the wizard to simulate the system more closely and test out design alternatives systematically. This interface, run on a Sun 'Sparc' workstation, allowed the wizard to simulate recognition errors and manipulate the synthesiser's use of enhanced prosody and text [3,4].

To generate output, the wizard selected an utterance template from a list of keywords on the screen, using a mouse. The task parameters for the scenario were inserted automatically by the simulation software, and the utterance was then synthesised using the TTS synthesiser. Utterances were also in some sense constructed in real time when, for example, confirming a series of flight parameters. The wizard would select a series of parameters to be confirmed, and then the appropriate confirmation phrase was constructed by the tool. Manipulated variables included the gender of the synthetic voice, enhancements to the output prosody and more elaborated versions of the text messages. In addition, various dialogue strategies were investigated in these experiments, for example, the use of an optional help facility for callers and different confirmation strategies.

Forty naive subjects have participated in these simulations. Subjects were ordinary members of the public and target user groups for flight enquiries (taxi companies and secretaries). Subjects were presented with text or pictorial flight-enquiry scenarios, based on actual flight enquiries to British Airways. For each scenario, subjects telephoned what they believed to be the system and tried to obtain the relevant flight information. They were instructed to note down the flight information given to them by the 'system', in order to make the task more realistic.

4. FEATURES OF DIALOGUES

As in the original simulation, dialogues with the bionic wizard showed evidence of naive callers simplifying their vocabulary and grammatical structures compared to human-human flight enquiry dialogues. However, in responding to a non-directive welcoming message from the bionic wizard, initial queries were often verbose; in the absence of guidance to the contrary, callers tended to include all the related information they had available (see Example dialogue). Using the error message "please be brief" was found to be counter-productive, generally resulting in an equally long query uttered in telegraphic speech (though it would eliminate irrelevant and problematic details such as "I have to meet my cousin at Heathrow tomorrow").

The bionic Wizard of Oz simulations have also shown that callers tolerate speech recognition errors where there is graceful error recovery, that is, successful recognition of a problematic word within three system correction turns, otherwise offering the number of a human agent to handle the enquiry instead.

Naive callers assume the dialogue will be unproblematic and, when offered introductory help, they fail to take up the offer. Indeed, the very naturalness of the synthetic speech appears to raise callers' expectations of the system's capabilities, and we found that it is essential to announce to the caller that they are talking to a computer. Even then, they credit it with human sophistication unless the dialogue breaks down completely. A common example of this is when the system asks for the spelling of a city name it cannot recognise and the caller replies "M for mother, I for ice cream, L for lemon, A for apple, N for nothing". Thus, they are using their knowledge of human speech recognition limitations over the telephone without considering the limitations of computer speech recognition.

The SUNDIAL system uses an endpointing tone to indicate to the caller when to start speaking. In simulation dialogues using a tone, naive callers would wait for the first tone emitted at the beginning of the dialogue but, thereafter, they would attempt to proceed with a response after the system asked them a question, forgetting to wait for the tone. Simply reminding the caller to speak after the tone often resulted in the caller repeating their initial enquiry rather than simply repeating their last utterance.

Example dialogue

Seven of the ten subjects in this trial, examining the effectiveness of optional help, did not request the help offered in the introductory message. There were comments that it sounded 'alarmist' but several subjects said they simply were not expecting such an option in the opening message and did not take it in. This suggests that they did not foresee any need for guidance on how to converse with a computer. There is a suggestion that subjects hearing the help message might interpret *"Please speak clearly and in short*

sentences" as an instruction to use telegraphic speech (lines 19-23).
Note that in this simulation there was no endpointing tone.

W = wizard C = caller (silences in seconds)

