Proceedings of the Institute of Acoustics

IS IT A BIRD? IS IT A PLANE? NO, IT'S A DIALOGUE SYSTEM: TOWARDS A GREATER USABILITY IN AUTOMATED DIALOGUE SYSTEMS

Christine Cheepen (1)
James Monaghan (2)
David Williams (3)

(1) University of Surrey, UK
(2) University of Hertfordshire, UK
(3) Vocalis, Great Shelford, UK (now at Motorola, Basingstoke, UK)

1. INTRODUCTION

Many of the automated telephone dialogue systems currently in commercial operation are intended to be used by the general public. This means that the typical human callers who use the system/s are essentially novice users who do not build up expertise over a period of time. Usability issues are of central importance for this kind of 'novice' user group, and dialogue designers expend considerable effort attempting to make dialogue systems easy and 'natural' for the caller to use. Until recently this has meant designing automated systems according to the human-human spoken dialogue model. Designers have, in fact, approached their task by trying to *disguise* automated systems as pseudo-humans. The literature refers to providing a system with a "likeable personality" [11]; it is claimed that "To develop usable real-world applications it is necessary to ... improve the naturalness of the interaction" [12]; there are recommendations to give dialogue systems "a warm and friendly 'personality'" [23]; some researchers take an extreme position, stating that "ideally the caller should not realise he is talking to a machine" [2].

At the University of Surrey, within the ESRC-funded project "Design guidelines for advanced voice dialogues" we have, over the past two years, carried out a series of experiments where the results challenge the notion of modelling human-machine dialogues according to the principles of human-human dialogue, and indicate that greater usability may be achieved by building into dialogue systems some explicit signalling to the caller that the system *is* a machine rather than a simulation of a friendly, helpful human being. These experiments focus on two aspects of spoken dialogue - the wording of system output utterances, and the meaning of different kinds of system silence.

2. NATURALNESS IN WORDING

Research into spoken discourse has established that dialogue can be viewed as primarily 'transactional' (i.e. goal-directed) or 'interactional' (i.e. person-directed) [5,6]. Between human co-conversationalists, even heavily transactional talk typically contains sections of interactional material, particularly at the opening and closing sections, and at points where there is a topic shift or some disruption to the ongoing flow because of a misunderstanding of some kind.

The interactional linguistic material which occurs in transactional telephone-mediated dialogues between human participants tends to be welcoming, friendly and polite. Human agents dealing with transactional telephone calls signal all these qualities in the lexical and grammatical content of their dialogic contributions, in order to foster a cooperative construction of a shared dialogue space and orientation. Analysis of such human-human dialogues shows that callers collaborate in this interactional exchange [7], and that it functions pragmatically to optimise information transfer and to allow the participants to move smoothly towards their overall transactional goal.

This strategy of progressing transactional dialogue by the inclusion of interactional material which is welcoming, friendly and polite, is evidently the natural way humans do this kind of dialogue. Designers have attempted to simulate this kind of naturalness in automated dialogues by including in system prompts a variety of lexicogrammatical tokens which, in a very general way, function in human-human dialogue to communicate friendliness, politeness and a degree of informality.

2.1 Guidelines experiment on naturalness in wording of system prompts

Our overall aim in this experiment was to test whether this kind of lexicogrammatical signalling of naturalness does enhance usability. We were concerned to investigate two particular areas - callers' ability to achieve transactional success, and callers' opinions of 'natural' and 'non-natural' (i.e. more machine-like) dialogues.

2.1.1 Experimental design. We selected the application domain of telephone banking, because of the large number of such systems already in operation in the commercial sector. We carried out our investigation by comparing the performance of a set of subjects when using two dialogue systems - one with an interactional strand built into the system prompts, which we call the 'original' version, and one where the interactional strand had been removed, which we call the 'denatured' version. Both versions were interfaced with an underlying dummy database of bank account information.

2.1.2 Dialogue prompts. The original version was the set of prompts which were recorded for a telephone banking system which is currently in use in a commercial situation.

