Adapting to Multiple Affective States in Spoken Dialogue

Kate Forbes-Riley
Learning R&D Ctr (LRDC)
University of Pittsburgh
Pittsburgh, PA 15260, USA
forbesk@cs.pitt.edu

Diane Litman
LRDC and Dept. Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
litman@cs.pitt.edu

Abstract

We evaluate a wizard-of-oz spoken dialogue system that adapts to multiple user affective states in real-time: user disengagement and uncertainty. We compare this version with the prior version of our system, which only adapts to user uncertainty. Our analysis investigates how iteratively adding new affect adaptation to an existing affect-adaptive system impacts global and local performance. We find a significant increase in motivation for users who most frequently received the disengagement adaptation. Moreover, responding to disengagement breaks its negative correlations with task success and user satisfaction, reduces uncertainty levels, and reduces the likelihood of continued disengagement.

1 Introduction

State of the art spoken dialogue system research focuses on responding not only to the literal content of users' speech but also to their affective state[1], such that the same literal content may receive one system response when the user is frustrated, and another when the user is confused, etc. The potential benefits are clear: affect-adaptive systems can increase task success (Forbes-Riley and Litman, 2011a; Wang et al., 2008) and other global performance metrics such as user satisfaction (Liu and Picard, 2005; Klein et al., 2002) and motivation (Aist et al., 2002). However, to date most researchers have focused on adapting to a single affective state. The next step is thus to develop and evaluate spoken dialogue systems that respond to multiple user affective states. The problem of how to develop effective affect adaptations is a complex one even as applied to a single affective state, and it multiples with every new state added. For example, it is not clear a priori how responding to one affective state may impact another's frequency and relationship to performance. In this paper we examine this problem in the context of the computer tutoring domain. We previously showed that adapting to user *uncertainty* during spoken dialogue computer tutoring improves task success, both in a wizard-of oz version where a hidden human performed the affect detection and natural language understanding (Forbes-Riley and Litman, 2011b), as well as in a fully automated system version (Forbes-Riley and Litman, 2011a).

We are now taking the next step by incorporating adaptation to a second user affective state: user *disengagement*. We target user disengagement for two reasons: first, our prior manual annotation showed disengagement and uncertainty to be the most frequent user affective states that occur in our system, and second, our prior analyses show that the occurrence of disengagement is negatively correlated with task success and user satisfaction (Forbes-Riley and Litman, 2012).[2] Thus, we hypothesized that providing appropriate system responses to both affective states could have multiple benefits: 1) reduce the frequency of one or both states, 2) "break" the nega-

[1] We use *affect* for emotions and attitudes that affect how users communicate. Other speech researchers also combine concepts of emotion, arousal, and attitudes where emotion is not full-blown (Cowie and Cornelius, 2003).

[2] Redesigning a system in light of correlational analyses can improve performance (Rotaru and Litman, 2009).

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 217–226,
Seoul, South Korea, 5-6 July 2012.

tive correlations with performance, and 3) yield further improvements in global and local performance.

In this paper, we test these hypotheses, presenting the results of a controlled experiment evaluating a wizard-of-oz version of our spoken dialogue computer tutor that adapts to both user uncertainty and user disengagement (Section 3). Although we address these states within the tutoring domain, speech researchers from other domains and applications are also focusing on detecting and adapting to user disengagement (e.g., (Schuller et al., 2010; Wang and Hirschberg, 2011)) and uncertainty (e.g. (Pon-Barry and Shieber, 2011; Paek and Ju, 2008)) to improve system performance. Our results should be of interest not only to these researchers but also more generally to any researchers working towards comprehensive affect-adaptive spoken dialogue systems. In particular, our results show that iteratively adding new affect adaptations to an existing affect-adaptive system can yield performance improvements. We find no increase (but also no decrease) in task success or user satisfaction, but we do find an increase in motivation for users who most frequently received the disengagement adaptation (Section 4). Furthermore, we find that responding to disengagement "breaks" negative correlations with task success and user satisfaction (Section 5), and also yields a reduction both in uncertainty levels (Section 4) and in the likelihood of continued disengagement (Section 6).

2 Related Work

User disengagement is highly undesirable because of its potential to increase dissatisfaction and task failure, and there is a growing awareness of its potential to negatively impact commercial applications; thus there has been substantial prior work focused on detecting disengagement (along with the closely related states of boredom and lack of interest) (e.g., (Schuller et al., 2010; Wang and Hirschberg, 2011; Bohus and Horvitz, 2009)). To date, however, only a few disengagement-adaptive systems have been evaluated, and within the tutoring domain these have focused on only one disengagement behavior: *gaming*. For example, responding to gaming with supplementary material reduced gaming and improved task success for users who most frequently gamed (Baker et al., 2006), while adding progress reports and productive learning tips at the end of problems (i.e., without specifically targeting gaming instances) increased task success, engagement, and user satisfaction (Arroyo et al., 2007). Our research builds on this work but is novel in that we focus on speech and dialogue-based disengagement and on adapting to multiple affective states.

