Empirically Evaluating an Adaptable Spoken Dialogue System

Diane J. Litman[1] and Shimei Pan[2]⋆

[1] AT&T Labs - Research, Florham Park, NJ, USA
[2] Computer Science Department, Columbia University, New York, NY, USA

Abstract. Recent technological advances have made it possible to build real-time, interactive spoken dialogue systems for a wide variety of applications. However, when users do not respect the limitations of such systems, performance typically degrades. Although users differ with respect to their knowledge of system limitations, and although different dialogue strategies make system limitations more apparent to users, most current systems do not try to improve performance by adapting dialogue behavior to individual users. This paper presents an empirical evaluation of TOOT, an adaptable spoken dialogue system for retrieving train schedules on the web. We conduct an experiment in which 20 users carry out 4 tasks with both adaptable and non-adaptable versions of TOOT, resulting in a corpus of 80 dialogues. The values for a wide range of evaluation measures are then extracted from this corpus. Our results show that adaptable TOOT generally outperforms non-adaptable TOOT, and that the utility of adaptation depends on TOOT's initial dialogue strategies.

1 Introduction

Recent technological advances have made it possible to build real-time, interactive spoken dialogue systems for a wide variety of applications. However, due to limitations in such areas as vocabulary, grammar, and application functionality, most deployed systems can only accept a small set of natural language inputs from users. When users do not respect the limitations of spoken dialogue systems, system performance typically degrades. For example, user utterances are more likely to be rejected or misunderstood, dialogue length typically increases, and users are less likely to achieve their task goals. Although individual users differ with respect to their knowledge of system limitations (Kamm et al., 1998), and although different dialogue strategies make system limitations more apparent to users (Walker et al., 1997a; Danieli and Gerbino, 1995), current spoken dialogue systems typically do not try to adapt their dialogue strategies to individual users. A major obstacle to having a spoken dialogue system perform such adaptation dynamically is that there is currently little understanding of the type of properties that would need to be represented in a user model for this task, let alone automatically acquired during the course of a dialogue. Nevertheless, users do seem to intuitively know when and how their interactions with a spoken dialogue system could be changed, in order to improve overall system performance. This observation has led us to explore the utility of an *adaptable* spoken dialogue system, where the *user* rather than the system controls the adaptation of the system's dialogue strategies. In particular, we hypothesize that if a user can effectively adapt the system's behavior, so that the system will use the dialogue strategies that best match the user's needs at any point in a dialogue, then the overall performance of the system should improve.

⋆ We thank J. Chu-Carroll, C. Kamm, D. Lewis, M. Walker, and S. Whittaker for helpful comments.

Figure 1 shows an example dialogue with an adaptable train-timetable spoken dialogue system called TOOT.[1] In this excerpt, TOOT initially uses a set of dialogue strategies often found

U1: I'd like to go to Chicago from Baltimore on Saturday at 8 am.
T1: Do you want to leave in the morning or in the evening?
U2: Uh morning.
T2: Sorry, I can't understand you. Please repeat your answer.
U3: In the morning.
T3: Do you want me to find the trains from Baltimore to Chicago on Saturday around 10 morning now?
U4: No.
T4: How may I help you?
U5: I would like to go from Baltimore to Chicago on Saturday at 8 o'clock.
T5: Do you want me to find the trains from Baltimore to DC on Tuesday around 11:45 am now?
U6: No.
T6: How may I help you?
U7: Change strategy.
T7: You are using the user initiative. Which initiative do you want to change to? System –
U8: System initiative please.
T8: You are using the no confirmation strategy. Which confirmation strategy do you want to change to? No confirmation, implicit confirmation, or explicit confirmation?
U9: Implicit please.
T9: I heard you say go to DC on Tuesday around 11:45 am. Do you want me to find the trains from Baltimore to DC on Tuesday around 11:45 am now?
U10: No.
T10: Which city do you want to go to?
U11: Chicago.
T11: I heard you say go to Chicago. Which city do you want to leave from?
U12: Baltimore.
T12: I heard you say leave from Baltimore. On which day of the week do you want to leave?
U13: Saturday.
T13: I heard you say leave on Saturday. What time do you want to leave?
U14: 8:00 in the morning.
T14: Sorry, I can't understand you. Please repeat your answer.
U15: 8:00 in the morning.
T15: I heard you say leave around 8:00 in the morning. Do you want me to find the trains from Baltimore to Chicago on Saturday around 8:00 in the morning now?
U16: Yes.

