Evaluating a Spoken Dialogue System that Detects and Adapts to User Affective States

Diane Litman
Computer Science Dept. & LRDC
University of Pittsburgh
Pittsburgh, PA 15260 USA
dlitman@pitt.edu

Kate Forbes-Riley
Learning Research & Development Center
University of Pittsburgh
Pittsburgh, PA 15260 USA
forbesk@pitt.edu

Abstract

We present an evaluation of a spoken dialogue system that detects and adapts to user disengagement and uncertainty in real-time. We compare this version of our system to a version that adapts to only user disengagement, and to a version that ignores user disengagement and uncertainty entirely. We find a significant increase in task success when comparing both affect-adaptive versions of our system to our non-adaptive baseline, but only for male users.

1 Introduction

There is increasing interest in building dialogue systems that can detect and adapt to user affective states.[1] However, while this line of research is promising, there is still much work to be done. For example, most research has focused on detecting user affective states, rather than on developing dialogue strategies that adapt to such states once detected. In addition, when affect-adaptive dialogue systems have been developed, most systems detect and adapt to only a single user state, and typically assume that the same affect-adaptive strategy will be equally effective for all users.

In this paper we take a step towards examining these issues, by presenting an evaluation of three versions of an affect-adaptive spoken tutorial dialogue system: one that detects and adapts to both user disengagement and uncertainty, one that adapts to only disengagement, and one that doesn't adapt to affect at all. Our evaluation examines the impact of adapting to differing numbers of affective states on task success, and also examines interactions with user gender. We target disengagement and uncertainty because these were the most frequent affective states in prior studies with our system and their presence was negatively correlated with task success[2] (Forbes-Riley and Litman, 2011; Forbes-Riley and Litman, 2012). The detection of these and similar states is also of interest to the larger speech and language processing communities, e.g. (Wang and Hirschberg, 2011; Bohus and Horvitz, 2009; Pon-Barry and Shieber, 2011). Our results suggest that while adapting to affect increases task success compared to not adapting at all, the utility of our current methods varies with user gender. Also, we find no difference between adapting to one or two states.

2 Related Work

2.1 Adapting to Multiple Affective States

While prior research has shown that users display a range of affective states during spoken dialogue (e.g. (Schuller et al., 2009)), only a few dialogue systems have been developed that can adapt to more than one user affective state (e.g., (D'Mello et al., 2010; Acosta and Ward, 2011)). Furthermore, prior evaluations have compared adapting to at least one affective state to not adapting to affect at all, but have not examined the benefits of adapting to one versus multiple affective states.

In a first evaluation comparing singly and multiply affect-adaptive dialogue systems, we compared an existing system that adapted to uncertainty to a new version that also adapted to disengagement (Forbes-Riley and Litman, 2012). The multiply-adaptive system increased motivation for users with high disengagement, and reduced both uncertainty and the likelihood of continued disengagement. However, this evaluation was only conducted in a "Wizard-of-Oz" scenario, where a hidden human replaced the speech recognition, semantic analysis, and affect detection components of our dialogue system. We also conducted a post-

[1] We use the term *affect* to describe emotions and attitudes that impact how people communicate. Other researchers also combine concepts of emotion, arousal, and attitudes where emotion is not full-blown (Cowie and Cornelius, 2003).

[2] Our success measure is learning gain (Section 4).

Proceedings of the SIGDIAL 2014 Conference, pages 181–185,
Philadelphia, U.S.A., 18-20 June 2014.

hoc correlational (rather than causal) study, using data from an earlier fully-automated version of the uncertainty-adaptive system. Regressions demonstrated that using both automatically labeled disengagement and uncertainty to predict task success significantly outperformed using only disengagement (Forbes-Riley et al., 2012). However, if manual labels were instead used, only disengagement was predictive of learning, and adding uncertainty didn't help. This suggests that detecting multiple affective states might compensate for the noise that is introduced in a fully-automated system. In this paper we further investigate this hypothesis, by evaluating the utility of adapting to zero, one, or two affective states in a controlled experiment involving fully-automated systems.

2.2 Gender Effects in Dialogue

Differences in dialogue structure have been found between male and female students talking to a human tutor (Boyer et al., 2007). Studies have also shown gender differences in conversational entrainment patterns, for acoustic-prosodic features in human-human dialogues (Levitan et al., 2012) and articles in movie conversations (Danescu-Niculescu-Mizil and Lee, 2011). For dialogue systems involving embodied conversational agents, gender effects have been found for facial displays, with females preferring more expressive agents (Foster and Oberlander, 2006). When used for tutoring, females report more positive affect when a learning companion is used, while males are more negative (Woolf et al., 2010).