More generally, while substantial spoken dialogue and affective systems research has shown that users display a range of affective states when interacting with a system (e.g. (Schuller et al., 2009; Conati and Maclaren, 2009)), to date only a few systems adapt to multiple affective states (e.g., (D'Mello et al., 2010; Aist et al., 2002; Tsukahara and Ward, 2001)). Most have been deployed with wizard-of-oz components, and none have yet shown significant improvements in task success, though other benefits have been shown, including increased user satisfaction (Tsukahara and Ward, 2001), rapport (Acosta and Ward, 2011) and motivation (Aist et al., 2002). Recently, D'Mello et al. (2010) showed that performance can depend on when and to whom the adaptations are provided; higher expertise users never benefited from system responses to their frustration, boredom and confusion, while lower expertise users only benefited after multiple system interactions. While this prior work showed the benefits of adapting to multiple affective states as compared to not adapting to affect at all, it did not test whether these benefits were due to having multiple adaptations, or if any one would have sufficed. Our work is novel in explicitly measuring the value of having multiple adaptations as compared to one.

3 The Experiment

Our prior work showed that our uncertainty-adaptive spoken dialogue system improves performance over not adapting to affect (Forbes-Riley and Litman, 2011b; Forbes-Riley and Litman, 2011a); this system serves as our baseline in the current work.

3.1 Baseline System: UNC_ADAPT ITSPOKE

UNC_ADAPT ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialog system)[3] tutors 5 Newtonian

[3]ITSPOKE is a speech-enhanced and modified version of the Why2-Atlas text-based tutor (VanLehn et al., 2002).

physics problems (one per dialogue), using a Tutor Question - User Answer - Tutor Response format.

In the fully automated system, the speech from the user's answer is digitized from head-mounted microphone input and sent to a speech recognizer. The answer's (in)correctness is then automatically classified based on the recognizer's transcription using a semantic analysis component, and the answer's (un)certainty is automatically classified by inputting features of the speech signal (e.g. prosody), the automatic transcript, and the dialogue context into a logistic regression model. The (in)correctness and (un)certainty detection components comprising our system's user model are described in detail elsewhere (Forbes-Riley and Litman, 2011a).

For the present experiment, the affect and (in)correctness labeling are performed by a hidden human wizard. As in our prior work, this allows us to first analyze the impact of an affect adaptation separately from the noise introduced by automating affect and semantic analysis (see Section 7). Figures 1-3 illustrate the binary (dis)engagement (ENG, DISE), (in)correctness (COR, INC), and (un)certainty (CER, UNC) labels.

Finally, the system automatically determines the appropriate response based on the answer's labeled (in)correctness and (un)certainty and this response is sent to the Cepstral text-to-speech system[4], whose audio output is played through the headphones and displayed on a web-based interface (see Figure 4).

The uncertainty label and system adaptation are described in detail elsewhere (Forbes-Riley and Litman, 2011b; Forbes-Riley and Litman, 2011a). Briefly, the *uncertain (UNC)* label is used for turns expressing uncertainty or confusion about the topic being discussed, and the *non-uncertain (CER)* label is used otherwise. The wizard in this experiment displayed interannotator agreement of 0.85 and 0.62 Kappa on correctness and uncertainty, respectively, in prior ITSPOKE corpora. Our uncertainty adaptation is based on the hypothesis that uncertainty and incorrectness are both points of impasse in a dialogue, and that providing additional knowledge can help resolve them. In UNC_ADAPT ITSPOKE, incorrect answers and uncertain answers both receive (in)correctness feedback (e.g., "Right" or "I don't think so"), followed by a (re)statement of the correct answer. Depending on topic difficulty, the system then either provides a brief explanation of reasoning ("Bottom Out") or a more lengthy dialogue exchange that walks the user through the steps of the reasoning ("Remediation Subdialogue"). An example is shown in Figure 1.

[4]an outgrowth of Festival (Black and Taylor, 1997).

