Some Usability Issues and Research Priorities in Spoken Dialog Applications

Nigel G. Ward, Anais G. Rivera, Karen Ward and **David G. Novick**
nigelward@acm.org, agrivera@utep.edu, ward@up.edu, novick@utep.edu
Department of Computer Science, University of Texas at El Paso

Technical Report UTEP-CS-05-23, June 15, 2005

Abstract

As a priority-setting exercise, we examined interactions between users and a simple spoken dialog system in comparison to interactions with a human operator. Based on analysis of the observed usability differences and their root causes we propose seven priority issues for spoken dialog systems research.

keywords: evaluation, human-computer interaction, human-human dialog, usability events, human factors, time, speech

1 Introduction

Commercial spoken dialog systems generally do not use the latest, most powerful models for dialog management. 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.

This paper is an expanded version of (Ward et al., 2005), with more details on the methods, additional observations, qualitative analysis of the time losses, consideration of development issues, and discussion of the limitations of state-based dialog management.

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 (Martinovsky and Traum, 2003).

We compare system performance to human performance because human-human dialogs often have many properties that seem worth emulating. More strongly, to the extent that the usual properties of human-human dialog are not just conventions but reflect fundamental capacities and limitations of human cognition, examination of these properties can be a good way to uncover useful directions. Incidentally, we note that we do not advocate mimicking human behavior as a goal in itself, nor because we think that ultimately computer-human dialogs must be like human-human ones — indeed it may be instead that even the ultimate spoken dialog systems will involve forms of interaction which today seem unfamiliar and unnatural (Heisterkamp, 2003).

Our approach follows work such as (Doran et al., 2001) which categorizes differences between human-computer interaction and human-human interaction, although our aim here is to go further and relate such differences to usability and to technical issues. Other studies have identified spoken dialog systems research issues, e.g. (Zue and Glass, 2000); here we go one step further and attempt to identify those specific issues most likely to have the largest impacts on usability. We also build on previous attempts to relate usability to a system's technical properties and to its objectively-measured performance (Walker et al., 2000; Möller, 2002); however our purpose is not to guide design nor to evaluate systems but rather to identify research priorities.

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.

2.1 Domain and System

We chose billing support as the test domain. While research in dialog management 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. Figure 1 shows a sample dialog between a user and the system.

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. If a subject reached a clearly hopeless state the experimenter intervened and either provided help or terminated the experiment. When possible such problems were fixed before the next subject run, as the analysis of easily fixable problems was not an aim of the study.

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. As the scenarios were unvarying, after a time the operator was able to anticipate the user's goals, although this did not appear to affect her behavior.

System: Welcome to the ISG credit card system. To log in I will first need your account number and your zip code. What is your account number?
Subject: 54 24 598 77 854 69 60
System: What was that?
Subject: 542 4 5 98 77 85 46 960
System: Okay, to verify your identity I will need your zip code.
Subject: 799 54
System: How may I help you?
Subject: Yes, umm I need to make my ... payment but I made some extra purchases and I want to know what's my new balance.
System: What was that?
Subject: I need to know my minimum payment.
System: To make a payment I will first need to record your bank's information, then I will need the check you would like to use and finally the amount you would like to pay. Ok, let's enter your bank's information; first I'll need your bank's routing number and finally your checking account number. What is your bank's routing number?
Subject: 5 4 8 7 96 3 2 1.
System: Is your bank ...

Figure 1: Start of a Typical Interaction with the System.

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 Labeling Usability Events

The dialogs and questionnaires were analyzed in several ways, of which one requires explanation: the process of examining usability events. These were of two kinds. First, there were times in the human-system dialogs where something unfortunate or suboptimal 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. Each usability event was characterized in several ways, including:

1. *Estimated impact*, in terms of time cost, user stress, and task completion.

2. *Usability issue*, for example as one of misunderstanding, violating expectations, giving inadequate guidance, being impolite, being slow, being rigid, giving inadequate feedback, violating turn-taking conventions, and so on. Thus this characterization was in terms of some common ways in which a dialog system can satisfy or disappoint its users (Dybkjaer and Bernsen, 2000).

3. *Technical description of the problem*, for example as one of failure to exploit prosodic cues, failure to adapt dialog pace or style, failure to model the user's feelings, misinforming the user regarding the system's abilities, failure to understand or produce non-lexical utterances, and so on.

4. *Identification of system components involved*, that is, a listing of the system modules which would need to be improved to eliminate the problem.

5. *Fixability*, that is, an estimate of whether the problem was easily fixable, fixable with some effort but doable within the Nuance system, fixable only within a (hypothetical) system incorporating more advanced techniques, or not fixable using any techniques known today.

