FROM NOVICE TO EXPERT: THE EFFECT OF TUTORIALS ON USER EXPERTISE WITH SPOKEN DIALOGUE SYSTEMS

Candace A. Kamm
Diane J. Litman
Marilyn A. Walker

AT&T Labs - Research
180 Park Ave., Bldg. 103
Florham Park, NJ 07932
{cak,diane,walker}@research.att.com

ABSTRACT

One of the challenges for the current state of the art in spoken dialogue systems is how to make the limitations of the system apparent to users. These limitations have many sources: limited vocabulary, limited grammar, or limitations in the application domain. This study explored the use of a 4-minute tutorial session to acquaint novice users with the features of a spoken dialogue system for accessing email. On a set of three scenario-based tasks, novice users who had the tutorial had task completion times and user satisfaction ratings that were comparable to those of expert users of the system. Novices who did *not* experience the tutorial had significantly longer task completion times on the initial task, but similar completion times to the tutorial group on the final task. User satisfaction ratings of the no-tutorial group were consistently lower than the ratings of the tutorial and the expert groups. Evaluation using the PARADISE [7] framework indicated that perceived task completion, mean recognition score, and number of help requests were significant predictors of user satisfaction with the system.

1. INTRODUCTION

Most currently deployed spoken language systems cannot handle unrestricted natural language input from the user. Despite efforts to support unrestricted input [1,2], typically users must know the system's limitations in vocabulary, grammar, and the application domain. Nevertheless, a number of recent studies suggest that users can *learn* the limitations of systems to accomplish tasks such as accessing voice mail, email or classified ads [4,5,6]. Previous work has shown that directive prompts is one way to make some limitations apparent [3] and that users learn from experience with the system [6,8]. In this paper, we examine the utility of a 4-minute tutorial conversation in helping novice users learn system limitations quickly.

The use of a tutorial dialogue is only appropriate for systems that the user will use repeatedly; for example, systems for accessing personal information such as voice mail, email, or a personal calendar. Thus our experiments involve testing users' performance with a spoken dialogue system for accessing email over the phone. We apply the PARADISE framework to evaluate the performance of three different user groups: (1) novice users who were not given a tutorial before doing the experimental tasks; (2) novice users who engaged in a tutorial interaction; and (3) expert users familiar with the limitations of the system who were also given a tutorial to remind them of the messaging commands.

Below, we show that novice users who were given a tutorial dialogue performed at almost expert level and their satisfaction with the system was much greater than the novice users who were not given a tutorial. Even though novice users who were not given a tutorial also learned the system limitations over three successive dialogues, their satisfaction with the system did not increase as their performance increased. We argue that this means that the use of tutorial dialogues may be critical to the success of spoken dialogue systems, since first impressions appear to have long term effects on users' perceptions of the system.

2. EXPERIMENTAL DESIGN

2.1 Email Retrieval via Spoken Dialogue

We used the same experimental setup that we had applied in other PARADISE evaluations [6,7]. The experiment required users to complete three tasks involving telephone access to email. The email retrieval system was a module in a voice-controlled personal communications assistant system called "Annie" [4]. This system used grammar-based speaker-independent automatic speech recognition for voice dialing and for voice-controlled message retrieval. The system supported barge-in; that is, the speech recognizer is active even when the system is playing a prompt, so the user can interrupt and take control of the interaction at any point by issuing a valid command. When the system recognizes a command, ongoing procedures (e.g., playing a long message) can be aborted in order to handle the user's newest request. The system also provided context-sensitive help messages, so that requests for help provided information about what the user could say at that point in the system hierarchy. The dialogue manager used a finite state machine to direct the interaction, based on the current state of the system and the recognition result.

To access the email retrieval module, a user had to call the system, login using a 10-digit account number (either by voice or by using the telephone keypad) and say "Play my messages" or one of its synonyms. Once inside the message retrieval module, available options included the commands "next message", "repeat the message", "delete it", "play message <*n*>" (where *n* is the message number), "cancel", "help", "I'm done here" (to exit the module) and "goodbye" (to end the session). A header indicating the sender and the subject preceded each message. The header information was identical to that provided in the ELVIS system [6], although the grammar and vocabulary were more limited. Each message was followed by a tone delimiting the end of the message body and a footer that provided the time that the message was received.

2.2 Subjects

Three groups of subjects were used. All of the subjects regularly used computers in the course of their everyday work and were familiar with email. The expert group (Expert) consisted of 12 researchers who had been using the communications assistant for voice-dialing and voice-mail retrieval for over 12 months. The novice groups each had 12 users who were a mix of administrative assistants and researchers. Novices were randomly assigned to either the group that participated in a 4-minute tutorial session about the system prior to performing the experimental tasks (Novice - Tutorial) or a group that did not experience the tutorial (Novice - No Tutorial).