3.2 UNC-DISE_ADAPT ITSPOKE

UNC-DISE_ADAPT ITSPOKE adds disengagement detection and adaptation to UNC_ADAPT ITSPOKE. Our disengagement annotation scheme is described in detail elsewhere (Forbes-Riley and Litman, 2011c). It was derived from empirical observation of our data and from prior work, including that mentioned in Section 2 and appraisal theory-based emotion models, which distinguish emotional behaviors from their underlying causes (e.g., (Conati and Maclaren, 2009)). Briefly, the *Disengaged (DISE)* label is used for turns expressing moderate to strong disengagement towards the interaction, i.e., responses given without much effort or caring about appropriateness, and might include signs of boredom or irritation. Clear examples include turns spoken in leaden monotone, with sarcasm, or off-task sounds such as electronics usage. The wizard in this experiment displayed interannotator agreement of 0.55 Kappa on the DISE label in prior ITSPOKE corpora, which is on par with prior affect research, where moderate agreement is common given the difficulty of the task (Forbes-Riley and Litman, 2011c).

Based on the results of the prior research discussed in Section 2 and our own prior research, we have developed one class of system responses for correct+disengaged (COR-DISE) answers and another for incorrect+disengaged (INC-DISE) answers (Forbes-Riley and Litman, 2011c)[5].

Our INC-DISE adaptation builds on the prior finding that supplementary information can help reduce some types of disengagement for highly disengaged users (Baker et al., 2006). We hypothesized that our UNC_ADAPT response to incorrectness (a Bottom Out or Remediation Subdialogue) was insufficient for an INC-DISE turn because the

[5]Originally we distinguished six DISE types, but found this too many to be reliably detected automatically and thus reduced the distinction to two using correctness. Our automatic disengagement detector is discussed further in Section 7.

user had already disengaged. To benefit from this supplementary knowledge, the user first had to reengage. Thus, the UNC-DISE_ADAPT system responds to INC-DISE answers with "productive interaction feedback"[6] followed by an easier "fill in the blank" version of the original question. The purpose of this two-pronged response is to regain the user's attention with the feedback and then provide a path through the impasse with the easier question, thereby keeping the user engaged. An example is shown in Figure 2, where **USER-1** is labeled INC-DISE because the user gives an irrelevant (and obviously incorrect) answer. Note that while most knowledge asymmetry spoken dialogue systems (e.g., problem-solving and troubleshooting (Janarthanam and Lemon, 2008)) use the concept of response *(in)correctness*, a more general version is response *(in)appropriateness*, which can be realized differently across applications, including as the user turn's speech recognition score (Kamm et al., 1998). Since misrecognitions are also a type of dialogue impasse, a similar version of our INC-DISE adaptation could be provided by other spoken dialogue systems for turns where users disengage and their response isn't recognized by the system.

Our COR-DISE adaptation builds on the prior findings that progress reports and productive learning tips can positively impact multiple performance metrics when used without specifically targeting disengagement (Arroyo et al., 2007), but not when used after every user turn (Walonoski and Heffernan, 2006). We hypothesized that these responses might be most beneficial if they targeted COR-DISE turns. Thus, the UNC-DISE_ADAPT system responds to COR-DISE answers with "productive interaction feedback" followed by a progress report graphing the user's correctness both in the current dialogue and over all prior dialogues. Examples are shown in Figures 3-4, where **USER-1** is labeled COR-DISE because the user unnecessarily repeats himself, signaling his lack of interest. As shown, we distinguish two classes of productive interaction feedback. That in "2a" shows the feedback given when the progress report indicates improvement on the current dialogue relative to the prior ones, while "2b" shows the feedback given when there is a decline. Note that a similar combination of productive interaction feedback and progress reports tailored to the domain (e.g., graphs showing subtasks accomplished so far) could be provided by most spoken dialogue systems on turns where users disengage and their response is recognized by the system.[7]

3.3 Experimental Procedure

College students with no college-level physics were recruited and randomly assigned to either the UNC_ADAPT or UNC-DISE_ADAPT condition after balancing for user expertise (pretest score) and gender. Users: (1) read a short physics text, (2) took a pretest and a pre-motivation survey, (3) worked 5 "training" problem dialogues with the system from their condition, (4) took a post-motivation survey and a user satisfaction survey, (5) took a posttest isomorphic to the pretest, and (6) worked a "test" problem dialogue with UNC_ADAPT.

The pre/post tests are the same as those used in multiple prior ITSPOKE experiments (c.f., (Forbes-Riley and Litman, 2011a)). The tests are isomorphic, each containing 26 multiple choice questions querying knowledge of the topics covered in the dialogues. Average pretest and posttest scores were 53% and 81% (out of 100%), respectively.

The pre/post motivation surveys are a reduced version of a widely used motivation survey in the tutoring domain (Pintrich and DeGroot, 1990); our selected questions were relevant to our system and also selected in other recent research (Ward, 2010; Roll, 2009). The two surveys are isomorphic, each containing 19 statements rated on a 7-point Likert scale. Average pre and post scores were 68% and 70% (out of 100%), respectively.