This process of characterizing events was not formalized. However we did use a checklist to provide some structure to the process (see `http://www.cs.utep.edu/nigel/dialog-usability/`).

The dialogs were observed from videotape by four people. One was the person who had built the system; he served as the authority on the technical causes of errors and difficulties. The second was the person who had taken the operator role; she was the authority on how she had interpreted the users' behavior. The other two were experts on human dialog; they were able to interpret the observations with respect to what is known more generally. The actual users were not involved in this stage, as there was no dearth of problems to study without considering individual subjective impressions. The identification of usability events was not hard — the labelers generally agreed easily on their locations and impacts — however the diagnosis of causes was more speculative, as seen below.

As this process of labeling usability events was largely subjective, it is conceivable that our preconceptions may have affected what we noticed. For the record, we expected that the important issues would include turn-taking, informative feedback, adaptation, and non-lexical utterances (Ward and Tsukahara, 2000; Ward and Tsukahara, 2003; Ward and Nakagawa, 2004; Ward, 2005 to appear). We also expected to see problems due to the limitations of the dialog manager, in particular the fact that it was based on dialog states.

3 Initial Observations

Twelve of the 20 users preferred to interact with the human operator instead of the system. The main reasons given were feeling more comfortable talking to a person and the faster task completion with the operator. On the other hand six users preferred interacting with the system, possibly in part because the subjects were computer science students. Their reasons included feeling more comfortable not discussing financial information with a person, likely better availability and no problem of feeling nervous. Two users had no preference.

From casual observation of the videotapes, there was a clear difference in users' expression and posture: they were clearly more relaxed when dealing with the operator. It would be interesting to quantify this and relate it to specific dialog properties on the one hand and the users' own impressions on the other.

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 appro-

	system-human		human-human		difference	
dialog activity	seconds	percent	seconds	percent	seconds	percent
normal system-side* utterances	1464	32%	693	37%	771	29%
normal user utterances	586	13%	759	41%	-173	-6%
error recovery -related utterances	787	17%	78	4%	710	27%
system-side* silences	654	13%	114	6%	540	20%
user silences	636	14%	214	11%	421	16%
simultaneous talking	46	1%	6	0%	40	1%
experimenter interventions	366	8%	0	0%	366	14%
total	4540	100%	1866	100%	2675	100%

Table 1: Total Times Attributable to Various Dialog Activities. The rightmost column shows the percent of the total difference attributable to each dialog activity. *In the operator condition the "system-side" utterances were those by the operator.

priate 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 after time-outs or recognition errors.

4 Time

The average time to complete a task was 130 seconds with the system and 40 seconds with the human operator. Dialogs with the system totaled 76 minutes in duration and dialogs with the human operator 31 minutes. Table 1 shows where the time was spent. Only one factor mitigated the overall tendency for the system-human dialogs to be longer: users used shorter utterances when talking to the system.

We begin our analysis of the causes of performance difficulties by looking at the data in terms of where the users took more time when interacting with the system.

First, even when things were going well, there was a time cost because the system utterances were much longer than the operator's utterances, accounting for about 29% of the total extra time. Thus, although in human-human dialogs the user's utterances cumulatively took more time than the operator's, in human-system dialogs the system's utterances cumulatively took more than twice as long as the users'.

Second, about 27% of the time difference was due to recognition errors. In part this was because the system made more errors, but mostly it was because it took much longer for users to correct an error with the system than with the operator.

Third, there was much more silence was in the system-human dialogs. We categorized these roughly into user silences and system-side silences, based on which side eventually took the turn. About 20% of the time difference was due to silence on the system's part, that is, delays before responding. These were largely attributable to delays due to processing time and to slow turn-taking. Another 16% 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.

Fourth, about 1% was due to times when both the system and the user were unproductively talking at the same time.

Fifth, about 14% was due to times when the experimenter stepped in to help the user through the task.

5 Issues

The usability events and time differences arise from a number of system properties and capabilities. These are complexly interrelated (Möller, 2002) and crosscut many traditional concerns of dialog management. For purposes of discussion we group them into seven key issues. The issues are ranked in order of importance, as judged by their frequency and their impact.

5.1 Recognition and Understanding

Recognition failures 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 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

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

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

Figure 3.

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 brought a substantial time cost.

5.2 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 (20 instances in total). 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 (40 instances in total).

To some extent these problems of time-outs could have been reduced by decreasing the user's confusion, for example by the use of more appropriate prompts. This is because many subjects used a time-outs as a last resort for getting an appropriate response when all else failed (or, viewed in a more positive light, subjects used silence as a strategy for shifting the system from mixed-initiative mode to directive mode). The problems of inappropriate time-outs could also be reduced directly. For example, it should be possible to make time-outs adaptive, so that they depend on the past context of the interaction.