Figure 1. Example dialogue with an adaptable spoken dialogue system.

in human-human interactions. However, the strategies do not work particularly well for this user. Utterances U1 and U5 are misunderstood by TOOT, and U2 is not understood at all. Recognizing a mismatch, the user changes the strategies used by TOOT (U7 - U9). As a result, the open ended prompt that TOOT used earlier in the dialogue (T4) is now replaced with a series of specific questions (T10, T11, T12, and T13). TOOT also now highlights any potential misrecog-

[1] This excerpt is taken from the experimental corpus described below.

nitions by prefacing its utterances with “I heard you say …”. Earlier in the dialogue, it took 5 utterances before the user was aware that TOOT had misrecognized the “8” in U1 as “10”. As a result of these changes, the dialogue proceeds more smoothly after the adaptation (e.g., TOOT's misrecognition rate is reduced), and a correct database query is soon generated (T15-U16).

In this paper, we present an evaluation of adaptability in TOOT. We conduct an experiment in which 20 novice users carry out 4 tasks with one of two versions of TOOT (*adaptable* and *non-adaptable*), resulting in a corpus of 80 dialogues. The values for a range of evaluation measures are then extracted from this corpus. Hypothesis testing shows that a variety of differences depend on the user's ability to adapt the system. A PARADISE assessment of the contribution of each evaluation measure to overall performance (Walker et al., 1997b) shows that the phenomena influenced by adaptation are also the major phenomena that significantly influence performance. Our results show that adaptable TOOT generally outperforms non-adaptable TOOT, and that the utility of adaptation depends on the initial configuration of dialogue strategies.

2 TOOT

TOOT is a voice-enabled dialogue system for accessing train schedules from the web via a telephone conversation. TOOT is implemented using a spoken dialogue system platform (Kamm et al., 1997) that combines automatic speech recognition (ASR), text-to-speech (TTS), a phone interface, and modules for specifying a dialogue manager and application functions. ASR in our platform is speaker-independent, grammar-based and supports *barge-in* (which allows users to interrupt TOOT when it is speaking, as in utterances T7 and U8 in Figure 1). The dialogue manager uses a finite state machine to control the interaction, based on the current system state and ASR results. TOOT's application functions access train schedules available at www.amtrak.com. Given a set of constraints, the functions return a table listing all matching trains in a specified temporal interval, or within an hour of a specified timepoint. This table is converted to a natural language response which can be realized by TTS through the use of templates.[2]

Depending on the user's needs during the dialogue, TOOT can use one of three dialogue strategies for managing initiative (“system”, “mixed” or “user”), and one of three strategies for managing confirmation (“explicit,” “implicit,” or “no”). TOOT's initiative strategy specifies who has control of the dialogue, while TOOT's confirmation strategy specifies how and whether TOOT lets the user know what it just understood. In Figure 1, TOOT initially used user initiative and no confirmation, then later used system initiative and implicit confirmation. The following fragments provide additional illustrations of how dialogues vary with strategy:

System Initiative, Explicit Confirmation	*User Initiative, No Confirmation*
T: Which city do you want to go to?	T: How may I help you?
U: Chicago.	U: I want to go to Chicago from Baltimore.
T: Do you want to go to Chicago?	T: On which day of the week do you want to leave?
U: Yes.	U: I want a train at 8:00.

[2] The current version of TOOT uses a literal response strategy (Litman et al., 1998). Informally, if the returned table contains 1-3 trains, TOOT lists the trains; if the table contains greater than 4 trains, TOOT lists the trains 3 at a time; if the table is empty, TOOT reports that no trains satisfy the constraints. TOOT then asks the user if she wants to continue and find a new set of trains.

