

Evaluating Automatic Dialogue Strategy Adaptation for a Spoken Dialogue System

Jennifer Chu-Carroll

Lucent Technologies Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974, U.S.A.
jenc@research.bell-labs.com

Jill Suzanne Nickerson

Harvard University
Cambridge, MA 02138, U.S.A.
nickerso@eecs.harvard.edu

Abstract

In this paper, we describe an empirical evaluation of an adaptive mixed initiative spoken dialogue system. We conducted two sets of experiments to evaluate the mixed initiative and automatic adaptation aspects of the system, and analyzed the resulting dialogues along three dimensions: performance factors, discourse features, and initiative distribution. Our results show that 1) both the mixed initiative and automatic adaptation aspects led to better system performance in terms of user satisfaction and dialogue efficiency, and 2) the system's adaptation behavior better matched user expectations, more efficiently resolved dialogue anomalies, and resulted in higher overall dialogue quality.

1 Introduction

Recent advances in speech technologies have enabled spoken dialogue systems to employ mixed initiative dialogue strategies (e.g. (Allen et al., 1996; Sadek et al., 1996; Meng et al., 1996)). Although these systems interact with users in a manner more similar to human-human interactions than earlier systems employing system initiative strategies, their response strategies are typically selected using only local dialogue context, disregarding dialogue history. Therefore, their gain in naturalness and performance under optimal conditions is often overshadowed by their inability to cope with anomalies in dialogues by automatically adapting dialogue strategies. In contrast, Figure 1 shows a dialogue in which the system automatically adapts dialogue strategies based on the current user utterance and dialogue history.¹ After failing to obtain a valid response to an information-seeking query in utterance (4), the system adapted dialogue strategies to provide additional information in (6) that assisted the user in responding to the query. Furthermore, after the user responded to a limited system prompt in (10) with a fully-specified query in (11), implicitly indicating her intention to take charge of the problem-

¹S and U indicate system and user utterances, respectively. The words appearing in square brackets are the output from the Lucent Automatic Speech Recognizer (Reichl and Chou, 1998; Ortmanns et al., 1999), configured to use class-based probabilistic n-gram language models. The task and dialogue initiative annotations are explained in Section 2.1.

solving process, the system again adapted strategies, hence providing an open-ended prompt in (13).

Previous work has shown that dialogue systems in which users can explicitly change the system's dialogue strategies result in better performance than non-adaptable systems (Litman and Pan, 1999). However, no earlier system allowed for initiative-oriented automatic strategy adaptation based on information dynamically extracted from the user's spoken input. In this paper, we briefly introduce MIMIC, a *mixed initiative* spoken dialogue system that *automatically adapts* dialogue strategies. We then describe two experiments that evaluated the effectiveness of MIMIC's mixed initiative and automatic adaptation capabilities. Our results show that, when analyzed along the *performance* dimension, MIMIC's mixed initiative and automatic adaptation features lead to more efficient dialogues and higher user satisfaction. Moreover, when analyzed along the *discourse* and *initiative* dimensions, MIMIC's adaptation capabilities result in dialogues in which system behavior better matches user expectations and dialogue anomalies are resolved more efficiently.

2 MIMIC: An Adaptive Mixed Initiative Dialogue System

MIMIC is a voice-enabled, telephone-based dialogue system that provides movie showtime information. Two main features distinguish MIMIC from other existing spoken dialogue systems. First, MIMIC automatically adapts dialogue strategies based on dialogue characteristics and thus tailors its behavior to the dialogue context. Second, MIMIC's initiative module is decoupled from the dialogue manager's goal and strategy selection processes, thereby enabling the system's adaptation behavior to be easily modified. We outline characteristics of MIMIC relevant to our experiments and analyses. For further details about the system, see (Chu-Carroll, 2000).