1 W Welcome to the computer service for
2 British Airways flight arrival and
3 departure enquiries.(.6)
4 If you would like information on what
5 to expect please say help (.) otherwise
6 go ahead with your enquiry. (4)
7 C help (1.3) Help ? (1)
8 .hhh huhh. .hh = -h
9 W -you are talking to
10 a computer (.) so please speak clearly
11 and in short sentences (.)
12 C .mm.
13 W it is very helpful if you can
14 provide a flight number in your enquiry
15 (.4) if at any time you cannot understand
16 my reply (.) you may ask me to repeat it
17 (.5) please proceed.
18 (1.7)
19 C f:light bee ay two two seven (2)
20 heathrow to montrea:l (1.7) thursday
21 (1.2) I think it's due to leave at two
22 forty five. (1) pee em (1)
23 can you tell me the time please ?
24 (2.5)
25 W You want to know the departure time ?
26 (.5)
27 C departure time (.) yes.
28 (6)
29 W bee ay two two seven from heathrow
30 to birmingham (.) on thursday october
31 the eighth,
32 (3.3)
33 C no:.
34 (7)
35 W please tell me where the flight
36 leaves from.
37 (.)
38 C heathrow
39 (4.4)
40 W please tell me where the flight
41 arrives.
42 (.)
43 C montrea:l:
44 (3)
45 W to melbourne ?
46 (.6)
47 C mon tray (.) al:
48 (4)
49 W I am having trouble with the name
50 of the arrival city (.6) can you
51 please spell it for me.
52 (.6)
53 C em: (.3) oh: (.3) en: (.5) tee: (.4)
54 ar:: (.3) ee:, (.3) ay: (.3) ell::.
55 (1.8)
56 W to montreal,
57 C yes:.
58 (4.3)
59 W flight bee ay two two seven to
60 montreal, (.) leaves london heathrow
61 terminal four, (.) at twelve thirty (.)
62 on thursday october the eighth
63 (2.3)
64 C thank you
65 (.8)
66 W goodbye,
67 C good bye,

5. DISCUSSION

In the absence of any experience with continuous speech dialogue systems, naive callers base their responses to what they believe to be a speech system on their conversational behaviour with other humans. They simplify their grammatical structures but they expect the system to handle long utterances and give an indication if the response will be delayed by more than a second or two. The callers' reaction to system silences has generally been a slight sense of discomfort that "it doesn't say anything while it's thinking" and a tendency to respond before hearing the system's endpointing tone. In human conversations, silences of more than one second represent transitional-relevant places in the dialogue [5]; turn transitions are suppressed by the use of "um" and "er" fillers. We are planning to provide the auditory equivalent of an hourglass cursor to reassure the caller that the system is still on the line and holding the turn.

It is interesting to note that telephone answering machines (with which all subjects were familiar) use a tone to prompt for a complete message. This may influence naive callers' tendency to give all relevant information after the first tone and to misinterpret requests to speak after the tone further into the dialogue. It would seem that the concept of endpointing requires some simple explanation to the callers, along the lines that the system cannot hear; it can only listen, and only when it is ready to do so, as indicated by the tone.

We have found that one strategy to overcome confusion about what to expect is to guide callers explicitly about 'how to get the best from the system', in the form of advertising. Current work suggests that this greatly improves callers' understanding of what to expect of the system.

Given a system which answered calls promptly and could give callers the solution to their query within three minutes of negotiation, callers say they would prefer it to the existing (human) British Airways flight enquiry service. Many of them positively enjoy using the simulated version, which fulfills these objectives, though it has until recently been somewhat futuristic in its sophistication.

CONCLUSIONS

The series of Wizard of Oz simulations described above have provided valuable input to the design of the SUNDIAL system. They have alerted us to naive callers' high expectations of current technology but also to their willingness to adapt their normal behaviour to the abilities of the system. Simulations have enabled us to focus our design efforts on the most problematic features of conversations with a spoken dialogue system without having to wait until the final evaluation to discover them.

Spontaneous speech poses a challenge to current speech recognition systems and it is evident that user guidelines with explicit examples help to constrain initial utterances, which would otherwise be unacceptably long. On the basis of experience gained from simulations, we have been able to build confirmation and error recovery strategies into the dialogue management within the SUNDIAL system. This allows callers to maintain a mixed-initiative, flexible dialogue with the system.

Acknowledgements

This project is partially funded by the Commission for the European Communities ESPRIT programme, as project P2218. The partners in this project are: CAP Gemini Innovation, CNET, CSELT, Daimler-Benz, Erlangen University, INFOVOX, IRISA, Logica, Politecnico di Torino, SARITEL, Siemens, Surrey University.

REFERENCES

[1] Wooffitt, R, Fraser, N, Gilbert, G N & McGlashan, S Designing Interaction: a conversation analytic study of human (simulated) computer interaction (London: Routledge, forthcoming).

[2] Fraser, N M & Gilbert, G N "Simulating Speech Systems", *Computer Speech & Language* 5, 81-99, 1991.

[3] MacDermid, C "Human factors aspects of the design of a speech understanding and dialogue system" Presented to the IEE Colloquium on Special Needs and the User Interface, London, January 1993.

[4] House, J, MacDermid, C, McGlashan, S, Simpson, A & Youd, N J "Evaluating synthesised prosody in simulations of an automated telephone enquiry service" Proc. *Eurospeech'93* (this issue), Berlin 1993.

[5] Fox, B Discourse Structure and Anaphora Cambridge University Press, 1987.