2.3 Test Scenarios

2.3.1 Tutorial The tutorial session consisted of following a web-based script that presented a task to the user (e.g., Find out the telephone number in the message from Kim about 'Call me tomorrow') and stepped the user through the features of the message retrieval module, telling the user what to say at each step and indicating what the system response would be. The tutorial exposed the user to the set of commands for navigating the module, as well as demonstrating the messages played when a user asked for help or said nothing, allowing the system's time-outs to expire. All subjects receiving the tutorial (Experts and Novice-Tutorial groups) completed a web-based User Survey after finishing the tutorial. Immediately following the completion of the tutorial session, subjects went on to the experimental tasks.

2.3.2 Experimental Tasks All three user groups completed the same set of experimental tasks. Instructions were given on three web pages, one for each task. Each web page consisted of a brief general description of the email retrieval module, hints for using the module, a task description, and information on how to dial into the system. Subjects read the instructions in their offices before calling from their office phone. Each user performed three tasks in sequence, and each task consisted of two subtasks done during the same conversation with the system. The tasks were identical to those used in the ELVIS experiments [6]. An example task scenario is shown below. For this scenario, subjects needed to determine the correct values for the attributes Meeting Time and Meeting Place.

- You are working at home in the morning and plan to go directly to a meeting when you go into work. Kim said she would send you a message telling you where and when the meeting is. Find out the Meeting Time and the Meeting Place.

The general description and the hints on the web page for each task were identical. The subjects were given a different account number for each task and told that they needed to talk to Annie to find out some information that had been sent to them in an email message. Specific examples of what users could say to retrieve messages were not provided because: (1) we wanted the instructions to be identical for all subjects; (2) users could get information about what they could say from the context-sensitive help messages; (3) we wanted to quantify the frequency with which the different user groups accessed information on what they could say. Subjects were told they could say "Help" to ask for help about what to say, "Cancel" to stop an incorrect action by the system, and "I'm done here" to exit the current context. They were also told they could interrupt system prompts and that the system would offer suggestions about what the user could do.

We collected four types of data to extract a number of variables relevant for spoken dialogue system evaluation using the PARADISE framework [7]. First, all dialogues were recorded. The recording was used to calculate the total time of the interaction (the variable named **Elapsed Time**). Transcripts of the recordings were used to count the number of times users barged-in on system prompts (**Barge-Ins**). Second, the system logged its dialogue behavior upon entering and exiting each state in the state transition table for the dialogue. For each state, the system logged the number of timeout prompts (**Timeout Prompts),** the number of times the confidence level for ASR was too low and the system played a special rejection message, e.g. *Sorry, I didn't understand you* (**ASR Rejections)**, the times the user told the system to cancel an action (**Cancellations**) and the times the system played a help message to the user (**Help Messages Played**). The number of **User Turns** in each dialogue was also calculated from this data. Third, users filled out the web page forms after each task specifying whether they had completed the task (**Perceived Completion**) and providing the information they had acquired from the agent. The values obtained for each task attribute were used to compute the **kappa** statistic [7], which was used as an objective measure of task success. Kappa represents the agreement between the subject's responses and the correct responses for each task scenario, adjusted for chance agreement. The system's understanding (concept accuracy) was calculated from the logged ASR results in combination with the recordings, to determine a mean recognition score for each dialogue (**Mean Recognition Score**). Finally, users responded to a survey on their subjective evaluation of their performance and their satisfaction with the system's performance with the following questions:

- Did you complete the task?
- Was Annie easy to understand in this conversation?
- In this conversation, did Annie understand what you said?
- In this conversation, was it easy to find the message you wanted?
- Was the pace of interaction with Annie appropriate in this conversation?
- In this conversation, did you know what you could say at each point of the dialogue?
- How often was Annie sluggish and slow to reply to you in this conversation?
- Did Annie work the way you expected her to in this conversation?
- In this conversation, how did Annie's voice interface compare to the touch-tone interface to voice mail?
- From your current experience with using Annie to get your email, do you think you'd use Annie regularly to access your mail when you are away from your desk?

The user satisfaction survey was multiple choice, and the possible responses to most questions ranged over values such as (*almost never, rarely, sometimes, often, almost always*), or an equivalent range. Each of these responses was mapped to an integer between 1 and 5. Some questions had (*yes, no, maybe*) responses. Each question emphasized the user's experience with the system in the current conversation, with the hope that satisfaction measures would indicate perceptions specific to each conversation, rather than reflecting an overall evaluation of the system over the three tasks. A **Cumulative Satisfaction** score for each dialogue was calculated by summing the scores for each question.