Although system initiative with explicit confirmation is the most cumbersome approach, it can help improve some aspects of performance for users who do not have a good understanding of the system's limitations. The use of system initiative helps reduce ASR misrecognitions and rejections (Walker et al., 1997a), by helping to keep the user's utterances within the system's vocabulary and grammar. The use of explicit confirmation helps increase the user's task success (Danieli and Gerbino, 1995), by making the user more aware of any ASR misrecognitions and making it easier for users to correct misrecognitions when they occur. On the other hand, system initiative and explicit confirmation typically increase total dialogue length (Walker et al., 1997a; Danieli and Gerbino, 1995). For users whose utterances are generally understood, other strategies might be more effective. Consider the use of user initiative with no confirmation, the most human-like approach. In user (as well as in mixed) initiative mode, TOOT can still ask the user specific questions, but can also ask open-ended questions such as "How may I help you?". Furthermore, in user (but not in mixed) initiative mode, TOOT even lets the user ignore TOOT's questions (as in the last user utterance in the example above). By allowing users to specify multiple attributes in a single utterance, and by not informing users of every potential misrecognition, this approach can lead to very short dialogues when ASR performance is not a problem.

In an earlier implementation of TOOT (as well as in other spoken dialogue systems that we have studied (Walker et al., 1997a; Kamm et al., 1998)), a set of initial dialogue strategies was assigned to the system as a default for each user, and could not be changed if inappropriate.[3] As discussed above, however, we hypothesize that we can improve TOOT's performance by dynamically adapting the choice of dialogue strategies, based on the circumstances at hand. Although one of our long-term goals is to have TOOT automatically control the adaptation process, this would require that we first solve several open research topics. For example, TOOT would need to be able to detect, in real time, dialogue situations suggesting system adaptation. As a result, our initial research has instead focused on giving users the ability to dynamically adapt TOOT's dialogue behaviors. For example, if a user's utterances are not being understood, the user could try to reduce the number of ASR rejections and misrecognitions by changing the strategies so that TOOT would take more initiative. Conversely, if a user's utterances are being correctly understood, the user could try to decrease the dialogue length by having TOOT perform less confirmations. To allow us to test whether such an adaptable system does indeed increase performance, we have created both "adaptable" and "non-adaptable" versions of TOOT. In adaptable TOOT, users are allowed to say "change strategy" at any point(s) in the dialogue. TOOT then asks the user to specify new initiative and confirmation strategies, as in utterances U7-U9 in Figure 1. In non-adaptable TOOT, the default dialogue strategies can not be changed.

3 Experimental Design

Our experiment was designed to test if adaptable TOOT performed better than non-adaptable TOOT, and whether any differences depended on TOOT's initial dialogue strategies and/or the user's task. Our design thus consisted of three factors: *adaptability*, *initial dialogue strategy*, and *task scenario*. Subjects were 20 AT&T technical summer employees not involved with the design or implementation of TOOT, who were also novice users of spoken dialogue systems in

[3] In particular, the previous version of TOOT always used system initiative with implicit confirmation (Litman et al., 1998).

general. 10 users were randomly assigned to *adaptable* TOOT and 10 to *non-adaptable* TOOT. For each of these groups, 5 users were randomly assigned to a version of TOOT with the initial dialogue strategies set to *system initiative* and *explicit confirmation* (SystemExplicit TOOT); the remaining 5 users were assigned to a version of TOOT with the initial dialogue strategies set to *user initiative* and *no confirmation* (UserNo TOOT). Each user performed the same 4 tasks in sequence. Our experiment yielded a corpus of 80 dialogues (2633 turns; 5.4 hours of speech).

Users used the web to read a set of experimental instructions in their office, then called TOOT from their phone. The experimental instructions consisted of a description of TOOT's functionality, hints for talking to TOOT, and links to 4 task scenarios. An example task scenario is as follows: "Try to find a train going *to* **Chicago** *from* **Baltimore** on **Saturday** at **8 o'clock am**. If you cannot find an exact match, find the one with the **closest** departure time. Please write down the **exact departure time** of the train you found as well as the **total travel time**." The instructions for adaptable TOOT also contained a brief tutorial explaining how to use "change strategy", and guidelines for doing so (e.g., " if you don't know what to do or say, try system initiative").

We collected three types of data to compute a number of measures relevant for spoken dialogue evaluation (Walker et al., 1997a). First, all dialogues were recorded. The recordings were used to calculate the total time of each dialogue (the evaluation measure **Elapsed Time**), and to (manually) count how many times per dialogue each user interrupted TOOT (**Barge Ins**).