In our own prior work, we compared two uncertainty-adaptive and one non-adaptive versions of a wizarded dialogue system. Our results demonstrated that only one method of adapting to user uncertainty increased task success, and only for female users (Forbes-Riley and Litman, 2009). In this paper we extend this line of research, by adding an affective dialogue system that adapts to two rather than just one user state to our evaluation, and by moving from wizarded to fully-automated systems.

3 System, Experiment and Corpus

Our corpus consists of dialogues between users and three different versions of ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialog system) (Forbes-Riley and Litman, 2011; Forbes-Riley and Litman, 2012). ITSPOKE is a speech-enhanced and otherwise modified version of the Why2-Atlas text-based qualitative physics tutor (VanLehn et al., 2002) that interacts with users using a system initiative dialogue strategy. User speech is first digitized from head-mounted microphone input and sent to the PocketSphinx recognizer.[3] The recognition output is then classified as (in)correct with respect to the anticipated physics content via semantic analysis (Jordan et al., 2007). Simultaneously, user uncertainty (UNC) and disengagement (DISE) are classified from prosodic, lexical and contextual features using two binary classification models (Forbes-Riley et al., 2012). All statistical components of the speech recognizer, the semantic analyzer, and the uncertainty and disengagement detectors were trained using prior ITSPOKE corpora.[4] Finally, ITSPOKE's response is determined based on the answer's automatically labeled (in)correctness, (un)certainty, and (dis)engagement and then sent to the Cepstral text-to-speech system,[5] as well as displayed on a web-based interface.

Our corpus was collected in an experiment consisting of three conditions (CONTROL, DISE, DISE+UNC), where ITSPOKE used a different method of affect-adaptation in each condition. The experiment was designed to compare the effectiveness of not adapting to user affect in ITSPOKE (CONTROL), adapting to user disengagement (DISE), and adapting to user disengagement as well as user uncertainty (DISE+UNC).[6]

In CONTROL, ITSPOKE's responses to user utterances were based on only the correctness of user answers. This version of the system thus ignored any automatically detected user disengagement or uncertainty. In particular, after each correct answer, ITSPOKE provided positive feedback then moved on to the next topic. After incorrect answers, ITSPOKE instead provided negative

[3] http://www.speech.cs.cmu.edu/pocketsphinx

[4] We have not yet performed the manual annotations needed to evaluate our current versions of these components in isolation. However, earlier versions of our affect detectors yielded FMeasures of 69% and 68% for disengagement and uncertainty, respectively, on par with the best performing affect detectors in the wider literature (Forbes-Riley and Litman, 2011; Forbes-Riley et al., 2012).

[5] http://www.cepstral.com

[6] We did not include an uncertainty-only condition (UNC) because in previous work we compared UNC versus CONTROL (Forbes-Riley and Litman, 2011) and DISE+UNC versus UNC (Forbes-Riley and Litman, 2012). Further details and motivation for all experimental conditions can be found in the description of our earlier Wizard-of-Oz experiment (Forbes-Riley and Litman, 2012).

feedback, then provided remediation tutoring before moving on to the next topic.

In DISE, two adaptive responses were developed to allow ITSPOKE's responses to consider user disengagement as well as the correctness of the user's answer;[7] however, this system version still ignored user uncertainty. In particular, after each disengaged+correct answer, ITSPOKE provided correctness feedback, a progress chart showing user correctness on prior problems and the current problem, and a brief re-engagement tip. After each disengaged+incorrect answer, ITSPOKE provided incorrectness feedback, a brief re-engagement tip, and an easier supplemental exercise, which consisted of an easy fill-in-the-blank type question to reengage the user, followed by remediation targeting the material on which the user disengaged and answered incorrectly. Examples of both types of adaptive responses are shown in A.1 and A.2 of Appendix A, respectively.

In DISE+UNC, ITSPOKE responded to disengagement as just described, but also adapted to uncertainty. In particular, after each uncertain+correct answer, ITSPOKE provided positive correctness feedback, but then added the remediation designed for incorrect answers with the goal of reducing the user's uncertainty. A dialogue excerpt illustrating this strategy is shown in A.3 of Appendix A. Note that when a single utterance is predicted to be both disengaged and uncertain, the DISE and UNC adaptations are combined.