2.1 Adaptation of Dialogue Strategies

Figure 1 showed an actual interaction between MIMIC and a user. The *task* and *dialogue* initiative holders annotated for each system turn are the factors that affect MIMIC's dialogue strategy adaptation. Following (Chu-Carroll and Brown, 1998), *task initiative* tracks the

	Utterance	Task Initiative	Dialogue Initiative
(1)	S: Hello, this is MIMIC, the movie information system.	User	System
(2)	How can I help you?		
(3)	U: What time is Analyze This playing in Montclair? [what time is analyze this playing in montclair]		
(4)	S: What theater would you like?	User	System
(5)	U: Montclair. [montclair]		
(6)	S: Choices for theater are Wellmont Theater and Clearviews Screening Zone.	System	System
(7)	What theater would you like?		
(8)	U: Clearviews Screening Zone. [um movie screening zone]		
(9)	S: Analyze This is playing at Clearviews Screening Zone in Montclair at 1:15pm, 4:10pm, and 6:30pm.	System	System
(10)	Please say the name of the movie or theater or town you would like information about.		
(11)	U: What time is Analyze This playing at the Wellmont Theater? [what time is analyze this playing at the wellmont theater]		
(12)	S: Analyze this is playing at Wellmont Theater in Montclair at 3:15pm, 5:45pm, 8:20pm, and 10:10pm.	User	System
(13)	Can I help you with anything else?		

Figure 1: An Adaptive Mixed Initiative Dialogue with MIMIC

lead in the process toward achieving the dialogue participants’ domain goal, while *dialogue initiative* models the lead in determining the current discourse focus. In our information query application domain, MIMIC has task (and thus dialogue) initiative when its utterances are intended to provide helpful guidance toward achieving the user’s domain goal, while it has dialogue but not task initiative if its utterances only specify the current discourse goal.² For example, as a result of MIMIC taking over task initiative in (6), helpful guidance, in the form of valid response choices, was provided in its attempt to obtain a theater name after the user failed to answer an earlier question intended to solicit this information. In (4), MIMIC specified the current discourse goal (requesting information about a missing theater) but did not suggest valid response choices since it only had dialogue initiative.

MIMIC’s ability to automatically adapt dialogue strategies is achieved by employing an initiative module that determines initiative distribution based on participant roles, cues detected during the current user utterance, and dialogue history (Chu-Carroll and Brown, 1998). This initiative framework utilizes the Dempster-Shafer theory (Shafer, 1976; Gordon and Shortliffe, 1984), and represents the current initiative distribution as two *basic probability assignments (bpas)* that signify the overall amount of evidence supporting each agent having task and dialogue initiatives. The effects that a cue has on changing the current task and dialogue initiative distribution are also represented as bpas, obtained using an iterative training procedure on a corpus of transcribed

²In the dialogues collected in our experiments, which are described in Section 3, there are system turns in which MIMIC had neither task nor dialogue initiative. However, such cases are rare in this domain and will not be discussed in this paper.

and annotated human-human dialogues. At the end of each user turn, the bpas representing the effects of cues detected during that turn are combined with the bpas representing the current initiative distribution to obtain the initiative distribution for the system’s next turn.

In Figure 1, utterance (3) triggered the cue *AmbiguousAction* since the town of Montclair has multiple theaters. Although *AmbiguousAction* contributed to the system having both initiatives in the next turn, the effect of this cue, when combined with the initiative distribution of (3), resulted in MIMIC having dialogue but not task initiative in (4). However, in addition to triggering *AmbiguousAction*, utterance (5) also triggered *NoNewInfo*, since it did not contribute any new information. The effect of these two cues, combined with the initiative distribution in (5), was enough to lead MIMIC to take over task initiative in (6).

2.2 Modification of Adaptation Behavior

MIMIC employs an initiative module which is decoupled from its goal selection process, while allowing the outcome of both processes to guide its response strategy selection. A major advantage of this design is that changes in the system’s strategy selection behavior, and thus its adaptation behavior, can be achieved by modifying parameters in the initiative module, while leaving the goal and strategy selection processes intact. The initiative module determines initiative distribution based on participant roles, cues observed in the current user utterance, and dialogue history. Although all three factors may play a role in modifying MIMIC’s adaptation behavior, we discuss how the effects of participant roles and dialogue history are modified to create systems used as bases of comparison in our experiments.