Second, the dialogue manager's behavior on entering and exiting each state in the finite state machine was logged. This log was used to calculate the total number of **System Turns** and **User Turns**, **Timeouts** (when the user doesn't say anything within a specified time frame, TOOT provides suggestions about what to say), **Helps** (when the user says "help", TOOT provides a context-sensitive help message), **Cancels** (when the user says "cancel", TOOT undoes its previous action), and **ASR Rejections** (when the confidence level of ASR is too low, TOOT asks the user to repeat the utterance). In addition, by listening to the recordings and comparing them to the logged ASR results, we calculated the concept accuracy (intuitively, semantic interpretation accuracy) for each utterance. This was then used, in combination with ASR rejections, to compute a **Mean Recognition** score per dialogue.

Third, users filled out a web survey after each dialogue. Users specified the departure and travel times that they obtained via the dialogue. Given that there was a single correct train to be retrieved for each task scenario, this allowed us to determine whether users successfully achieved their task goal or not (**Task Success**). Users also responded to the following questionnaire:

– Was the system easy to understand? (**TTS Performance**)
– Did the system understand what you said? (**ASR Performance**)
– Was it easy to find the schedule you wanted? (**Task Ease**)
– Was the pace of interaction with the system appropriate? (**Interaction Pace**)
– Did you know what you could say at each point of the dialogue? (**User Expertise**)
– How often was the system sluggish and slow to reply to you? (**System Response**)
– Did the system work the way you expected it to? (**Expected Behavior**)
– From your current experience with using our system, do you think you'd use this regularly to access train schedules when you are away from your desk? (**Future Use**)

Each question measured a particular usability factor, e.g., **TTS Performance**. Responses ranged over *n* pre-defined values (e.g., *almost never, rarely, sometimes, often, almost always*), and were mapped to an integer in $1 \ldots 5$ (with 5 representing optimal performance). **User Satisfaction** was computed by summing each question's score, and thus ranged in value from 8 to 40.

4 Results

We use analysis of variance (ANOVA) (Cohen, 1995) to determine whether the adaptability of TOOT produces significant differences in any of the evaluation measures for our experiment. We also use the PARADISE evaluation framework (Walker et al., 1997b) to understand which of our evaluation measures best predicts overall performance in TOOT. Following PARADISE, we organize our evaluation measures along the following four performance dimensions:

- *task success*: Task Success
- *dialogue quality*: Helps, ASR Rejections, Timeouts, Mean Recognition, Barge Ins, Cancels
- *dialogue efficiency*: System Turns, User Turns, Elapsed Time
- *system usability*: User Satisfaction (based on TTS Performance, ASR Performance, Task Ease, Interaction Pace, User Expertise, System Response, Expected Behavior, Future Use)

4.1 Adaptability Effects

Recall that our mixed[4] experimental design consisted of three factors: *adaptability, initial dialogue strategy*, and *task scenario*. Each of our evaluation measures is analyzed using a three-way ANOVA for these three factors. The ANOVAs demonstrate a *main effect of adaptability* for the task success and system usability dimensions of performance. These main adaptability effects are independent of TOOT's initial dialogue strategy as well as of the task scenario being executed by the user. The ANOVAs also demonstrate *interaction effects of adaptability and initial dialogue strategy* for the dialogue quality and system usability performance dimensions. In contrast to the main effects, these adaptability effects are not independent of TOOT's initial dialogue strategy (i.e., the effects of adaptability and initial strategy are not additive).[5]

Table 1 summarizes the means for each evaluation measure that shows a main effect of adaptability, and that cannot be further explained by any interaction effects. The first row in the table indicates that Task Success is significantly higher for adaptable TOOT than for non-adaptable TOOT. Users successfully achieve the goals specified in the task scenario in 80% of the dialogues with adaptable TOOT, but in only 55% of the dialogues with non-adaptable TOOT. The probability $p<.03$ indicates that the difference is statistically significance (the standard upper bound for calling a result statistically significant is $p<.05$ (Cohen, 1995)). The second row indicates that with respect to User Satisfaction, users also rate adaptable TOOT more highly than non-adaptable TOOT. Recall that User Satisfaction takes all of the factors in the usability questionnaire into account. As will be discussed below, PARADISE correlates overall system performance with this measure. In sum, our ANOVAs indicate that making TOOT adaptable increases users' rates of task success as well as users' perceptions of overall system usability.