The interactional material which occurred in the original prompt set fell into three major categories:

- politeness tokens

e.g. please, thank you

- personal pronouns

e.g. I, you, your

- conversational constructions

e.g. in a moment I will ask you to tell me the account number

Working from the initial set of prompts, we wrote a parallel set, removing all the purely interactional material, to produce our denatured version. In some cases this was simply a matter of removing an explicit politeness token, for example *After the tone please select the account* became *After the tone select the account*, and some personal pronouns were simply replaced by the, for example *speak your Telephone Banking Number* became *speak the Telephone Banking number*. In other cases, however, in order to avoid using personal pronouns, the grammatical structures were also changed, for instance *I'm sorry I didn't understand that* became *Not understood*.

Throughout the 'translation' process we kept our denatured version as close to the original as possible. At no point did we make any additions to the original prompts. Where more than minimal changes were required in order to denature the prompt, the alteration was always in the direction of reducing the prompt, rather than augmenting it, while retaining all the transactional content necessary for the caller's task to be achieved. From a purely transactional (i.e. directly goal-driven) point of view, then, the denatured version of the system provided the same functionality as the original.

2.1.3 Subjects and tasks. We carried out two phases of the experiment, with 12 callers for the initial, pilot phase, and a further 22 for the main phase. The subjects had general domain knowledge (e.g. familiarity with banking accounts etc.) but no specific system knowledge. We asked the subjects to call both versions of the system and carry out a set of four, scenario-based tasks - bill payment, balance enquiry, transfer of funds and statement request.

The dialogues between system and caller were taped, and after completion of the task set with both versions, callers were asked to evaluate the dialogues. Specifically, this involved filling out a questionnaire which simply asked which version they preferred and why.

2.1.4 Results. We used both objective and subjective measures to analyse the results of our experiments. Our objective measures involved counting transactional errors (i.e. breakdown of dialogue, inability to carry out particular tasks), and timing the task-based dialogues for the different versions. We found that the denatured version was as efficient as the original version in terms of transactional success - i.e. callers encountered no particular difficulties in carrying out their tasks. We also found that the denatured version resulted in a slightly shorter transaction time (though this was not statistically significant), due to the removal of some of the linguistic content in the system prompts.

The subjective measure was the sum of the callers' evaluations. The overwhelming majority of callers in both the pilot and the main phase of the experiment preferred the denatured version to the original. No subjects preferred the original to the denatured. The reasons given for preferring the denatured version were consistent across the evaluation questionnaires. Although, as we have said, the denatured version did, in fact, result in slightly shorter dialogue times, callers perceived them as considerably shorter. This response was reinforced by many subjects, who commented on their evaluation forms that the denatured version (they knew it simply as the 'second version') was "not so long winded", "more to the point", even (in one case) "more user friendly". (For a more detailed account of the experiments, see [15].)

2.1.5 Design recommendations. The experiments described in the previous section have indicated that the use of interactional language tokens in heavily transactional dialogues offers none of the advantages designers aim for, and that equivalent denatured prompt sets perform equally well in terms of overall transactional success. Using interactional linguistic material provides no advantage in transaction times, and results in negative evaluations in terms of perceived speed and general preference for users.

Clearly, the absence of interactional language - or 'naturalness tokens' - in automated system prompts does not result in usability problems for novice callers. On the contrary, our experimental subjects indicated a clear preference for talking to a machine which did not - even to a minimal extent - emulate the interactional behaviour of a human agent.

3. THE MEANING OF SYSTEM SILENCE

At the lexical level, a typical human-computer dialogue in an aural-only spoken language system consists of two stages, system output and user input. As with human-human conversation, a good proportion of turn taking clues are given by lapses in talk. Unfortunately, in telephone-based automated spoken dialogues, silences on the system's part are not always easily resolved.