Participant roles affect the default initiative distribu-

	Utterance	Task Initiative	Dialogue Initiative
(14)	S: Hello, this is MIMIC, the movie information system.	System	System
(15)	Please say the name of the movie or theater or town you would like information about.		
(16)	U: Analyze This. <i>[analyze this]</i>		
(17)	S: Choices for task are when and where.	System	System
(18)	What would you like to know about Analyze This?		
(19)	U: When. <i>[when is]</i>		
(20)	S: Choices for theater are Middlebrook Galeria, Loews East Hanover, Montgomery Cinemas, Atlantic Triplex Cinema, and others.	System	System
(21)	What theater would you like?		
(22)	U: Montclair. <i>[montclair]</i>		
(23)	S: Choices for theater are Wellmont Theater and Clearviews Screening Zone.	System	System
(24)	What theater would you like?		
(25)	U: Wellmont. <i>[wellmont]</i>		
(26)	S: Analyze This is playing at Wellmont Theater in Montclair at 3:15pm, 5:45pm, 8:20pm, and 10:10pm.	System	System
(27)	Please say the name of the movie or theater or town you would like information about.		
(28)	U: Analyze This in Montclair. <i>[analyze this in montclair]</i>		
(29)	S: Choices for task are location, what, where, and when.	System	System
(30)	What would you like to know about Montclair?		
(31)	U: When. <i>[when]</i>		
(32)			
(33)	S: Choices for theater are Wellmont Theater and Clearviews Screening Zone	System	System
(34)	What theater would you like?		
(35)	U: Clearviews Screening Zone. <i>[screening zone]</i>		
(36)	S: Analyze This is playing at Clearviews Screening Zone in Montclair at 1:15pm, 4:10pm, and 6:30pm.	System	System
(37)	Please say the name of the movie or theater or town you would like information about.		

Figure 2: A System Initiative Dialogue with MIMIC-SI

tion for MIMIC’s initial utterance. For naive users, it may be more appropriate for MIMIC to employ system-initiative strategies. This system, MIMIC-SI, can be achieved by setting the default initiative distribution to provide definite evidence for the system having both task and dialogue initiatives.³ Figure 2 shows a dialogue in which a user interacts with MIMIC-SI to obtain the same information as in Figure 1. MIMIC-SI prompts the user for one piece of information at a time and provides (sometimes unnecessary) assistance, in the form of valid dialogue moves, during each turn.

Taking into account dialogue history allows MIMIC to determine the initiative distribution based on the cumulative effect of previously observed cues. This effect can be disabled so that only local dialogue context affects the initiative distribution for the next turn. Based on MIMIC’s parameter settings, this adjustment results in MIMIC-MI, a system which always has dia-

³Based on the Dempster-Shafer theory, if the bpas of the default initiative distribution or those of a detected cue provide definite evidence for drawing a certain conclusion, then no subsequent cues have any effect on changing that conclusion.

logue initiative but never has task initiative. MIMIC-MI behaves similarly to many existing non-adaptive mixed initiative dialogue systems (e.g., (Bennacef et al., 1996; Papineni et al., 1999)). Figure 3 shows a dialogue involving MIMIC-MI in which the user again attempts to acquire the same information as in the previous two dialogues. Notice that (41)-(49) constitute a failed attempt to solicit a theater name from the user. Instead of providing helpful information as in (6) in Figure 1, MIMIC-MI relied on the user to change her problem-solving strategy in (50) to obtain the information needed to answer the system’s question.

3 Experimental Design

Our main goal in evaluating MIMIC is to determine whether users find the mixed initiative and automatic adaptation aspects of its dialogue strategies useful. We compared MIMIC to two control systems: MIMIC-SI and MIMIC-MI, since they employ dialogue management strategies similar to those in many existing systems. The comparison between MIMIC and MIMIC-SI