The purpose of the experiment was to evaluate the effects of the tutorial session on performance measures (both task completion and measures reflecting the quality and efficiency of the interaction) and on user satisfaction. Thus, our primary independent variable was user expertise: whether the user was an expert, a novice who had the tutorial, or a novice who did not experience the tutorial. In addition, we were interested in examining how performance and user satisfaction changed as the novices became more familiar with the system.

3. RESULTS AND DISCUSSION

The PARADISE evaluation framework [7] posits that system performance can be modelled by determining the contributions of task success measures and a range of cost measures to user satisfaction. Table 1 shows mean results on each task for each user group (Expert, Novice - Tutorial, Novice - No Tutorial), for the task success measures of perceived completion and kappa, for the cost measures of elapsed time, mean recognition score, user turns, ASR rejections, ASR time outs, cancellations, barge-ins and help messages, and for cumulative satisfaction. A two-way ANOVA for the mixed design, with task as the between-groups factor and user expertise as the within-group factor, was performed for each measure. None of the ANOVA demonstrated a significant interaction between user expertise and task. Post-hoc comparisons of main effects described below use the Scheffe ratio to control overall error rate at $p<0.05$.

3.1 Task Success Measure

The ANOVA for both perceived completion and kappa demonstrated a significant main effect of user expertise. Post-hoc comparisons demonstrated that task completion rate and kappa for the Novice – No Tutorial group were significantly lower than the completion rates for the other two groups. The Novice – Tutorial group and the Expert group did not differ significantly on these task success measures. This result indicates that the tutorial was an effective way to increase task success for novice users.

3.2 Cost Measures

The ANOVA for the cost measures overwhelmingly demonstrated a significant main effect of user expertise. The Novice – No Tutorial group had significant higher elapsed time, user turns, help requests and cancellation requests than the Expert and Novice - Tutorial groups, and significantly more time-outs and ASR rejections than the Expert group. In addition, mean recognition score for the Novice – No Tutorial group was significantly lower than mean recognition score for the Expert group. The mean results for the Novice – Tutorial group were not significantly different from the Expert group on any cost measure. These results suggest that the use of the tutorial improved the quality and efficiency of interactions of novice users with the system.

The significant main effects for task demonstrated that Task 2 took significantly longer to complete (elapsed time) and required more user turns than Task 1 and Task 3. This result is attributable to the fact that Task 2 required an exhaustive exploration of the message set to complete the component tasks (e.g., "Find out if you need to call anyone, and if so, what the number is."), whereas Task 1 and Task 3 directed the users to find items in specific messages (e.g., "You have a message from Lee about a meeting. Find out the meeting place and time."). Once those items were encountered, the user could terminate the session, leaving any remaining messages unheard.

Table 1. Mean Results for Each Measure and Subject Group.

	Task	**Expert**	**Novice with Tutorial**	**Novice without Tutorial**
Perceived Completion (%)	**1**	100.0	95.8	70.8
	2	95.8	95.8	75.0
	3	95.8	83.3	79.2
Kappa	**1**	0.915	0.860	0.638
	2	0.958	0.896	0.604
	3	0.875	0.875	0.688
Elapsed Time (sec.)	**1**	123.2	173.7	305.8
	2	204.3	251.0	326.5
	3	142.1	162.5	199.6
Mean Recognition Score (%)	**1**	79.4	69.2	61.4
	2	77.2	72.7	69.0
	3	84.3	79.7	71.6
User Turns	**1**	11.0	15.7	33.8
	2	14.8	23.0	31.7
	3	11.75	16.1	17.6
ASR Rejections	**1**	2.5	5.1	13.5
	2	2.6	7.0	11.0
	3	1.8	4.8	5.5
ASR Time Outs	**1**	0.2	1.8	2.5
	2	0.8	0.6	2.6
	3	0.0	0.5	0.8
Cancels	**1**	0.17	0.0	1.0
	2	0.17	0.25	0.5
	3	0.17	0.25	0.33
Barge-ins	**1**	2.5	2.3	4.3
	2	5.5	5.7	6.5
	3	5.5	5.7	6.7
Help Messages Played	**1**	0.2	1.5	6.8
	2	0.1	1.6	3.7
	3	0.0	0.3	1.6
Cumulative Satisfaction	**1**	37.6	33.6	24.4
	2	35.2	34.5	25.1
	3	36.0	34.5	28.0

There were also significant main effects of task for the cost factors of mean recognition score, barge-ins, ASR rejections, and time-outs. The trend over task was an increase in mean recognition score and number of barge-ins and a decrease in number of rejections and time-outs. ASR rejections and mean recognition score may reflect, in part, how consistently users used valid grammar when speaking to the system. System time-out messages may reflect user confusion or uncertainty about which commands are valid during the interaction. The number of help messages played may also reflect user uncertainty about how to use the system. The improvement in these measures over the course of the experiment is consistent with subjects learning improved strategies for interacting with the system.