Current spoken dialogue systems in the commercial sector typically fail to represent different system modes to the caller, and usability problems can arise when a caller misunderstands the meaning of system silence. The caller may, for example, assume that the system is silent because it is processing some caller input, whereas the system may not have 'heard' the preceding input and be waiting for the caller to speak. Stifelman identified this problem in Apple's VoiceNotes prototype. In this system, out-of-vocabulary recognition caused the system to remain silent and await a correct utterance. Stifelman notes:

"if the user spoke a command and VoiceNotes did not respond, rather than repeat the command, [as was expected], users waited for a response, thinking the system was still processing the input or busy performing the task" [Stifelman, [14]: pp 184].

Other researchers have carried out work on this problem. Gaver [10] and Mynatt & Weber [13] used 'auditory icons', actual or stylised samples of real-world sounds, e.g. machine sounds, in a process control interface. Gaver emphasised the need to use sounds which were present in the real world since they allowed users to make typical interpretations. Informal analysis showed these sounds were good at representing parallel events. However, if continuous sounds were used to represent system state - e.g. Gaver used a continuous machine sound whose pitch was a function of the system's activity - they could be intrusive. Dutton et al [8] used a mixture of auditory icons and more arbitrary sounds to represent system events and modes. Subjects were tested for recall and preference. Results showed that the more concrete and less arbitrary the icon the more easily they were remembered by the subjects.

Within the spoken dialogue community, AlTech's [1] banking demonstrator uses a processing sound to identify that the recogniser is processing an utterance, a solution analogous to the visual egg timer. The WAXHOLM conversational system [3] uses paralinguistic utterances to represent different processing states, e.g. "Ummm" to signify the start of a long process and "Aha" to signify the system recognising a change in conversational topic. No formal investigation of the effectiveness of the sounds was carried out.

In summary, the use of auditory icons is more effective where there are concrete referents. If this is the case, users can bring their world knowledge to bear on interpreting icons. Icons also provide a short-hand for system output which may reduce transaction time in spoken dialogues. Given the effectiveness of auditory icons, the question remains, which ones are most suitable? In the case of the recognition phase of spoken dialogues, the processing state is a concrete one, which lends itself to an aural representation reasonably easily. The listening mode/state, however, is much more abstract, and calls for either a metaphoric or abstract aural representation. Assigning aural representations to system states is not, then, a simple matter. Some key questions are:

- Do the sounds encourage the correct behaviour within a particular context?
- Does correct sound interpretation depend upon user familiarity with spoken dialogues?

3.1 Guidelines experiment on signalling different meanings of silence

Our overall aim was to compare usability in two versions of a dialogue - one using beeps and silence - which we called our traditional version, and the other using a 'processing sound' and a 'listening sound' - the auditory icon version [4]. Specifically, we focused on callers' initial responses to unfamiliar sounds in automated dialogues, and measured the time taken by callers to respond appropriately to the system prompt. We used talkover technology for all conditions including the 'speak after the beep' control condition. Therefore it was possible for the caller to successfully complete the dialogue even though they spoke before the beep.

3.1.1 Experimental design. The research team decided to keep the dialogue for this experiment as simple as possible, in order to avoid the possibility of domain-dependent effects which could confound the results. If, for example, we had selected a banking application, this may have resulted in longer pauses (while callers thought about the implications, or perhaps carried out calculations) before responding to the system prompts. We therefore designed a dialogue which required the caller to say a single digit between zero and nine. The system then processed the caller input and provided a confirmation in the form "was that X? say yes or no". If the caller said no then the dialogue loop was repeated. One repair mechanism was included; if the system failed to obtain caller input (if the caller did not speak when asked), it asked the caller to repeat the last utterance.

3.1.2 Subjects and auditory representations. We used 28 subjects for the experiment. These were selected to represent the novice and the expert caller (expert in this case referring to expertise with automated dialogue systems). In our traditional version we used beep-silence to represent the listening state, and silence for the processing state. In our auditory icon version we used a sonar sound for the listening state and a stylised processing sound for the processing state.