The user satisfaction survey was recently developed and validated for use with spoken dialogue computer tutors (Dzikovska et al., 2011). It contains 40 statements rated on a 5-point Likert scale. Average score was 68% (out of 100%).

The "test" dialogue is isomorphic to the fifth training dialogue, such that all questions are identical except for the identities of the objects discussed. In this way, we can measure how the disengagement

[6]This is our generalization of the concept of "productive learning tip" used in prior work (Arroyo et al., 2007).

[7]Note that our DISE and UNC adaptations are combined if the two states occur simultaneously.

adaptations from the fifth dialogue impact user turns when the questions are repeated in the test dialogue (where no disengagement adaptation is given). We have also used this test dialogue in our prior work (c.f., (Forbes-Riley and Litman, 2011a)).

3.4 Corpus

The resulting corpus contains 228 dialogues (6 per user) and 3518 turns from 38 users, 22 female and 16 male, with 19 subjects per condition.[8] Table 1 shows the distribution of the labeled turns in the corpus.

Table 1: Corpus Description (N=3518)

Turn Label	Total	Percent
Disengaged	622	17.7%
Correct	2825	80.3%
Disengaged+Correct	247	7.0%
Uncertain	537	15.3%

4 Global Performance Evaluation

We use the test and survey instruments described in Section 3.3 to evaluate global performance in UNC-DISE_ADAPT. We measure task success via learning gain; as is typical in the tutoring community, we compute normalized learning gain as (posttest-pretest)/(1-pretest). We compute percent user satisfaction from the survey as (user score)/(maximum possible score). We compute raw motivation gain from the surveys as (post score-pre score).[9] For each metric, we ran a one-way ANOVA with condition as the between-subjects factor. The first two rows of Table 2 show the number of users (N), means (Mn) and standard deviations (sd) for these metrics across condition. Although UNC-DISE_ADAPT shows a small decrease in means for learning gain and user satisfaction, there were no significant differences ($p \leq .05$) or trends ($p \leq .10$) for differences between conditions for any global metric.

As a further comparison, we compared the performance of UNC-DISE_ADAPT to our non-adaptive wizard-of-oz version of ITSPOKE (NO_ADAPT), using the corpus collected from our prior user study comparing UNC_ADAPT and NO_ADAPT; that study showed UNC_ADAPT had significantly higher learning gain than NO_ADAPT (p=.001) (Forbes-Riley and Litman, 2011b).[10] The goal here was to ascertain in a post-hoc way whether adapting to multiple affective states yielded higher task success than not adapting to affect at all. As shown last in Table 2, UNC-DISE_ADAPT and UNC_ADAPT both significantly outperform NO_ADAPT ($p \leq .003$), suggesting that while iteratively adding new affect adaptations to an existing affect-adaptive system does not necessarily yield additive improvements to global performance, it also does not decrease performance.

Table 2: Global Performance Metrics Across Conditions (All UNC vs. UNC-DISE Differences Yield $p >= .274$; All NO-ADAPT Differences Yield $p \leq .003$)

Cond	N	LearnGain		UserSat		MotGain	
		Mn	sd	Mn	sd	Mn	sd
Unc	19	.65	.20	.69	.11	.01	.07
Unc-Dise	19	.58	.19	.66	.09	.01	.07
NoAdapt	21	.38	.20	-	-	-	-

The frequency of disengagement and other affective states can vary widely across system users. In our case, some users showed disengagement on the majority of turns in later dialogues while others showed almost none at all; the average and standard deviation of per user %DISE over conditions are 17.7% and 10.1%, respectively (Table 5 breaks this down by condition). Thus we hypothesized that the global performance improvements of UNC-DISE_ADAPT might have been weakened by including users with low or no disengagement who rarely received the adaptation and thus could not be expected to show improvement. To test this hypothesis, we split users into *high* and *low* DISE based on the median %DISE in the corpus. We ran a two-way ANOVA for each global metric with DISE split and condition as factors. We found a significant interaction effect between condition and DISE

[8] One outlier with negative learning was removed from each condition, because our goal is to investigate the role of affect adaptation when learning is successful.

[9] Total, average or percent satisfaction yielded comparable results, as did raw or normalized motivation and learning gains.

[10] Because this prior corpus was collected in a different experiment, the conclusions here are tenuous. However, both experiments had similar subject populations (local college students) and mean pretest scores (p=.84). The prior experiment used a smaller satisfaction survey and no motivational surveys, so we can only compare learning.

split (F(1,38) = 4.84, p=0.035) for motivation gain. Means for these groups are shown in Table 3. As shown, *low* DISE users had higher motivation gain in UNC_ADAPT, while *high* DISE users had higher motivation gain in UNC-DISE_ADAPT.

