September, 2005

Root causes of lost time and user stress in a simple dialog system

Nigel Ward, *The University of Texas at El Paso*
Anais Rivera, *The University of Texas at El Paso*
Karen Ward, *The University of Texas at El Paso*
David G. Novick, *The University of Texas at El Paso*

Root Causes of Lost Time and User Stress in a Simple Dialog System

Nigel G. Ward, Anais G. Rivera, Karen Ward, David G. Novick

University of Texas at El Paso
El Paso, Texas, USA
nigelward@acm.org, agrivera@utep.edu, ward@up.edu, novick@utep.edu

Abstract

As a priority-setting exercise, we compared interactions between users and a simple spoken dialog system to interactions between users and a human operator. We observed usability events, places in which system behavior differed from human behavior, and for each we noted the impact, root causes, and prospects for improvement. We suggest some priority issues for research, involving not only such core areas as speech recognition and synthesis and language understanding and generation, but also less-studied topics such as adaptive or flexible time-outs, turn-taking and speaking rate.

1. Introduction

Commercial spoken dialog systems generally do not use the latest, most powerful techniques. It is also the case that the user experience for today's spoken dialog systems falls short of the ideal. This suggests a question: to what extent do the weaknesses of common dialog systems reflect, on the one hand, a lag in the commercial application of capabilities already demonstrated in research systems, or, on the other hand, a need for further research advances. In either case, we also wish to identify the specific issues that need attention.

Thus the main aim of this study is to determine some priorities for both practitioners and researchers, in order to ultimately make spoken dialog systems more usable.

2. Methods

To achieve this goal we developed a new way to analyze dialogs and systems. The basic idea is to have subjects perform the same task with both a spoken dialog system and a human operator. This enables within-subject comparisons of the two interactions and enables us to go beyond the identification of clear errors, to also identify missed opportunities for better performance. We compare system performance to human performance simply because human-human dialogs often have many properties worth emulating.

When attempting to set priorities for a research field, there can be a tendency to be visionary, targeting very challenging goals, or a tendency to be grounded, targeting problems salient in existing systems. This study takes a compromise approach: it is visionary in that it uses human performance as the gold standard, but grounded in that it focuses on a practical domain and observed needs. We build on previous attempts to relate usability to a system's technical properties [1, 2, 3]; however our purpose is not to guide design nor to evaluate systems but rather to identify research priorities.

2.1. Domain and System

We chose billing support as the test domain. While dialog systems research has largely moved beyond such simple tasks, interactions at this level of complexity are of great practical importance and are still challenging to implement well.

To make the comparisons simple and provide more direct answers to our questions, we would have liked to build the best system possible with currently available commercial technology. In fact we could only devote about 200 person-hours to development, including design, coding, testing, and debugging. The resulting system, built on Nuance's Voice Platform, has about 48 states. The back-end was stubbed. Although functional, the system was not highly polished. In particular the prompts, grammars, and time-outs were not tuned, and the prompts were synthesized rather than recorded speech. However, in initial evaluation the system's overall performance seemed to be in the same league as many deployed systems — certainly not comparable to the best but good enough to use as a proxy for the commercial state-of-the-art.

2.2. Protocol

The subjects were 20 lower-division Computer Science students, of whom 11 were native speakers of English, all with little or no experience using spoken dialog systems.

For each interaction, subjects were given a mock credit-card statement, a mock bank statement, and a brief checklist of three tasks to complete. They were also instructed verbally regarding the tasks, which were to obtain balance information, to review the most recent transactions, and to make a payment. Instructions were kept simple so that subjects would know what they needed to accomplish but not how. In the system-based interactions the subjects were informed that they would be using a spoken dialog system and that they should speak to it as they would with a person. The interactions with a human operator were constrained to be roughly comparable by showing the operator the system's prompts and dialog flow and asking her to use mostly the same vocabulary and roughly the same level of formality.

Each subject performed the task with both the system and the human operator, in balanced order. The scenarios were similar, although with different names and numbers. Interactions were recorded and videotaped in both conditions, giving two subcorpora. After both interactions subjects completed a written questionnaire and were debriefed.

2.3. Analysis

The dialogs and questionnaires were analyzed in the usual ways, plus two others.

First was the process of examining usability events. These

Operator: Okay you just requested a payment of 50 dollars using check number 51. Is this correct?
Subject: Uhm *451*.
Operator: *451*?
Subject: Uh-hm.
Operator: Okay, your payment has been processed.
Operator: Is there anything else I can help you with?
Subject: Umh . . . can I know, umh . . . about other purchases that I did?
Operator: Certainly. You have a debit . . .