	Utterance	Task Initiative	Dialogue Initiative
(38)	S: Hello, this is MIMIC, the movie information system.	User	System
(39)	How can I help you?		
(40)	U: When is Analyze This playing in Montclair? <i>[when analyze this playing in montclair]</i>		
(41)	S: What theater would you like?	User	System
(42)	U: <i>[]</i>		
(43)	S: What theater would you like?	User	System
(44)	U: Um. <i>[um]</i>		
(45)	S: What theater would you like?	User	System
(46)	U: Montclair. <i>[montclair]</i>		
(47)	S: What theater would you like?	User	System
(48)	U: Any. <i>[any]</i>		
(49)	S: What theater would you like?	User	System
(50)	U: Where in Montclair is Analyze This playing? <i>[where in montclair is analyze this playing]</i>		
(51)	S: Analyze This is playing at Wellmont Theater and Clearviews Screening Zone in Montclair.	User	System
(52)	U: When is Analyze This playing at Wellmont Theater? <i>[when is analyze this playing at wellmont theater in]</i>		
(53)	S: Analyze This is playing at Wellmont Theater in Montclair at 3:15pm, 5:45pm, 8:20pm, and 10:10pm.	User	System
(54)	Can I help you with anything else?		
(55)	U: When is it playing at Clearviews Screening Zone? <i>[when is it playing at movie screening zone]</i>		
(56)	S: Analyze This is playing at Clearviews Screening Zone in Montclair at 1:15pm, 4:10pm, and 6:30pm.	User	System
(57)	Can I help you with anything else?		

Figure 3: A Non-Adaptive Mixed Initiative Dialogue with MIMIC-MI

focused on the contribution of mixed-initiative dialogue management, while the comparison between MIMIC and MIMIC-MI emphasized the contribution of automatic strategy adaptation. The following three factors were controlled in our experiments:

1. **System version:** For each experiment, two systems were used: MIMIC and a control system. In the first experiment MIMIC was compared with MIMIC-SI, and in the second experiment, with MIMIC-MI.
2. **Order:** For each experiment, all subjects were randomly divided into two groups. One group performed tasks using MIMIC first, and the other group used the control system first.
3. **Task difficulty:** 3-4 tasks which highlighted differences between systems were used for each experiment. Based on the amount of information to be acquired, we divided the tasks into two groups: *easy* and *difficult*; an example of each is shown in Figure 4.

Town	Theater (if playing)	Movie	Times after 5:10pm (if playing)
Hoboken		Antz	

(a) Easy Task

Town	Theater (if playing)	Movie	Two Times (if playing)
Millburn		Analyze This	
Berkeley Hgts		Analyze This	
Mountainside		Analyze This	
Madison		True Crime	
Hoboken		True Crime	

(b) Difficult Task

Figure 4: Sample Tasks for Evaluation Experiments

Eight subjects⁴ participated in each experiment. Each of the subjects interacted with both systems to perform

⁴The subjects were Bell Labs researchers, summer students, and their friends. Most of them are computer scientists, electrical engi-

all tasks. The subjects completed one task per call so that the dialogue history for one task did not affect the next task. Once they had completed all tasks in sequence using one system, they filled out a questionnaire to assess **user satisfaction** by rating 8-9 statements, similar to those in (Walker et al., 1997), on a scale of 1-5, where 5 indicated highest satisfaction. Approximately two days later, they attempted the same tasks using the other system.⁵ These experiments resulted in 112 dialogues with approximately 2,800 dialogue turns.

In addition to user satisfaction ratings, we automatically logged, derived, and manually annotated a number of features (shown in boldface below). For each task/subject/system triplet, we computed the **task success rate** based on the percentage of slots correctly filled in on the task worksheet, and counted the **# of calls** needed to complete each task.⁶ For each call, the user-side of the dialogue was recorded, and the **elapsed time** of the call was automatically computed. All user utterances were logged as recognized by our automatic speech recognizer (ASR) and manually transcribed from the recordings. We computed the **ASR word error rate**, **ASR rejection rate**, and **ASR timeout rate**, as well as **# of user turns** and **average sentence length** for each task/subject/system triplet. Additionally, we recorded the **cues** that the system automatically detected from each user utterance. All system utterances were also logged, along with the **initiative distribution** for each system turn and the **dialogue acts** selected to generate each system response.

4 Results and Discussion

Based on the features described above, we compared MIMIC and the control systems, MIMIC-SI and MIMIC-MI, along three dimensions: *performance features*, in which comparisons were made using previously proposed features relevant to system performance (e.g., (Price et al., 1992; Simpson and Fraser, 1993; Danieli and Gerbino, 1995; Walker et al., 1997)); *discourse features*, in which comparisons were made using characteristics of the resulting dialogues; and *initiative distribution*, where initiative characteristics of all dialogues involving MIMIC from both experiments were examined.

4.1 Performance Features

For our performance evaluation, we first applied a three-way analysis of variance (ANOVA) (Cohen, 1995) to each feature using three factors: system version, order,

neers, or linguists, and none had prior knowledge of MIMIC.