Table 3: Motivation Gain Differences Across Condition for High and Low DISE Users (p=.035)

Condition	Split	N	MotGain	
			Mn	sd
UNC	high DISE	9	-.01	.04
UNC-DISE	high DISE	7	.04	.07
UNC	low DISE	10	.03	.08
UNC-DISE	low DISE	12	-.01	.06

In contrast to the tests and surveys, which do not necessarily reflect user performance during the dialogues, the "test" dialogue enables us to measure global performance using dialogue-based metrics. The test dialogue was isomorphic with the final training dialogue, except that the disengagement adaptation was not given; moreover, different system questions could appear in the test dialogue if the user answered a question differently.[11] We hypothesized that responding to the user's disengagement during the training dialogue (UNC-DISE_ADAPT) would yield increased correctness as well as reduced uncertainty and disengagement in the test dialogue.

We tested this hypothesis by computing percent correctness, disengagement, and uncertainty for each user, both alone and in combination, over user answers to tutor questions that were repeated between the training and test dialogues. We ran ANOVAs comparing these metrics across the two conditions. Table 4 presents our results. Interestingly, no differences between conditions were found for transitions from DISE turns. However, the disengagement adaptation did impact other turns in the dialogues apart from the (DISE) ones that triggered it. The first row shows that uncertain answers are more likely to remain uncertain in UNC_ADAPT than in UNC-DISE_ADAPT. The second row shows that incorrect+uncertain+engaged answers are more likely to become correct and certain in UNC-DISE_ADAPT. By more fully engaging users, the disengagement adaptation may thereby enable them to benefit more from the uncertainty adaptation. However, the third row suggests that the adaptation can have a negative impact when users are originally certain about their incorrect answers: incorrect+certain+engaged users turns are more likely to become disengaged in UNC-DISE_ADAPT. This suggests that the disengagement adaptation does not more fully engage certain users (particularly those whose certainty does not reflect correctness).

[11] For example, if a user answered a question incorrectly during training and then answered its isomorph correctly during testing, s/he would not receive the remediation during the test dialogue that s/he received during training.

Table 4: Differences Across Condition for Test Dialogue

Metric	Condition	Mn	sd	p
UNC → UNC	UNC	.06	.09	.05
	UNC-DISE	.01	.04	
INC+UNC+ENG →	UNC	.01	.03	.10
COR+CER+ENG	UNC-DISE	.03	.05	
INC+CER+ENG →	UNC	.00	.00	.04
INC+CER+DISE	UNC-DISE	.02	.03	

5 Breaking Negative Correlations

As noted in Section 1, in our prior ITSPOKE corpora we found that user disengagement was negatively correlated with task success (measured as learning gain) (p=.01) and user satisfaction (p=.03) (Forbes-Riley and Litman, 2011c; Forbes-Riley and Litman, 2012). Thus, one important standard of evaluation for our disengagement adaptation is to determine whether or not it "breaks" these negative correlations when it is employed with real users (Rotaru and Litman, 2009). A broken correlation would mean that even though disengagement may still occur, it no longer relates to decreased performance.

UNC-DISE_ADAPT responds differently to correct and incorrect DISE turns (Section 3.2). To compare the impacts of these responses both combined and individually, we computed %DISE, %correctDISE (CDISE) and %incorrectDISE (IDISE) for each user (over all five training problems). We then computed bivariate Pearson's correlations within each condition between each DISE metric and both learning and user satisfaction.

Table 5 shows the mean (Mn) and standard deviations (sd) for the DISE metrics within each con-

dition, the coefficient (R) for each correlation, and its significance (p). Consider first task success. The first pair of rows shows that the negative correlation between DISE and learning is still present whether or not the disengagement adaptation is received. However, the second pair of rows shows that the negative correlation between %correctDISE and learning is broken when the disengagement adaptation is received (UNC-DISE), but is still present when not received (UNC). The third pair of rows shows that the disengagement adaptation does not break the negative correlation between %incorrectDISE and learning. Consider next user satisfaction. The first pair of rows shows that the negative correlation between DISE and user satisfaction is broken when the disengagement adaptation is received (UNC-DISE), but is still present when not received (UNC). The third pair of rows shows that the the negative correlation between %incorrectDISE and user satisfaction is also broken when the disengagement adaptation is received (UNC-DISE), but is still present when not received (UNC). These results suggest that for improving task success, adapting to disengagement is more effective for correct turns than incorrect turns[12], while for improving user satisfaction, adapting to disengagement is effective for incorrect turns and for the dialogue as a whole without considering correctness. Finally, Table 5 shows that while %correctDISE is reduced in UNC-DISE as compared to UNC, %incorrectDISE actually increases in UNC-DISE. This suggests that while a reduction in disengagement due to the adaptation partially explains the broken correlations, the adaptation may also ameliorate the negative performance impact of user disengagement.