Table 2 summarizes the means for each evaluation measure that shows an interaction effect of adaptability and initial dialogue strategy. A similar pattern of interaction emerges in the first and

[4] *Task scenario* is between-groups and *initial dialogue strategy* and *adaptability* are within-group.
[5] Effects of *initial dialogue strategy* and *task scenario* are beyond the scope of this paper.

Table 1. Main effects of adaptability.

Measure	Non-Adaptable (n=40)	Adaptable (n=40)
Task Success (%) ($p<.03$)	55.00	80.00
User Satisfaction ($p<.03$)	26.68	31.60

Table 2. Interaction effects of adaptability and initial dialogue strategy.

	Non-Adaptable (n=40)		Adaptable (n=40)	
Measure	SystemExplicit	UserNo	SystemExplicit	UserNo
Mean Recognition (%) ($p<.01$)	88.44	57.94	82.55	75.85
User Expertise ($p<.05$)	4.69	3.01	4.45	3.85
Future Use ($p<.02$)	3.50	1.70	3.60	3.80

second rows of the table. When users are given the capability to adapt TOOT, Mean Recognition decreases for SystemExplicit TOOT (88.44% versus 82.55%) but increases for UserNo TOOT (57.94% versus 75.85%). Perceptions of User Expertise also decrease for adaptable SystemExplicit TOOT (4.69 versus 4.45) but increase for adaptable UserNo TOOT (3.01 versus 3.85). In contrast, Future Use is higher for adaptable TOOT than for non-adaptable TOOT, for both initial strategies. Thus, users of adaptable TOOT are more likely than users of non-adaptable TOOT to think that they would use TOOT on a regular basis. However, the increase in Future Use is smaller for SystemExplicit TOOT (3.5 to 3.6) than for UserNo TOOT (1.7 to 3.8). In sum, Table 2 indicates that differences reflecting both dialogue quality and system usability are an effect of the interaction of the adaptability of TOOT and TOOT's initial dialogue strategy. For the UserNo version of TOOT, making TOOT adaptable increases the means for all of the measures shown in Table 2. For the SystemExplicit version of TOOT, despite the Mean Recognition and User Expertise results in Table 2, users are nevertheless at least if not more likely to use adaptable System Explicit TOOT in the future. We speculate that users are willing to tolerate minor levels of particular types of performance degradations in SystemExplicit TOOT, in order to obtain the sense of control provided by adaptability. We also speculate that the utility of adaptable SystemExplicit TOOT would increase for expert users. In conjunction with Table 1, our results with novice users suggest that adaptability is an extremely useful capability to add to UserNo TOOT, and a capability that is still worth adding to SystemExplicit TOOT.

It is interesting to also examine the way in which adaptation is performed for each initial dialogue strategy. Of the 20 dialogues with adaptable SystemExplicit TOOT, 5 dialogues contained 1 adaptation and a 6th dialogue contained 2 adaptations. 3 of the 5 users adapted at least 1 dialogue, and overall, confirmation was changed more times than initiative. Of the 20 dialogues with adaptable UserNo TOOT, 10 dialogues contained 1 adaptation and an 11th dialogue contained 2 adaptations. All 5 users of UserNo TOOT adapted at least 1 dialogue. Users of UserNo TOOT changed initiative more than they changed confirmation, and also changed initiative more drastically. In conjunction with our ANOVA results, these observations lead us to speculate that adapting a poorly performing system is both more feasible and more important for novice users than adapting a reasonably performing system.