⁵We used the exact same set of tasks rather than designing tasks of similar difficulty levels because we intended to compare all available features between the two system versions, including ASR word error rate, which would have been affected by the choice of movie/theater names in the tasks.

⁶Although the vast majority of tasks were completed in one call, some subjects, when unable to make progress, did not change strategies as in (41)-(49) in Figure 3; instead, they hung up and started the task over.

Performance Feature	MIMIC	SI	p
<i># of user turns</i>	<i>10.3</i>	<i>13.6</i>	<i>0.0075</i>
<i>Elapsed time (sec.)</i>	<i>229.5</i>	<i>277.5</i>	<i>0.0162</i>
<i>ASR timeout (%)</i>	<i>12.5</i>	<i>6.9</i>	<i>0.0239</i>
<i>User satisfaction (n=8)</i>	<i>21.9</i>	<i>19.8</i>	<i>0.0447</i>
ASR rejection (%)	5.4	8.1	0.1911
Task success (%)	100	98.8	0.3251
# of calls	1.0	1.1	0.572
ASR word error (%)	28.1	31.1	0.8475

(a) MIMIC vs. MIMIC-SI (n=32)

Performance Feature	MIMIC	MI	p
<i>ASR timeout (%)</i>	<i>5.7</i>	<i>15.6</i>	<i>0.001</i>
<i># of user turns</i>	<i>10.3</i>	<i>14.3</i>	<i>0.0199</i>
<i>User satisfaction (n=8)</i>	<i>29.5</i>	<i>24.4</i>	<i>0.0364</i>
<i>Elapsed time (sec.)</i>	<i>200.6</i>	<i>246.4</i>	<i>0.0457</i>
ASR word error (%)	23.0	30.6	0.0588
Task success (%)	100	98.4	0.1639
# of calls	1.21	1.21	0.5
ASR rejection (%)	8.4	7.7	0.8271

(b) MIMIC vs. MIMIC-MI (n=24)

Table 1: Comparison of Performance Features

and task difficulty.⁷ If no interaction effects emerged, we compared system versions using paired sample t-tests.⁸

Following the PARADISE evaluation scheme (Walker et al., 1997), we divided performance features into four groups:

- Task success: task success rate, # of calls.
- Dialogue quality: ASR rejection rate, ASR timeout rate, ASR word error rate.
- Dialogue efficiency: # of user turns, elapsed time.
- System usability: user satisfaction.

For both experiments, the ANOVAs showed no interaction effects among the controlled factors. Tables 1(a) and 1(b) summarize the results of the paired sample t-tests based on performance features, where features that differed significantly between systems are shown in italics.⁹ These results show that, when compared with either

⁷User satisfaction was a per subject as opposed to a per task performance feature; thus, we performed a two-way ANOVA using the factors system version and order.

⁸This paper focuses on evaluating the effect of MIMIC’s mixed initiative and automatic adaptation capabilities. We assess these effects based on comparisons between system version when no interaction effects emerged from the ANOVA tests using the factors system version, order, and task difficulty. Effects based on system order and task difficulty alone are beyond the scope of this paper.

⁹Typically $p < 0.05$ is considered statistically significant (Cohen, 1995).

control system, users were more satisfied with MIMIC¹⁰ and that MIMIC helped users complete tasks more efficiently. Users were able to complete tasks in fewer turns and in a more timely manner using MIMIC.

When comparing MIMIC and MIMIC-MI, dialogues involving MIMIC had a lower timeout rate. When MIMIC detected cues signaling anomalies in the dialogue, it adapted strategies to provide assistance, which in addition to leading to fewer timeouts, saved users time and effort when they did not know what to say. In contrast, users interacting with MIMIC-MI had to iteratively reformulate questions until they obtained the desired information from the system, leading to more timeouts (see (41)-(49) in Figure 3). However, when comparing MIMIC and MIMIC-SI, even though users accomplished tasks more efficiently with MIMIC, the resulting dialogues contained more timeouts. As opposed to MIMIC-SI, which always prompted users for one piece of information at a time, MIMIC typically provided more open-ended prompts when the user had task initiative. Even though this required more effort on the user’s part in formulating utterances and led to more timeouts, MIMIC quickly adapted strategies to assist users when recognized cues indicated that they were having trouble.