Table 5: Disengagement-Performance Correlations Across Conditions (Bold Indicates "Broken" Correlation)

	Mn	sd	LGain		UserSat	
			R	p	R	p
%DISE in:						
UNC	17.2	12.1	-.77	.01	-.48	**.04**
UNC-DISE	16.9	7.9	-.65	.01	-.16	**.51**
%CDISE in:						
UNC	7.7	7.6	-.45	**.05**	-.14	.56
UNC-DISE	6.1	3.3	.25	**.31**	-.27	.27
%IDISE in:						
UNC	9.5	7.7	-.76	.01	-.61	**.01**
UNC-DISE	10.8	7.7	-.78	.01	-.05	**.83**

6 Local Affect Transition Analyses

In addition to global performance analyses, the impact of affect adaptation can also be evaluated *locally*, i.e., in terms of its immediate impact in the dialogue. We investigate this local effect by computing the likelihoods of transitioning from each user disengagement state in turn *n* (DISE or ENG) to each user disengagement state in turn *n+1* (DISE or ENG). We use the *transition likelihood L* metric (D'Mello et al., 2007), which has also previously been used by ourselves and others to compute the likelihood of transitioning from one affective state to another in a dialogue corpus and to compare these likelihoods across different system versions (Forbes-Riley and Litman, 2011a; McQuiggan et al., 2008; D'Mello et al., 2007). As in this prior work, we compute the transition likelihoods for each user (over all 5 training dialogues), then use ANOVAs to determine if there were differences in the likelihoods of all possible transitions from the user state in turn *n*.

Transition likelihood L is computed as shown below, where *n* refers to the disengagement state in turn *n* and *n+1* refers to the state in turn *n+1*. As shown, L computes the likelihood that the n→n+1 transition will occur. L=1 indicates that n+1 always follows n, while L=0 and L<0 indicate that the likelihood of transitioning from n to n+1 is equal to chance, and less than chance, respectively.

$$\mathbf{L(n \rightarrow n+1)} = \frac{P(n+1|n) - P(n+1)}{1 - P(n+1)}$$

We hypothesized that users in the UNC-DISE_ADAPT condition would be less likely to transition into disengagement in turn *n+1*. Mean L values across users for each transition are shown in Table 6 for the two conditions, where the rows represent each turn *n* state and the columns represent each turn *n+1* state. The p-value from the ANOVA for each transition likelihood comparison is also shown. The table shows that in both conditions, an engaged

[12]Users who are more often correct may also be predisposed to learn more. This may explain why %correctDISE has a lesser negative impact on learning than %DISE and %incorrectDISE in UNC and UNC-DISE. However, only the disengagement adaptation can explain why %correctDISE has a lesser negative impact on learning in UNC-DISE than in UNC.

user in turn *n* is significantly more likely to remain engaged in turn *n+1* than s/he is to become disengaged. However, in UNC_ADAPT, a disengaged user is more likely (as a trend, p=.06) to remain disengaged than to become engaged in turn *n+1*. In contrast, in UNC-DISE_ADAPT, a disengaged user is equally likely (p=.14) to become disengaged or remain engaged in turn *n+1*. This analysis thus indicates that the disengagement adaptation also has a benefit at the local performance level, in that it reduces the likelihood of continued disengagement.

Table 6: Mean L Values for Disengagement State Transitions

Condition	Turn n	Turn n+1		
		ENG	DISE	p
UNC-DISE	ENG	.06	-.01	.04
	DISE	-.35	.06	.14
UNC	ENG	.09	-.03	.01
	DISE	-.41	.09	.06

7 Summary and Current Directions

We investigated how iteratively adding new affect adaptation to an affect-adaptive spoken dialogue system impacts global and local performance. We presented a disengagement adaptation that can generalize across domains, and discussed its incorporation into our uncertainty-adaptive computer tutor. We then presented a controlled evaluation comparing these multiply and singly adaptive systems. Our results showed that while the disengagement adaptation did not increase (or decrease) task success or user satisfaction, it demonstrated a slight but significant increase in motivation gain for users with high disengagement. Future analyses will shed further light on how disengagement mediates the effect of condition on motivation. The adaptation also reduced user uncertainty and increased correctness for uncertain answers when repeated in the test dialogue, but increased disengagement for repeated answers that were originally certain and incorrect. It also broke negative correlations between disengaged turns and performance, when measured both as task success and user satisfaction, and showed a trend to reduce disengagement at the local dialogue level.