5.3 Responsiveness

The human operator was fast; there was seldom dead time between the user's utterance and her response. In part this was because she was 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 (Ferrer et al., 2003). Another cause of slow responses was the processing time required for speech recognition. Beyond direct speed-ups, some behaviors of the human operator

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 4: The operator starting to respond before fully considering what to say.

suggest another way to alleviate this; she seemed to be giving some responses before fully processing the user's utterances. Figure 4 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 (McInnes and Attwater, 2004), 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 (Lemon et al., 2003).

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.

5.4 Speech 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 resulted in 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.

5.5 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.

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 3, the operator's tone of voice seemed to convey reassurance that the dialog was back on track.

Another common type of feedback was the use of words like "yeah," "okay," "absolutely," "sure thing" and "certainly" in response to requests, for example at the end of Figure 3. These seemed to become more common if the operator judged that the user needed reassurance. From the user's perspective, it seems that these show an understanding not just of the user's words, but what he or she was trying to accomplish. Users seemed more comfortable and confident when they received feedback of this sort. It is probably significant that these were not always the same token, but seemed to be chosen based on the user's current state, as inferred from the specific dialog context, and their exact words and prosody, in ways that remain to be elucidated.

5.6 Adaptation

The human operator was good at adapting her 'dialog style' to that of the user. Some of the adaptations seemed easy 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 (Ward and Nakagawa, 2004) could potentially reduce by half the time cost due to the system prompts.

There were also more complex adaptations. The need for these was easy to see in the system-human dialogs, where the system worked acceptably for users with certain dialog strategies but not others. (In part we noticed these inadvertently, due to labeler fatigue. After examining a number of dialogs the system behavior came to seem almost normal, and the labelers began to notice more the differences among users.) Some interesting difference included: Some users were clearly testing the limits of the system and the operator, while others were trying only to accomplish the tasks. Some users responded to the "how may I help you" prompt with a full description and justification, while others responded with a specific request. Some users laughed at communications breakdowns, while others seemed to take them as personal failures. Some users tried to learn the system's characteristics and adapt, while others tried to persist in their own speaking style. Some users used filled pauses while others delimited their utterances with silence. Some users seemed to want encouragement or reassurance while others were indifferent. Some users tended to take control of the conversation while others wanted to be guided.

5.7 Prosody, Tone of Voice, and Non-Lexical Utterances

The operator was clearly sensitive to the prosody of the user's utterances. For example, she responded correctly to user saying "thank you" in a tone indicating "I'm done, good bye." She also could detect when the user was talking to himself or herself and, as noted above, when the user felt unsure and needed reassurance or guidance. Finally she responded swiftly to corrections, which were often marked in subtle ways (for example, the "uhm 451" in Example 3), and similarly for turn-grabs and yields.

5.8 Other

Some of the issues identified above, including the duration of system prompts and tailored feedback, could be addressed in part by dynamic generation of suitable prompts. Other usability events observed relate to issues such as: 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.

We note that our study was designed only to uncover usability issues. In practice, other factors, such as the utility of the information available or the attractiveness of the system's voice, may have larger effects on user satisfaction.

6 Development Know-How

Recently many resources for spoken dialog systems design have appeared, including (Suhm, 2003; Cohen et al., 2004; Harris, 2005), which detail how to produce useful systems despite the limitations of today's technology. As compendiums of human-factors know-how, such resources are valuable. Certainly for our system, had we faithfully followed all the design guidelines and development steps, many usability problems would have been seen less often. Of course, given the labor-intensive nature of the design-iterations needed to improve usability, the need for better toolkits and development support is clearly another priority research area.

However, as noted earlier, we found that most of the time our users ended up interacting with the system at a basic level. By comparison, 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. For example, users often seemed to be trying to use guessed keywords to navigate the system into the desired state. Indeed, this was more common than the appropriate exchanges and smooth flow we had envisioned when designing the system. 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 know-how 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 another priority is the development of useful human-factors knowledge on such topics.

7 Beyond State-Based Dialog Management

Most commercial spoken dialog systems are structured around a state transition network, where a state is a packet of information typically including: 1. a set of next states and conditions for choosing which, 2. a prompt, 3. a grammar (a language model), 4. possibly a mapping to a semantic interpretation or an action involving the back-end, and 5. turn-taking parameters, such as a time-out value and a flag indicating whether barge-in is allowed.