To sum up, our experiments show that both MIMIC’s mixed initiative and automatic adaptation aspects resulted in better performance along the *dialogue efficiency* and *system usability* dimensions. Moreover, its adaptation capabilities contributed to better performance in terms of *dialogue quality*. MIMIC, however, did not contribute to higher performance in the *task success* dimension. In our movie information domain, the tasks were sufficiently simple; thus, all but one user in each experiment achieved a 100% task success rate.

4.2 Discourse Features

Our second evaluation dimension concerns characteristics of resulting dialogues. We analyzed features of user utterances in terms of utterance length and cues observed and features of system utterances in terms of dialogue acts. For each feature, we again applied a three-way ANOVA test, and if no interaction effects emerged, we performed a paired sample t-test to compare system versions.

The cues detected in user utterances provide insight into both user intentions and system capabilities. The cues that MIMIC automatically detects are a subset of those discussed in (Chu-Carroll and Brown, 1998):¹¹

- *TakeOverTask*: triggered when the user provides more information than expected; an implicit indication that the user wants to take control of the

¹⁰The range of user satisfaction scores was 8-40 for experiment one and 9-45 for experiment two.

¹¹A subset of these cues corresponds loosely to previously proposed evaluation metrics (e.g., (Danieli and Gerbino, 1995)). However, our system automatically detects these features instead of requiring manual annotation by experts.

Discourse Feature	MIMIC	SI	p
<i>Cue: TakeOverTask</i>	1.84	5	0
<i>Cue: AmbiguousActResolved</i>	1.69	4.59	0
<i>Cue: AmbiguousAction</i>	3	6.59	0.0008
Avg sentence length (words)	6.82	5.45	0.0016
<i>Cue: InvalidAction</i>	1.16	0.94	0.1738
<i>Cue: NoNewInfo</i>	1.28	1.38	0.766

(a) MIMIC vs. MIMIC-SI (n=32)

Discourse Feature	MIMIC	MI	p
<i>Cue: TakeOverTask</i>	2.33	0	0
<i>Cue: InvalidAction</i>	2.04	3.75	0.0011
<i>Cue: NoNewInfo</i>	2.25	4.79	0.0161
<i>Cue: AmbiguousActResolved</i>	2.08	1.13	0.0297
Avg sentence length (words)	5.26	5.63	0.1771
<i>Cue: AmbiguousAction</i>	4.13	4.38	0.8767

(b) MIMIC vs. MIMIC-MI (n=24)

Table 2: Comparison of User Utterance Features

problem-solving process.

- *NoNewInfo*: triggered when the user is unable to make progress toward task completion, either when the user does not know what to say or the ASR engine fails to recognize the user’s utterance.
- *InvalidAction/InvalidActionResolved*: triggered when the user utterance makes an invalid assumption about the domain and when the invalid assumption is corrected, respectively.
- *AmbiguousAction/AmbiguousActionResolved*: triggered when the user query is ambiguous and when the ambiguity is resolved, respectively.

Tables 2(a) and (b) summarize the results of the paired sample t-tests based on user utterance features where features whose numbers of occurrences were significantly different according to system version used are shown in italics.¹² Table 2(a) shows that users expected the system to adapt its strategies when they attempted to take control of the dialogue. Even though MIMIC-SI did not behave as expected, the users continued their attempts, resulting in significantly more occurrences of *TakeOverTask* in dialogues with MIMIC-SI than with MIMIC. Furthermore, the average sentence length in dialogues with MIMIC was only 1.5 words per turn longer than in dialogues with MIMIC-SI, providing further evidence that users

¹²Since system dialogue acts are often selected based on cues detected in user utterances, we only discuss results of our user utterance feature analysis, using dialogue act analysis results as additional support for our conclusions.

preferred to provide free-formed queries, regardless of system version used.