Our next step is to repeat the experiment with fully automated versions of our affect-adaptive spoken dialogue systems, to determine the impact of adding new affect adaptation when the system performs the affect detection and natural language understanding tasks. We are currently in the last stages of building an automatic disengagement detector that will then be implemented in UNC-DISE_ITSPOKE. Interestingly, our prior work suggests that the fully automated UNC-DISE_ADAPT system may yield greater global performance improvements relative to UNC_ADAPT (Forbes-Riley and Litman, 2012) than the wizard-of-oz version of the system; it may be that users are more responsive to the disengagement adaptation when the affect detection and natural language understanding outputs are "noisier". Future work will also consider other experimental designs to help determine the separate and joint effects of the two affect adaptations.

Acknowledgments

This work is funded by NSF award 0914615. We thank Scott Silliman for experimental support.

References

J. C. Acosta and N. G. Ward. 2011. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication*, 53(9-10):1137–1148.

G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard. 2002. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proc. IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 483–492, Washington, DC.

I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Merheranian, D. Fisher, A. Barto, S. Mahadevan, and B. Woolf. 2007. Repairing disengagement with non-invasive interventions. In *Proc. Artificial Intelligence in Education (AIED)*, pages 195–202.

R. S. Baker, A. Corbett, K. Koedinger, S. Evenson, I. Roll, A. Wagner, M. Naim, J. Raspat, D. Baker, and J. Beck. 2006. Adapting to when students game an intelligent tutoring system. In *Proceedings Intelligent Tutoring Systems*, pages 392–401.

A. Black and P. Taylor. 1997. Festival speech synthesis system: system documentation (1.1.1). The Centre for Speech Technology Research, University of Edinburgh, http://www.cstr.ed.ac.uk/projects/festival/.

D. Bohus and E. Horvitz. 2009. Models for multiparty engagement in open-world dialog. In *Proceedings of SIGdial*, pages 225–234, London, UK.

C. Conati and H. Maclaren. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3):267–303.

R. Cowie and R. R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32.

S. D'Mello, R. S. Taylor, and A. Graesser. 2007. Monitoring affective trajectories during complex learning. In *Proc. Cognitive Science Society*, pages 203–208.

S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser. 2010. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *Proc. Intelligent Tutoring Systems Conference*, pages 245–254, June.

M. Dzikovska, J. Moore, N. Steinhauser, and G. Campbell. 2011. Exploring user satisfaction in a tutorial dialogue system. In *Proc. SIGDIAL*, pages 162–172, Portland, Oregon, June.

K. Forbes-Riley and D. Litman. 2011a. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9–10):1115–1136.

K. Forbes-Riley and D. Litman. 2011b. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language (CSL)*, 25(1):105–126.

K. Forbes-Riley and D. Litman. 2011c. When does disengagement correlate with learning in spoken dialog computer tutoring? In *Proceedings of AIED*, Auckland, NZ, June.

K. Forbes-Riley and D. Litman. 2012. Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proc. NAACL-HLT*, Montreal, June.

S. Janarthanam and O. Lemon. 2008. User simulations for online adaptation and knowledge-alignment in troubleshooting dialogue systems. In *Proc. SEMdial*.

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

J. Klein, Y. Moon, and R. Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14:119–140.

K. Liu and R. W. Picard. 2005. Embedded empathy in continuous, interactive health assessment. In *CHI Workshop on HCI Challenges in Health Assessment*.

S. W. McQuiggan, J. L Robison, and J. C. Lester. 2008. Affective transitions in narrative-centered learning environments. In *Proc. Intelligent Tutoring Systems Conference*, pages 490–499.

T. Paek and Y.-C. Ju. 2008. Accommodating explicit user expressions of uncertainty in voice search or something like that. In *Proceedings Interspeech)*, pages 1165–1168, Brisbane, Australia, September.

P. Pintrich and E. DeGroot. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1):33–40.

H. Pon-Barry and S. Shieber. 2011. Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*.

I. Roll. 2009. *Structured Invention Tasks to Prepare Students for Future Learning: Means, Mechanisms, and Cognitive Processes*. Ph.D. thesis, Carnegie Mellon University.

M. Rotaru and D. Litman. 2009. Discourse structure and performance analysis: Beyond the correlation. In *Proc. SIGDIAL*, pages 178–187, London, UK.

B. Schuller, S. Steidl, and A. Batliner. 2009. The Interspeech 2009 emotion challenge. In *Proc. Interspeech*, pages 312–315, ISCA, Brighton, UK, September.

B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan. 2010. The Interspeech 2010 paralinguistic challenge. In *Proc. Interspeech*, pages 2794–2797, Chiba, Japan, September.