4.2 Contributors to Performance

To quantify the relative importance of our multiple evaluation measures to performance, we use the PARADISE evaluation framework to derive a performance function from our data. The PARADISE model posits that performance can be correlated with a meaningful external criterion of usability such as User Satisfaction. PARADISE then uses stepwise multiple linear regression to model User Satisfaction from measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency:

$$\text{User Satisfaction} = \sum_{i=1}^{n} w_i * \mathcal{N}(measure_i)$$

Linear regression produces coefficients (i.e., weights w_i) describing the relative contribution of predictor factors in accounting for the variance in a predicted factor. In PARADISE, the task success and dialogue cost measures are predictors, while User Satisfaction is predicted. The normalization function $\mathcal{N}$ guarantees that the coefficients directly indicate the relative contributions.

The application of PARADISE to the TOOT data shows that the most significant contributors to User Satisfaction are Mean Recognition, Task Success, and Elapsed Time, respectively. In addition, PARADISE shows that the following performance function provides the best fit to our data, accounting for 55% of the variance in User Satisfaction:[6]

$$\text{User Satisfaction} = .45\mathcal{N}(\text{Mean Recognition}) + .33\mathcal{N}(\text{Task Success}) - .14\mathcal{N}(\text{Elapsed Time})$$

Our performance function demonstrates that TOOT performance (estimated using subjective usability ratings) can be best predicted using a weighted combination of objective measures of dialogue quality, task success, and dialogue efficiency. In particular, more accurate speech recognition, more success in achieving task goals, and shorter dialogues all contribute to increasing perceived performance in TOOT.

Our performance equation helps explain the main effect of adaptability for User Satisfaction that was shown in Table 1. Recall that our ANOVAs for both Mean Recognition and Task Success showed adaptability effects (Tables 2 and 1, respectively). Our PARADISE analysis showed that these measures were also the most important measures in explaining the variance in User Satisfaction. It is thus not surprising that User Satisfaction shows an effect of adaptability, with users rating the performance of adaptable TOOT more highly than non-adaptable TOOT.

A result that was not apparent from the analysis of variance is that Elapsed Time is a performance predictor. However, the weighting of the measures in our performance function suggests that Mean Recognition and Task Success are more important measures of overall performance than Elapsed Time. These findings are consistent with our previous PARADISE evaluations, where measures of task success and dialogue quality were also the most important performance predictors (Litman et al., 1998; Walker et al., 1998; Kamm et al., 1998). Our findings draw into question a frequently made assumption in the field regarding the centrality of efficiency to performance, and like other recent work, demonstrates that there are important tradeoffs between efficiency and other performance dimensions (Danieli and Gerbino, 1995; Walker et al., 1997a).

[6] Linear regression assumes that predictors are not highly correlated (e.g., because correlations above .70 can affect the coefficients, deletion of redundant predictors is advised (Monge and Cappella, 1980)). There is only 1 positive correlation among our predictors (between Mean Recognition and Task Success), and it is well below .70.

5 Related Work

In the area of spoken dialogue, van Zanten (1998) has proposed a method for adapting initiative in form-filling dialogues. Whenever the system rejects a user's utterance, the system takes more initiative; whenever the user gives an over-informative answer, the system yields some initiative. While this method has the potential of being automated, the method has been neither fully implemented nor empirically evaluated. Smith (1998) has evaluated strategies for dynamically deciding whether to confirm each user utterance during a task-oriented dialogue. Simulation results suggest that context-dependent adaptation strategies can improve performance, especially when the system has greater initiative. Walker et al. (1998) and Levin and Pieraccini (1997) have used reinforcement learning to adapt dialogue behavior over time such that system performance improves. We have instead focused on optimizing performance during a single dialogue.

The empirical evaluation of an adaptive interface in a commercial software system (Strachan et al., 1997) is also similar to our work. Analysis of variance demonstrated that an adaptive interface based on minimal user modeling improved subjective user satisfaction ratings.

6 Conclusion

We have presented an empirical evaluation of adaptability in TOOT, a spoken dialogue system that retrieves train schedules from the web. Our results suggest that adaptable TOOT generally outperforms non-adaptable TOOT for novice users, and that the utility of adaptation is greater for UserNo TOOT than for SystemExplicit TOOT. By using analysis of variance to examine how a set of evaluation measures differ as a function of adaptability, we elaborate the conditions under which adaptability leads to greater performance. When users interact with adaptable rather than non-adaptable TOOT, User Satisfaction and Task Success are significantly higher. These results are independent of TOOT's initial dialogue strategy and task scenario. In contrast, Mean Recognition, User Expertise, and Future Use illustrate an interaction between initial dialogue strategy and adaptability. For SystemExplicit TOOT, the adaptable version does not outperform the non-adaptable version, or does not outperform the non-adaptable version very strongly. For UserNo TOOT, the adaptable version outperforms the non-adaptable version on all three measures.