3.1.3 Results. Subjects tended to respond more quickly to system prompts in the traditional version than in the auditory icon version. This result, however, is because the expert subjects responded much more quickly than the novices, which biased the overall figures. However, when we considered subjects' reponses to the auditory icon version in isolation, it was clear that there was a stronger effect in the opposite direction. Experts tended to listen to the auditory icons because they were at odds with their extensive experience of spoken dialogue applications. It was also noticeable that on average the novice users spoke slightly *before* the auditory icon was heard. This suggests that the auditory icons for the recogniser listening state did not contribute to usability, because the imperative intonation of the system prompt induced correct caller behaviour without any additional auditory clues.

Enhanced usability was, however, displayed in the auditory icon version when we considered the case of caller errors. Fewer errors were made by novice and expert callers in the auditory icon condition, and there were no caller interruptions during the processing state. For novice users this result can be attributed to the auditory icon in the second system request, "Was that X? Say yes or no", where unlike the first request, subjects did not interrupt and so heard the icon.

3.1.4 Conclusions and future work. In summary, the experiment showed that auditory icons need to be carefully chosen for the particular state under consideration and the typical user experience. Overall, icons caused fewer errors than the control condition for all users. The experience of callers with spoken dialogue systems had a strong effect on the success of auditory icon deployment. Expert users *listened* to 'recogniser listening' icons rather than immediately speaking, which is in stark contrast to their speed of response to the traditional version. This suggests that a period of acclimatisation is required.

For novice users, icons may not be required where the imperative intention of the system is made clear by prompt wording and/or intonation. There was some indication that in less well signalled systems with states such as processing, an auditory icon will deter both novice and expert users from making futile verbalisations.

Future work in this area should include experiments to examine the use of sound ecologies which represent a wider range of system states and events, such as particular recogniser errors, and application events, e.g. a new message has arrived, an operator is unavailable and the like. Of particular interest is the possibility of using auditory representations to replace verbal prompts thus shortening transaction time for repetitive tasks.

4. REFERENCES

[1] Applied Language Technologies, http://www.altech.com/

[2] Aust, H. 1996, Dialog Modelling, Elsnet Summer School, Budapest

[3] Blomberg, M, 1993, An Experimental Dialogue System: Waxholm, Proceedings of Eurospeech '93

[4] Brewster , S A., P.C.Wright, A.J.Dix & A.D.N.Edwards, 1995, The Sonic Enhancement of Graphical Buttons, Proc. of Interact'95

[5] Brown, G. & G.Yule, 1983, Teaching the Spoken Language, CUP

[6] Cheepen, C., 1988, The Predictability of Informal Conversation, Pinter Publishers

[7] Cheepen, C. & J.Monaghan, 1998, Designing for Naturalness in Automated Dialogues: some problems and solutions, in Wilks, Y. (ed) 1998 (in press), Machine Conversations, Kluwer

[8] Dutton, D, C.Kamm & B.Boyce, 1997, Recall Memory for Earcons, Proc. of EuroSpeech '97

[9] EAGLES Handbook on Spoken Dialogue Systems, 1996

[10] Gaver, W., 1989, The SonicFinder: An Interface that Uses Auditory Icons, Journal of HCI, Vol. 4

[11] House, J., C. MacDermid, S. McGlashan, A. Simpson & N. Youd, 1993, Evaluating synthesised prosody in simulations of an automated telephone enquiry service, SUNDIAL Report

[12] Junqua, J-C. & P. Morin, 1994, Naturalness of the interaction in multimodal applications, Proceedings of ICSLP-94

[13] Mynatt, E.D. & G.Weber, 1994, Nonvisual Presentation of Graphical User Interfaces: Contrasting Two Approaches, Proc. of ACM SIGCHI'94, New York: ACM Press

[14] Stifelman et al., 1993, Voice Notes, Proc. of INTERCHI'95, New York: ACM Press

[15] Williams, D. & C.Cheepen, 1998, 'Just speak naturally': Designing for naturalness in automated spoken dialogues, Proceedings of CHI 98

5. ACKNOWLEDGEMENTS

This research has been carried out within the context of the ESRC-funded project Design guidelines for advanced voice dialogues, under the Cognitive Engineering programme, project no. L127251012. Project partners are Department of Sociology, University of Surrey, Guildford, and Vocalis Ltd., Great Shelford, Cambridge.