W. Tsukahara and N. Ward. 2001. Responding to subtle, fleeting changes in the user's internal state. In *Proceedings of the SIG-CHI on Human factors in computing systems*, pages 77–84, Seattle, WA. ACM.

K. VanLehn, P. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems Conference*, pages 158–167.

J. Walonoski and N. Heffernan. 2006. Prevention of off-task gaming behavior in intelligent tutoring systems. In *Proc. Intelligent Tutoring Systems Conference (ITS)*, pages 722–724.

W. Wang and J. Hirschberg. 2011. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proc. SIGDIAL*, pages 152–161, Portland, Oregon, June.

N. Wang, W.L. Johnson, R. E. Mayer, P. Rizzo, E. Shaw, and H. Collins. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*, 66(2):98–112.

A. Ward. 2010. *Reflection and Learning Robustness in a Natural Language Conceptual Physics Tutoring System*. Ph.D. thesis, University of Pittsburgh.

APPENDIX

UNC_ADAPT ITSPOKE: What are the forces exerted on the man after he releases his keys?

USER: gravity??? (COR, UNC, ENG)

UNC_ADAPT ITSPOKE: Right. *[Correctness Feedback]* Gravity is the only force acting on the man. This force is also called the weight of the man. It's always present when an object is in earth's gravitational field. *[Bottom Out]*. So what's the direction of the force of gravity on the man (e.g., vertically up)? *[Next Topic Question]*

Figure 1: Annotated Dialogue Example Showing Uncertainty Adaptation

UNC-DISE_ADAPT ITSPOKE-1: Let's begin by looking at the motion of the man and his keys while he's holding them. How does his velocity compare to that of his keys?

USER-1: vertical (INC, UNC, DISE)

UNC-DISE_ADAPT ITSPOKE-2: That doesn't sound right. Let's think about this a little more. *[Productive Interaction Feedback]* Since the man is holding his keys, they aren't moving relative to each other. So their velocities must be WHAT? *[Supplementary Question]*.

Figure 2: Dialogue Example Showing Adaptation for Disengaged+Incorrect Turns

UNC-DISE_ADAPT ITSPOKE-1: Let's begin by looking at the motion of the man and his keys while he's holding them. How does his velocity compare to that of his keys?

USER-1: same same same (COR, CER, DISE)

. . . . . .

UNC-DISE_ADAPT ITSPOKE-2a: Well done. Here's a quick progress update. Good effort so far!![*Productive Interaction Feedback When Progress Report Shows Improvement]* Now let's see what happens after the man releases his keys. [...] *[Next Topic Question]*

. . . . . .

UNC-DISE_ADAPT ITSPOKE-2b: Right. Here's a quick progress update. It might help to remember we will build on the topics we're discussing now.*[Productive Interaction Feedback When Progress Report Shows Decline]* Now let's see what happens after the man releases his keys. [...] *[Next Topic Question]*

Figure 3: Dialogue Example Showing Adaptation for Disengaged+Correct Users

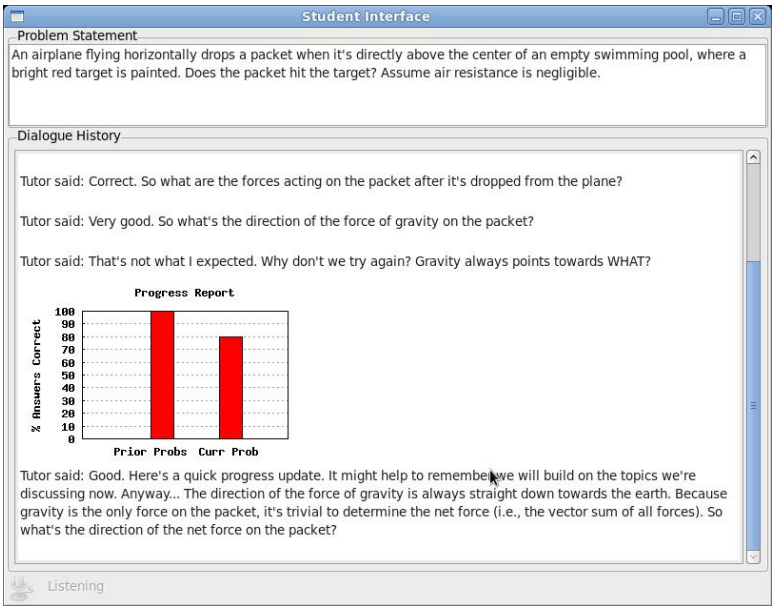

Figure 4: Example Progress Report after Disengaged+Correct Turn