Figure 1: Error Recovery and Context-Appropriate Feedback in a Human-Human Dialog

were of two kinds. First, there were times in the human-system dialogs where something unfortunate or sub-optimal happened: specifically occasions where a human operator could have done better than the system. Second, conversely, there were times in the human-human dialogs where the human operator did something appropriate that the system could not have done. Sometimes direct comparisons between the two dialogs for one user were possible, but other observations relied on comparing patterns seen across the two subcorpora. A total of 115 usability events were noted in the human-system dialogs and 62 in the human-human dialogs. This process was not formalized, but we did use a checklist to provide some structure to the process (see `http://www.cs.utep.edu/nigel/dialog-usability/`). More about this process appears elsewhere [4].

We also examined the dialog activites to determine where time was lost when interacting with the system. Overall the average time to complete a task was 130 seconds with the system and 40 seconds with the human operator. We did a coarse hand-labeling of the dialog activities to see where the time was being lost [4]. Some of the time losses were easy to attribute to specific causes, but an unexpectedly high 16% of the time loss was due to silence on the user's part, that is, declining to take a turn or delaying before a response. This was a symptom of several underlying problems, discussed below.

3. Initial Observations

Overall, the subjects completed a total of 35 tasks with the system and 46 with the human operator, out of 60 possible in each case. A few of the non-completions were due to unrecoverable system failure, all of which were easily fixable in retrospect. However many were due to subjects simply forgetting, or not bothering, to do a task. In particular, some non-completions seemed to be due to subjects' being disconcerted, annoyed, or stressed by the unsatisfactory nature of the system's interactions, of which more later. If this is a general phenomenon, it means that failure to provide "the niceties of dialog" can affect task completion, a bottom-line aspect of user satisfaction, at least when users are not strongly motivated.

One striking aspect of the dialogs was that they took place at two levels. One was the desired level, where the system utterances were timely and appropriate responses to user utterances, and the dialog flowed much as seen in the human-human interactions. However most dialogs were only intermittently at this level: most of the time interactions were at a more basic level, with the user producing single-word commands and the system giving simple reprompts that basically just informed the user of the system's current dialog state. This typically happened

System: You just requested a payment of one hundred United States dollars, with check number five hundred and twenty one. Is this correct?
Subject: No
System: Ok, what is the correct check number?
Subject: five eleven
[. . .]
System: You just requested a payment of one hundred United States dollars, with check number five hundred and eleven. Is this correct?
Subject: Yes
System: What was that?

Figure 2: Recovering from Error in a Human-System Dialog

after time-outs or recognition errors. In contrast, after breakdowns was where the operator really shined; at such times clear concise feedback, deft use of non-lexicals, swift turn-taking, and other behaviors enabled quick, painless recovery. Figure 1 presents an example.

It was sobering to find that most of the user experience was at this basic level, especially since most of our design effort had been at the higher level. Unfortunately, much of current design know-how, e.g. [5, 6], seems to be similarly focused on the better interactions, which are less frequent in practice. For example, in all the literature we surveyed, there were but two paragraphs giving concrete guidance on appropriate values for time-outs, a parameter of great importance whenever the smooth dialog flow breaks down and users revert to basic level interaction. Thus a clear priority is the development of useful human-factors knowledge on such topics.

4. Issues

The usability events and time differences arise from a number of system properties and capabilities. These are complexly interrelated [3], and crosscut many traditional concerns of dialog management. The issues are ranked in order of importance, as judged by their frequency and their impact.

4.1. Recognition and Understanding

The top issue was accuracy of recognition and understanding; failures of course had high impact in terms of task completion, time, and user satisfaction. Much of the impact of the recognition errors was felt during error recovery, which accounted for about 27% of the lost time. Error recovery was often awkward and time-consuming with the system, as illustrated in Figure 2. In contrast, the human operator could easily detect, diagnose, and recover from errors, as seen in Figure 1.

Recognition failures may also have brought a hidden cost in terms of demands on the user. For example it seemed that non-native speakers made more effort to use a standard accent when dealing with the system rather than the operator. Changes in user behavior also led to other problems. Misrecognition of fillers, or users' fear of them, probably accounted for the limited use of fillers in the system dialogs, leading sometimes to awkward turn-taking. In addition, recognition problems, or the fear of them, probably also accounted for much of the user silences. These often seemed to be due to users spending time thinking what to say and how to say it, presumably because they thought the system would do better with utterances that were specific, well-formed and concise.

The problem of recognition failures indirectly caused another problem: the need to guide users to produce utterances easy to recognize. During development we did this by making some of the prompts rather detailed and explicit, which accounted for much of the 29% of the time cost which was due to longer system utterances.