3.3 User Satisfaction

The ANOVA for the combined satisfaction score demonstrated a significant effect of user expertise. The mean combined satisfaction score for the Novice – No Tutorial group was significantly lower than that of the other two groups. There was no difference in mean combined satisfaction score between the Expert and Novice – Tutorial groups.

3.4 PARADISE Performance Function

The PARADISE framework [7] proposes modeling the performance of a dialogue system by estimating the relative contribution of a set of potential predictors to an externally valid criterion that reflects the "goodness" or utility of the system. In our experiment, we assume that user satisfaction is the external performance criterion. Multivariate linear regression was used to determine which of the task success and cost factors are most predictive of user satisfaction. First, all measures were normalized so that the magnitude of the regression coefficients would reflect the relative contribution of that factor to the satisfaction measure. An initial stepwise regression over all the factors suggested that perceived task completion, mean recognition score, number of help requests and number of ASR rejections were the only significant predictors. A subsequent regression on those four factors demonstrated that only perceived completion, mean recognition score and help requests were significant. A final regression on these three factors accounted for 41.3% of the variance in the data, and yielded the following equation:

$$\text{PERFORMANCE} = .25\ \text{MRS} + .33\ \text{COMP} - .33\ \text{HELP}$$

where MRS is mean recognition score, COMP is perceived completion, and HELP is number of help messages. The finding that recognition score and perceived completion are significant factors is consistent with previous results for both an email reading task and a train timetable task [6,7]. The importance of help requests most likely reflects the fact that the current study sampled subjects with different levels of expertise. Predicted satisfaction scores were computed for each subject. ANOVA for the predicted scores demonstrated significant main effects of user expertise and task that mirrored the results of the analyses described above; that is, system performance was significantly poorer for the Novice – No Tutorial group than for the other two groups, and system performance tended to increase over the three experimental tasks.

4. SUMMARY

This paper examined the effect of a 4-minute tutorial session with a voice-enabled email retrieval system on users' performance and satisfaction with the spoken dialogue system. The results support our hypothesis that novice users who experienced the tutorial would outperform novice users who did not receive the tutorial, even when both groups had access to the same system help messages and hints on how to use the system. The results also showed that user satisfaction is higher for subjects who have the benefit of the tutorial session. In addition, even though performance of the Novice – No Tutorial group improved as they gained experience with the system, their user satisfaction scores did not increase. This result indicates that initial interactions with a spoken dialogue system may have a persistent influence on subjective reactions to the system. The results of this study suggest that a short duration tutorial can serve as a simple procedure for ensuring a successful initial experience with a spoken dialogue system that may have real impact on customer retention for spoken dialogue services.

ACKNOWLEDGEMENT

The authors would like to thank Dawn Dutton and Russ Ritenour for their work in the design and implementation of the spoken dialogue messaging retrieval system used in this study.

REFERENCES

1. Boyce, S. and Gorin, A. "User interface issues for natural spoken dialog systems." *Proc. ISSD 96*, 65-68, 1996.

2. Gorin, A., Parker, B., Sachs, R., and Wilpon, J. "How may I help you?" *Proc. IVTTA-96*, 61-64, 1996.

3. Kamm, C. "User interfaces for voice applications". In D. Roe & J. Wilpon (Eds.) *Voice Communication between Humans and Machines* (pp. 422-442). Washington, DC: National Academy Press, 1994.

4. Kamm, C., Narayanan, S., Dutton, D., and Ritenour, R "Evaluating spoken dialog systems for telecommunications services". *Proc. Eurospeech 97* , 2203-2207, 1997.

5. Meng, H. , Busayapongchi, S., Glass, J., Goddeau, D., Hetherington, L. Hurley, E. Pao, C., Polifroni, J., Seneff, S. and Zue, V. "WHEELS: A conversational system in the automobile classifieds domain." Proceedings ISSD 96, 165-168, 1996.

6. Walker, M., Fromer, J., Di Fabbrizio, G., Mestel, C. and Hindle, D. "What can I say: Evaluating a spoken language interface to email." *Proceedings of the Conference on Human Factors in Computing Systems, CHI98,* 1998.

7. Walker, M., Litman, D., Kamm, C., and Abella, A. "Evaluating spoken dialogue agents with PARADISE: two case studies". *Computer Speech and Language,* in press.

8. Yankelovich, N., Levow, G. and Marx, M. "Designing speech acts: Issues in speech user interfaces." *Proceedings of the Conference on Human Factors in Computing Systems, CHI95*, 1995.