We believe that the use of state machines to model dialog is ultimately indefensible; the idea that a system needs to make decisions only at a few time points and the idea that the information needed can be neatly associated with states both seem insupportable. Instead a system could, at every moment, be

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

Table 2: Some Research Issues and their Estimated Potential Impact on Dialog System Usability

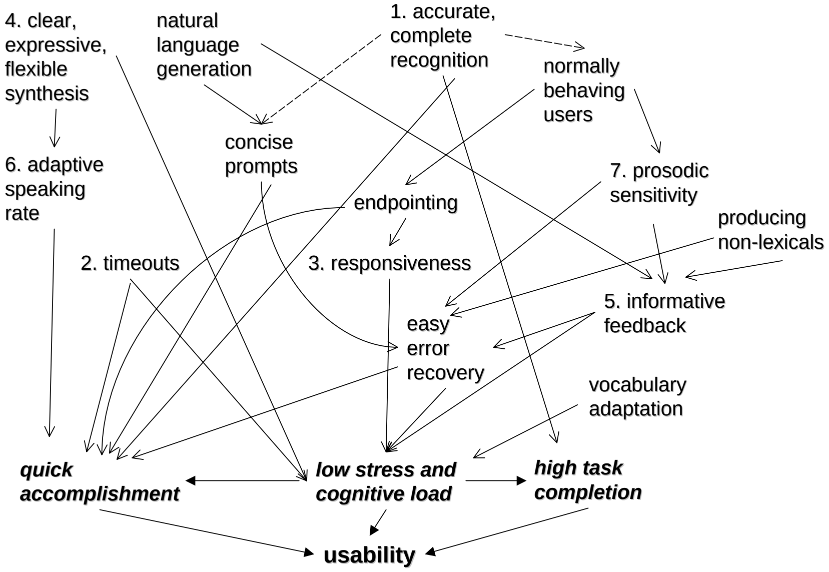

Figure 5: Relations Among some System Capabilities and Usability

deciding what to do next, based on all of the information in the signal so far and on all of the context. Certainly a number of researchers have explored ways to go beyond state machine models of dialog.

Yet contrary to our initial expectation, most of the issues which seem most important here do not relate to limitations of the state model of dialog. Moreover, since state-based systems are intrinsically easier to design, develop, and debug, they will probably be with us for a while.

8 Summary

Although this study was exploratory, it is possible to suggest some answers to the question we raised in the introduction: Why are many spoken dialog systems difficult to use? Table 2 summarizes our rough estimates of the potential for usability improvements based on foreseeable advances on each of the issues discussed above. Figure 5 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.

Some of these priority issues require industry to use recent research findings, in areas such 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 developed here, together with other approaches (Walker et al., 2000; Möller, 2002), will make this day come sooner, leading to more focused basic research and ultimately more usable systems.

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.

References

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

Christine Doran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogs. In *2nd SigDial Workshop on Discourse and Dialogue*.

Laila Dybkjaer and Niels Ole Bernsen. 2000. Usability issues in spoken language dialogue systems. *Natural Language Engineering*, 6:243–272.

Luciana Ferrer, Elizabeth Shriberg, and Andreas Stolcke. 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *ICAASP*.

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

Paul Heisterkamp. 2003. “Do not attempt to light with match!”: Some thoughts on progress and research goals in spoken dialog systems. In *Eurospeech*.

Oliver Lemon, Lawrence Cavedon, and Barbara Kelly. 2003. Managing dialogue interaction: A multi-layered approach. In *4th (ACL) SigDial Workshop on Discourse and Dialogue*.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *ITRW on Error Handling in Spoken Dialog Systems*. ISCA.

Fergus McInnes and David Attwater. 2004. Turn-taking and grounding in spoken telephone number transfers. *Speech Communication*, 43:205–223.

Sebastian Möller. 2002. A new taxonomy for the quality of telephone services based on spoken dialog systems. In *3rd SigDial Workshop on Discourse and Dialogue*, pages 142–153.

Bernhard Suhm. 2003. Towards best practices for speech user interface design. In *Eurospeech*.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000. Towards developing general models of usability with Paradise. *Natural Language Engineering*, 6:363–377.

Nigel Ward and Satoshi Nakagawa. 2004. Automatic user-adaptive speaking rate selection. *International Journal of Speech Technology*, 7:235–238.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel feedback in English and Japanese. *Journal of Pragmatics*, 32:1177–1207.

Nigel Ward and Wataru Tsukahara. 2003. A study in responsiveness in spoken dialog. *International Journal of Human-Computer Studies*, 59:603–630.

Nigel G. Ward, Anais G. Rivera, Karen Ward, and David G. Novick. 2005. Root causes of lost time and user stress in a simple dialog system. In *Interspeech*.

Nigel Ward. 2005, to appear. Non-lexical conversational sounds in American English. *Pragmatics and Cognition*.

Victor W. Zue and James R. Glass. 2000. Conversational interfaces: Advances and challenges. *Proceedings of the IEEE*, 88:1166–1180.