4.2. Time-Outs

One recurring problem was inappropriate time-outs. Waiting for time-outs is of course awkward in that each party is silent and waiting for the other, a situation that is generally avoided in human-human dialog.

Our system used a fixed time-out; that is, after a fixed amount of user silence the system reprompted. Sometimes this was too short, resulting in the system re-prompting during the users' "think time," interrupting as they were trying to understand what the system expected, formulate their own next goal, or decide what to say. This occurred more often for those users who, when interacting with the system, did not use fillers to claim the floor nor disfluency markers to keep it. At other times the time-out was too long, meaning that users wanting the system to give follow-up help were left waiting. And sometimes it seemed to be both, in cases where users seemed willing either to be guided or to think things through themselves, but the time-out was an awkward intermediate value, with the result that both user and machine started talking at the same time.

4.3. Responsiveness

The human operator was fast; there was seldom dead time between the user's utterance and her response. In contrast, the system often delayed before responding; these delays accounted for about 20% of the time cost.

The operator was clearly sensitive to the turn-taking cues. She could usually tell whether the user had more to say or was finished. In contrast, the system often responded too slowly and sometimes too quickly, cutting off the user. Some of these problems seemed to be due in part to unsophisticated endpointing [7].

Another cause of slow responses was the processing time required for speech recognition. Beyond direct speed-ups, some behaviors of the human operator suggest another strategy to alleviate this; she seemed to be giving some responses before fully processing the user's utterances. Figure 3 presents an example where the operator gave a swift response that was appropriate at one level, and then recovered gracefully when she more fully realized what the situation required. Two common cases of this were her interpolation of back-channels between number chunks [8], and her use of fillers, actions that appeared to effectively meet user expectations. Thus she seemed to be processing and responding to the input 'asynchronously' on multiple levels at once [9].

Responsiveness seemed to become relatively more important when the dialog departed from the desired path. In particular, swift exchanges were common during error recovery in the human-human dialogs but painfully absent during error recovery with the system.

4.4. Synthesis

Although intelligible and not unpleasant, the synthesized utterances of the system were inferior to those of the operator. First, the speaking rate of the synthesized voice was fixed at a moderate pace. Although necessary for intelligibility, this was a cause

Operator: . . . How may I help you?
Subject: Hi, I just, I have a, payment due tomorrow. I just need, to know the uh, the uh amount I need to pay. And to do a payment.
Operator: Your minimum payment due, um, what is your account number?

Figure 3: An Example of Responsiveness

of longer system prompts and thus lost time. Second, sometimes the prompts confused the users, probably because the prosody of the system utterances was not always what the users expected for the discourse context. For example the prosody of the system prompts strongly discouraged users from barging in, although barge-in would have been a valuable way for users to deal with over-long prompts.

4.5. Feedback

One of the reasons why users were sometimes confused and slow to respond may have been system utterances that were inappropriate for the local dialog context. This problem was not at the semantic or task levels; indeed, the system generally succeeded in conveying the information required to accomplish the task and in indicating task progress and dialog structure (with discourse markers like "okay" and "now"). Rather the problem was that the system failed to provide utterances that were entirely situation-appropriate. By comparison, the operator's utterances were generally appropriate for the local context and also at an interpersonal level; the "certainly" at the bottom of Figure 1 is an example.

One common type of feedback indicated dialog status. The operator let the user know (that the operator knew) what the current activity was, such as finding and fixing an error, or returning to the main task after a sub-dialog. For example, at "okay" back in turn 5 of Figure 1, the operator's tone of voice seemed to convey reassurance that the dialog was back on track.

4.6. Adaptation

The human operator was good at adapting her 'dialog style' to that of the user. Although some of these adaptations seemed to relate to user's personality, most were relatively straightforward to characterize, such as adopting the user's vocabulary, matching the user's level of formality, and adjusting her speaking rate to the user's language proficiency. The latter holds great promise: adaptation of speaking rate [10] could potentially reduce by half the time cost due to the system prompts.

4.7. Other

The other major issues identified were use of prosodic information, use of non-lexical tokens [11], and dynamic generation of prompts.

Other usability-related issues, observed but less important in this domain, include: recognizing dialog acts, managing initiative, modeling complex dialog structure and tracking multiple subgoals, choosing confirmation strategy, understanding in the face of user disfluencies and self-corrections, negotiating meaning, using unsolicited information, managing pre-closings, and handling clarification sub-dialogs. This list of course includes only issues which arose in this study; others would be seen in other domains and with other user populations.