Table 2(b) shows that MIMIC was more effective at resolving dialogue anomalies than MIMIC-MI. More specifically, there were significantly fewer occurrences of *NoNewInfo* in dialogues with MIMIC than with MIMIC-MI. In addition, while the number of occurrences of *AmbiguousAction* was not significantly different for the two systems, the number that were resolved (*AmbiguousActionResolved*) was significantly higher in interactions with MIMIC than with MIMIC-MI. Since *NoNewInfo* and *AmbiguousAction* both prompted MIMIC to adapt strategies and, as a result, provide additional useful information, the user was able to quickly resolve the problem at hand. This is further supported by the higher frequency of the system dialogue act *GiveOptions* in MIMIC ($p=0$), which provides helpful information based on dialogue context.

In sum, the results of our discourse feature analysis further confirm the usefulness of MIMIC’s adaptation capabilities. Comparisons with MIMIC-SI provide evidence that MIMIC’s ability to give up initiative better matched user expectations. Moreover, comparisons with MIMIC-MI show that MIMIC’s ability to opportunistically take over initiative resulted in dialogues in which anomalies were more efficiently resolved and progress toward task completion was more consistently made.

4.3 Initiative Analysis

Our final analysis concerns the task initiative distribution in our adaptive system in relation to the features previously discussed. For each dialogue involving MIMIC, we computed the percentage of turns in which MIMIC had task initiative and the *correlation coefficient* (r) between the initiative percentage and each performance/discourse feature. To determine if this correlation was significant, we performed *Fisher’s r to z transform*, upon which a conventional Z test was performed (Cohen, 1995).

Tables 3(a) and (b) summarize the correlation between the performance and discourse features and the percentage of turns in which MIMIC has task initiative, respectively.¹³ Again, those correlations which are statistically significant are shown in italics. Table 3(a) shows a strong positive correlation between task initiative distribution and the number of user turns as well as the elapsed time of the dialogues. Although earlier results (Table 1(a)) show that dialogues in which the system always had task initiative tended to be longer, we believe that this correlation also suggests that MIMIC took over task initiative more often in longer dialogues, those in which the user was more likely to be having difficulty. Table 3(a) further shows moderate correlation between task initiative distribution and ASR rejection rate as well as ASR word error rate. It is possible that such a correlation exists

¹³This test was not performed for user satisfaction, since user satisfaction was a per subject and not a per dialogue feature.

Performance Feature	r	p
<i># of user turns</i>	<i>0.71</i>	<i>0</i>
<i>ASR rejection</i>	<i>0.55</i>	<i>0</i>
<i>Elapsed time</i>	<i>0.51</i>	<i>0.00002</i>
<i>ASR word error</i>	<i>0.46</i>	<i>0.00012</i>
# of calls	0.15	0.1352
ASR timeout	-0.003	0.4911
Task success rate	0	0.5

(a) Performance Features

Discourse Feature	r	p
<i>Cue: AmbiguousActionResolved</i>	<i>0.61</i>	<i>0</i>
<i>Cue: NoNewInfo</i>	<i>0.59</i>	<i>0</i>
<i>Cue: TakeOverTask</i>	<i>0.44</i>	<i>0.00028</i>
<i>Cue: InvalidAction</i>	<i>0.42</i>	<i>0.00057</i>
<i>Average sentence length</i>	<i>-0.40</i>	<i>0.00099</i>
<i>Cue: AmbiguousAction</i>	<i>0.38</i>	<i>0.00169</i>

(b) Discourse Features

Table 3: Correlation Between Task Initiative Distribution and Features (n=56)

because ASR performance worsens when MIMIC takes over task initiative. However, in that case, we would have expected the results in Section 4.1 to show that the ASR rejection and word error rates for MIMIC-SI are significantly greater than those for MIMIC, which are in turn significantly greater than those for MIMIC-MI, since in MIMIC-SI the system always had task initiative and in MIMIC-MI the system never took over task initiative. To the contrary, Tables 1(a) and 1(b) showed that the differences in ASR rejection rate and ASR word error rate were not significant between system versions, and Table 1(b) showed that ASR word error rate for MIMIC-MI was in fact quite substantially higher than that for MIMIC. This suggests that the causal relationship is the other way around, i.e., MIMIC’s adaptation capabilities allowed it to opportunistically take over task initiative when ASR performance was poor.

Table 3(b) shows that all cues are positively correlated with task initiative distribution. For *AmbiguousAction*, *InvalidAction*, and *NoNewInfo*, this correlation exists because observation of these cues contributed to MIMIC having task initiative. However, note that *AmbiguousActionResolved* has a stronger positive correlation with task initiative distribution than does *AmbiguousAction*, again indicating that MIMIC’s adaptive strategies contributed to more efficient resolution of ambiguous actions.