By using PARADISE to derive a performance function from data, we show that Mean Recognition, Task Success, and Elapsed Time best predict a user's overall satisfaction with TOOT. These results help explain why adaptability in TOOT leads to overall greater performance, and allow us to make predictions about future performance. For example, we predict that a SystemImplicit strategy is likely to outperform our SystemExplicit strategy, since we expect that Mean Recognition and Task Success will remain constant but that Elapsed Time will decrease.

Currently, we are extending our results along two dimensions. First, we have made a first step towards automating the adaptation process in TOOT, by using machine learning to develop a classifier for detecting dialogues with poor speech recognition (Litman et al., 1999). (Recall that our PARADISE evaluation suggested that recognition accuracy was our best performance predictor.) We hope to use this classifier to determine the features that need to be represented in a user model, and to tell us when the user model indicates the need for adaptation. Guided by our empirical results, we can then develop an initial adaptation algorithm that takes dialogue strategy into account. For example, based on our experiment, we would like UserNo TOOT to adapt itself fairly aggressively when it recognizes that the user is having a problem. Second, the experiments reported here considered only our two most extreme initial dialogue strategy configurations. To generalize our results, we are currently experimenting with other dialogue strategies. To date we have collected 40 dialogues using a mixed initiative, implicit confirmation version of TOOT, with initial promising results. For example, user satisfaction continues to exhibit the same main effect of adaptability when our corpus is augmented with these new dialogues.

References

Cohen, P. (1995). *Empirical Methods for Artificial Intelligence*. MIT Press, Boston.

Danieli, M., and Gerbino, E. (1995). Metrics for evaluating dialogue strategies in a spoken language system. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 34–39.

Kamm, C., Narayanan, S., Dutton, D., and Ritenour, R. (1997). Evaluating spoken dialog systems for telecommunication services. In *Proc. 5th European Conf. on Speech Communication and Technology*.

Kamm, C., Litman, D., and Walker, M. (1998). From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. In *Proc. 5th International Conf. on Spoken Language Processing*, 1211–1214.

Levin, E., and Pieraccini, R. (1997). A stochastic model of computer-human interaction for learning dialogue strategies. In *Proc. 5th European Conf. on Speech Communication and Technology*.

Litman, D., Pan, S., and Walker, M. (1998). Evaluating Response Strategies in a Web-Based Spoken Dialogue Agent. In *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conf. on Computational Linguistics*, 780–786.

Litman, D., Walker, M., and Kearns, M. (1999). Automatic detection of poor speech recognition at the dialogue level. Manuscript submitted for publication.

Monge, P., and Cappella, J., eds. (1980). *Multivariate Techniques in Human Communication Research*. Academic Press, New York.

Smith, R. W. (1998). An evaluation of strategies for selectively verifying utterance meanings in spoken natural language dialog. *International Journal of Human-Computer Studies* 48:627–647.

Strachan, L., Anderson, J., Sneesby, M., and Evans, M. (1997). Pragmatic user modelling in a commercial software system. In *Proc. 6th International Conf. on User Modeling*, 189–200.

van Zanten, G. V. (1998). Adaptive mixed-initiative dialogue management. Technical Report 52, IPO, Center for Research on User-System Interaction.

Walker, M., Hindle, D., Fromer, J., Fabbrizio, G. D., and Mestel, C. (1997a). Evaluating competing agent strategies for a voice email agent. In *Proc. 5th European Conf. on Speech Communication and Technology*.

Walker, M., Litman, D., Kamm, C., and Abella, A. (1997b). PARADISE: A general framework for evaluating spoken dialogue agents. In *Proc. 35th Annual Meeting of the Association for Computational Linguistics and 8th Conf. of the European Chapter of the Association for Computational Linguistics*, 271–280.

Walker, M., Fromer, J., and Narayanan, S. (1998). Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conf. on Computational Linguistics*, 1345–1352.