Finally, our experimental procedure was as follows. College students who were native English speakers and who had no college-level physics read a short physics text, took a pretest, worked 5 physics problems (one problem per dialogue) with the version of ITSPOKE from their experimental condition, and took a posttest isomorphic to the pretest. The pretest and posttest were taken from our Wizard-of-Oz experiment and each contained 26 multiple choice physics questions. Our experiment yielded a corpus of 335 dialogues (5 per user) from 67 users (39 female and 28 male). Average pretest[8] and posttest scores were 50.4% and 74.7% (out of 100%), respectively.

4 Performance Analysis

Based on the prior research discussed in Section 2, we had two experimental hypotheses:

[7] Engaged answers were treated as in CONTROL.

[8] Pretest did not differ across conditions (p = .92).

Condition	**Learning Gain**		**N**
	Mean (%)	*Std Err*	
DISE+UNC	53.2	5.0	23
DISE	51.4	4.8	22
CONTROL	46.6	4.7	22
Gender	**Learning Gain**		**N**
Male	53.2	4.3	28
Female	47.6	3.6	39

Table 1: No effect of experimental condition (p=.62) or gender (p=.32) on learning gain.

Gender	**Condition**	**Learning Gain**		**N**
		Mn (%)	*Std Err*	
Male	DISE+UNC	58.8	8.4	7
	DISE	62.2	7.0	10
	CONTROL	38.7	6.7	11
Female	DISE+UNC	47.5	5.6	16
	DISE	40.6	6.4	12
	CONTROL	54.6	6.7	11

Table 2: Significant interaction between the effects of gender and condition on learning (p=.02).

H1: Responding to multiple affective states will yield greater task success than responding to only a single state (DISE+UNC $>$ DISE), which in turn will outperform not responding to affect at all (DISE $>$ CONTROL).

H2: The effects of ITSPOKE's affect-adaptation method and of gender will interact.

A two-way analysis of variance (ANOVA) was thus conducted to examine the effect of experimental condition (DISE+UNC, DISE, CONTROL) and user gender (Male, Female) on task success. As is typical in the tutoring domain, task success was computed as (normalized) learning gain: $\frac{posttest-pretest}{100-pretest}$.

Table 1 shows that although our results patterned as hypothesized when considering all users, the differences in learning gains were not statistically different across experimental conditions, F (2, 61) = .487, p = .617. There were also no main effects of gender, F (1, 61) = 1.014, p = .318.

In contrast, as shown in Table 2, there was a statistically significant interaction between the effects of user gender and experimental condition on learning gains, F (2, 61) = 4.141, p = .021. We thus tested the simple effects of condition within each level of gender to yield further insights.

For males, simple main effects analysis showed

that there were statistically significant differences in learning gains between experimental conditions (p = .042). In particular, males in the DISE condition had significantly higher learning gains than males in the CONTROL condition (p = .019). Males in the DISE+UNC condition also showed a trend for higher learning gains than males in the CONTROL condition (p = .066). However, males in the DISE and DISE+UNC conditions showed no difference in learning gains (p= .760).

For females, in contrast, simple main effects analysis showed no statistically significant differences in learning gains between any experimental conditions (p = .327).

In sum, hypothesis H1 regarding the utility of affect adaptations was only partially supported by our results, where (DISE+UNC = DISE) > CONTROL, and only for males. That is, adapting to affect was indeed better than not adapting at all, but only for males (supporting hypothesis H2). Contrary to H1, adapting to uncertainty over and above disengagement did not provide any benefit compared to adapting to disengagement alone (DISE+UNC = DISE), for both genders.

5 Discussion and Future Directions

Our results contribute to the increasing body of literature demonstrating the utility of adding fully-automated affect-adaptation to existing spoken dialogue systems. In particular, males in our two affect-adaptive conditions (DISE+UNC and DISE) learned more than males in the non-adaptive CONTROL. While our prior work demonstrated the benefits of adapting to uncertainty, the current results demonstrate the importance of adapting to disengagement either alone or in conjunction with uncertainty. However, we also predicted that DISE+UNC should outperform DISE, which was not the case. In future work we will examine other performance measures besides learning, and will manually annotate true disengagement and uncertainty in order to group students by amount of disengagement. Furthermore, since the motivating prior studies discussed in Section 2 were based on older versions of our system, annotation could identify problematic differences in training and testing data. A final potential issue is that the re-engagement tips do not convey exactly the same information.