In brief, our initiative analysis lends additional support to the conclusions drawn in our performance and discourse feature analyses and provides new evidence for the advantages of MIMIC’s adaptation capabilities.

In addition to taking over task initiative when previously identified dialogue anomalies were encountered (e.g., detection of ambiguous or invalid actions), our analysis shows that MIMIC took over task initiative when ASR performance was poor, allowing the system to better constrain user utterances.¹⁴

5 Conclusions

This paper described an empirical evaluation of MIMIC, an adaptive mixed initiative spoken dialogue system. We conducted two experiments that focused on evaluating the mixed initiative and automatic adaptation aspects of MIMIC and analyzed the results along three dimensions: performance features, discourse features, and initiative distribution. Our results showed that both the mixed initiative and automatic adaptation aspects of the system led to better performance in terms of user satisfaction and dialogue efficiency. In addition, we found that MIMIC's adaptation behavior better matched user expectations, more efficiently resolved anomalies in dialogues, and led to higher overall dialogue quality.

Acknowledgments

We would like to thank Bob Carpenter and Christine Nakatani for their help on experimental design, Jan van Santen for discussion on statistical analysis, and Bob Carpenter for his comments on an earlier draft of this paper. Support for the second author is provided by an NSF graduate fellowship and a Lucent Technologies GRPW grant.

References

James F. Allen, Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. 1996. A robust system for natural spoken dialogue. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 62–70.

S. Bennacef, L. Devillers, S. Rosset, and L. Lamel. 1996. Dialog in the RAILTEL telephone-based system. In *Proceedings of the 4th International Conference on Spoken Language Processing*.

Jennifer Chu-Carroll and Michael K. Brown. 1998. An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction*, 8(3-4):215–253.

Jennifer Chu-Carroll. 2000. MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. In *Proceedings of the 6th ACL Conference on Applied Natural Language Processing*. To appear.

Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press.

Morena Danieli and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language

system. In *Proceedings of the AAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39.

- Jean Gordon and Edward H. Shortliffe. 1984. The Dempster-Shafer theory of evidence. In Bruce Buchanan and Edward Shortliffe, editors, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 13, pages 272–292. Addison-Wesley.
- Diane J. Litman and Shimei Pan. 1999. Empirically evaluating an adaptable spoken dialogue system. In *Proceedings of the 7th International Conference on User Modeling*, pages 55–64.
- H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao, J. Polifroni, S. Seneff, and V. Zue. 1996. WHEELS: A conversational system in the automobile classifieds domain. In *Proceedings of the International Conference on Spoken Language Processing*, pages 542–545.
- Stefan Ortmanns, Wolfgang Reichl, and Wu Chou. 1999. An efficient decoding method for real time speech recognition. In *Proceedings of the 5th European Conference on Speech Communication and Technology*.
- K.A. Papineni, S. Roukos, and R.T. Ward. 1999. Free-flow dialog management using forms. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 1411–1414.
- Patti Price, Lynette Hirschman, Elizabeth Shriberg, and Elizabeth Wade. 1992. Subject-based evaluation measures for interactive spoken language systems. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 34–39.
- Wolfgang Reichl and Wu Chou. 1998. Decision tree state tying based on segmental clustering for acoustic modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.
- M.D. Sadek, A. Ferrieux, A. Cozannet, P. Bretier, F. Panaget, and J. Simonin. 1996. Effective human-computer cooperative spoken dialogue: The AGS demonstrator. In *Proceedings of the International Conference on Spoken Language Processing*.
- Glenn Shafer. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.
- Andrew Simpson and Norman M. Fraser. 1993. Black box and glass box evaluation of the SUNDIAL system. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 1423–1426.
- Gert Veldhuijzen van Zanten. 1999. User modelling in adaptive dialogue management. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 1183–1186.
- Marilyn A. Walker, Diane J. Litman, Candance A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 271–280.

¹⁴Although not currently utilized, the ability to adapt dialogue strategies when ASR performance is poor enables the system to employ dialogue strategy specific language models for ASR.