Issue	Potential Impact		
	Time	Completion	Stress
Recognition, Understanding	+++	+++	+++
Time-Outs	+++	+	++
Responsiveness	+++	+	++
Generation, Synthesis	+++	+	+
Feedback	+		++
Adaptation	++		+
Prosody, Non-Lexicals	+		+
Other	+	+	+

Table 1: Some Research Issues and their Estimated Potential Impact on Dialog System Usability

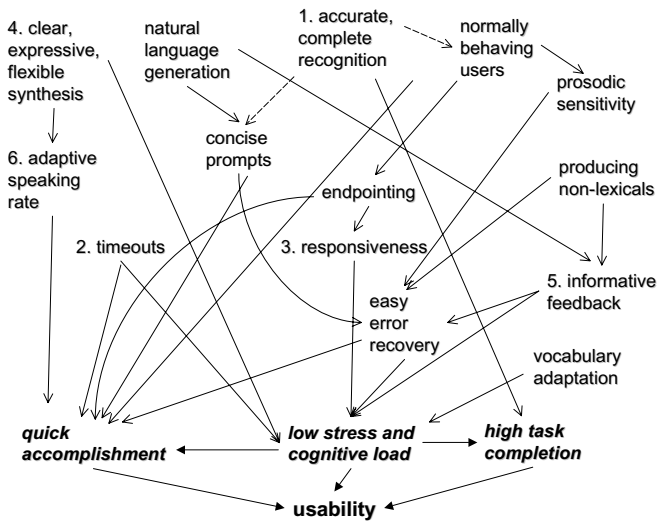

Figure 4: Some Relations Among some System Capabilities and Usability

5. Summary

Although this study was exploratory and the limitations are clear, it is possible to suggest some answers to the question we raised in the introduction: Why are many spoken dialog systems unpleasant to use? Table 1 summarizes our rough estimates of the potential for usability improvements based on foreseeable advances on each of the issues discussed above. Figure 4 suggests some of the ways these seven issues are interrelated, how they relate to some other issues identified in the literature, and how they relate to the bottom line.

Clearly some of these priority issues require industry to utilize recent research findings, for example in such areas as speaking rate control and accurate endpointing. Most issues indicate a need for more basic research: in such core areas as speech recognition and synthesis and language understanding and generation, on more recent areas of interest such as prosody and turn-taking; and on one topic that seems to have been largely neglected: time-outs.

Of course, this sort of analysis is not something to do just once; as the field advances different issues will arise. Ultimately we would like to close the loop: to arrive at a model or method to make the connections between system capabilities and user satisfaction clear and even quantitative. We hope that the methods used here, together with other approaches [1, 3], will make this day come sooner, leading to more focused basic research and ultimately more usable systems.

6. Acknowledgements

We thank Nuance for providing Voice Builder and Javier Aldaz-Salmon for his help in building the system. This material is based upon work supported by the National Science Foundation under Grant No. 0415150.

7. References

[1] Walker, Marilyn, Candace Kamm and Diane Litman. Towards Developing General Models of Usability with Paradise. *Natural Language Engineering*, 6, pp 363-377. 2000.

[2] Dybkjaer, Laila and Niels Ole Bernsen. Usability Issues in Spoken Language Dialogue Systems. *Natural Language Engineering*, 6, pp 243–272, 2000.

[3] Möller, Sebastian. A New Taxonomy for the Quality of Telephone Services Based on Spoken Dialog Systems. 3rd SigDial Workshop on Discourse and Dialog, pp 142-153, 2002.

[4] Ward, Nigel, Anais G. Rivera, Karen Ward, David G. Novick. Some Usability Issues and Research Priorities in Spoken Dialog Applications, Technical Report UTEP-CS-05-23, 2005.

[5] Cohen, Michael H., James P. Giangola and Jennifer Balogh. *Voice User Interface Design*. Addison-Wesley, 2004.

[6] Harris, Randy Allen. *Voice Interaction Design: Crafting the New Conversational Speech Systems*. Morgan Kaufmann, 2005.

[7] Ferrer, Luciana, Elizabeth Shriberg, and Andreas Stolcke. A Prosody-Based Approach to End-Of-Utterance Detection that does not Require Speech Recognition. ICAASP 2003.

[8] McInnes, Fergus and David Attwater. Turn-taking and grounding in spoken telephone number transfers. *Speech Communication*, 43, pp 205-223, 2004.

[9] Lemon, Oliver, Lawrence Cavedon and Barbara Kelly. Managing Dialogue Interaction: A Multi-Layered Approach. 4th SigDial Workshop on Discourse and Dialogue, 2003.

[10] Ward, Nigel and Satoshi Nakagawa. Automatic User-Adaptive Speaking Rate Selection, *International Journal of Speech Technology*, 7, pp 235-238. 2004.

[11] Ward, Nigel. Non-Lexical Conversational Sounds in American English. *Pragmatics and Cognition*, 2005, to appear.