Second, our results contribute to the literature suggesting that gender effects should be considered when designing dialogue systems. We see similar results as in our prior work; namely our current results continue to suggest that males don't benefit from adapting to their uncertainty as compared to ignoring it, but our current results also suggest that males do benefit from adapting to their disengagement. On the other hand, our current results suggest that females do not benefit from our disengagement adaptation and moreover, combining it with our uncertainty adaptation reduces the benefit of the uncertainty adaptation for them. This suggests the possibility of a differing affective hierarchy, in terms of how affective states may impact the learning process of the two genders differently. Our results yield an empirical basis for future investigations into whether adaptive system performance can improve by adapting to affect differently based on gender. However, further research is needed to determine more effective combinations of disengagement and uncertainty adaptations for both males and females, and to investigate whether gender differences might be related to other types of measurable user factors.

Acknowledgments

This work is funded by NSF 0914615. We thank S. Silliman for experimental support, and H. Nguyen, W. Xiong, and the reviewers for feedback.

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.

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

K. Boyer, M. Vouk, and J. Lester. 2007. The influence of learner characteristics on task-oriented tutorial dialogue. *Frontiers in Artificial Intelligence and Applications*, 158:365.

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

C. Danescu-Niculescu-Mizil and L. Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

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 *Intelligent Tutoring Systems Conference*, pages 245–254, Pittsburgh, PA, USA.

K. Forbes-Riley and D. Litman. 2009. A user modeling-based performance analysis of a wizarded uncertainty-adaptive dialogue system corpus. In *Proceedings Interspeech*, pages 2467–2470.

K. Forbes-Riley and D. Litman. 2011. 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. 2012. Adapting to multiple affective states in spoken dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '12, pages 217–226.

K. Forbes-Riley, D. Litman, H. Friedberg, and J. Drummond. 2012. Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 91–102.

M. Foster and J. Oberlander. 2006. Data-driven generation of emphatic facial displays. *Proceedings of EACL*, 2006:353–363.

P. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C.P. Rose. 2007. Tools for authoring a dialogue agent that participates in learning studies. In *Proc. Artificial Intelligence in Education (AIED)*, pages 43–50.

R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19.

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

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

K. VanLehn, P. W. 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. Intl. Conf. on Intelligent Tutoring Systems*.

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.

B. Woolf, I. Arroyo, D. Cooper, W. Burleson, and K. Muldner. 2010. Affective tutors: Automatic detection of and response to student emotion. In *Advances in Intelligent Tutoring Systems*, pages 207–227. Springer.

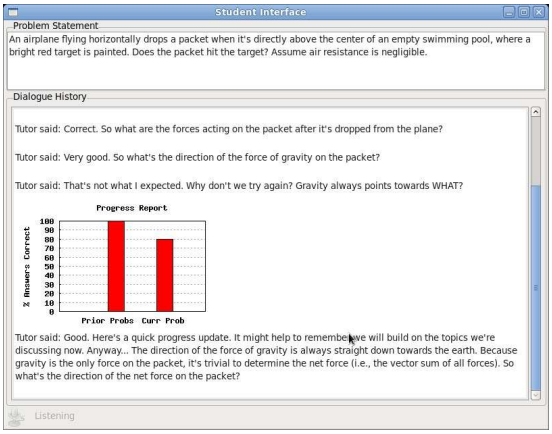
Appendix A. Affect-Adaptive Examples.

A.1 After a Predicted Disengaged+Correct Turn

ITSPOKE: How does his velocity compare to that of his keys?

USER: same same same

ITSPOKE: Right. *[Feedback]* Here's a quick progress update. *[See Screenshot]* It might help to remember we will build on the topics we're discussing now. *[Re-engagement Tip]* Now let's see what happens after the man releases his keys. *[Next Topic]*



A.2 After Predicting Disengaged+Incorrect

ITSPOKE: How does his velocity compare to that of his keys?

USER: vertical

ITSPOKE: That doesn't sound right. *[Feedback]* Let's think about this a little more. *[Re-engagement Tip]* Since the man is holding his keys, they aren't moving relative to each other. So their velocities must be WHAT? *[Supplemental Exercise]*.

A.3 After a Predicted Uncertain+Correct Turn

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

USER: gravity???

ITSPOKE: Right. *[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. *[Remediation]*. So what's the direction of the force of gravity on the man ? *[Next Topic]*