Intrinsic and Extrinsic Evaluation of an Automatic User Disengagement Detector for an Uncertainty-Adaptive Spoken Dialogue System

Kate Forbes-Riley and **Diane Litman** and **Heather Friedberg** and **Joanna Drummond***
University of Pittsburgh
Pittsburgh, PA 15260, USA
forbesk@pitt.edu, litman@pitt.edu, haf13@pitt.edu

Abstract

We present a model for detecting user disengagement during spoken dialogue interactions. Intrinsic evaluation of our model (i.e., with respect to a gold standard) yields results on par with prior work. However, since our goal is immediate implementation in a system that already detects and adapts to user uncertainty, we go further than prior work and present an extrinsic evaluation of our model (i.e., with respect to the real-world task). Correlation analyses show crucially that our automatic disengagement labels correlate with system performance in the same way as the gold standard (manual) labels, while regression analyses show that detecting user disengagement adds value over and above detecting only user uncertainty when modeling performance. Our results suggest that automatically detecting and adapting to user disengagement has the potential to significantly improve performance even in the presence of noise, when compared with only adapting to one affective state or ignoring affect entirely.

1 Introduction

Spoken dialogue systems that can detect and adapt to user affect[1] are fast becoming reality (Schuller et al., 2009b; Batliner et al., 2008; Prendinger and Ishizuka, 2005; Vidrascu and Devillers, 2005; Lee and Narayanan, 2005; Shafran et al., 2003). The benefits are clear: affect-adaptive systems have been shown to increase task success (Forbes-Riley and Litman, 2011a; D'Mello et al., 2010; Wang et al., 2008) or improve other system performance metrics such as user satisfaction (Liu and Picard, 2005; Klein et al., 2002). However, to date most affective systems researchers have focused either only on affect detection, or only on detecting and adapting to a single affective state. The next step is thus to develop and evaluate spoken dialogue systems that detect and respond to multiple affective states.

We previously showed that detecting and responding to user *uncertainty* during spoken dialogue computer tutoring significantly improves task success (Forbes-Riley and Litman, 2011a). We are now taking the next step: incorporating automatic detection and adaptation to user *disengagement* as well, with the goal of further improving task success. We targeted user uncertainty and disengagement because manual annotation showed them to be the two most common user affective states in our system and both are negatively correlated with task success (Litman and Forbes-Riley, 2009; Forbes-Riley and Litman, 2011b). Thus, we hypothesize that providing appropriate responses to these states would reduce their frequency, consequently improving task success. Although we address these user states in the tutoring domain, spoken dialogue researchers across domains and applications have investigated the automatic detection of both user uncertainty (e.g. (Drummond and Litman, 2011; Pon-Barry and Shieber, 2011; Paek and Ju, 2008; Alwan et al., 2007)) and user disengagement (e.g., (Schuller

*Now at Univ. Toronto: jdrummond@cs.toronto.edu

[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).

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 91–102,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

et al., 2010; Wang and Hirschberg, 2011; Schuller et al., 2009a)), to improve system performance. The detection of user disengagement in particular has received substantial attention in recent years, due to growing awareness of its potential for negatively impacting commercial applications (Wang and Hirschberg, 2011; Schuller et al., 2009a).

In this paper we present a model for automatically detecting user disengagement during spoken dialogue interactions. Intrinsic evaluation of our model yields results on par with those of prior work. However, we argue that while intrinsic evaluations are necessary, they aren't sufficient when immediate implementation is the goal, because there is no a priori way to know when the model's performance is acceptable to use in a working system. This problem is particularly relevant to affect detection because it is such a difficult task, where no one achieves near-perfect results. We argue that for such tasks some extrinsic evaluation is also necessary, to show that the automatic labels are useful and/or are a reasonable substitute for a gold standard *before* undertaking a labor-intensive and time-consuming evaluation with real users. Here we use correlational analyses to show that our automatic disengagement labels are related to system performance in the same way as the gold standard (manual) labels. We further show through regression analyses that detecting user disengagement adds value over and above detecting only user uncertainty when modeling performance. These results provide strong evidence that enhancing a spoken dialogue system to detect and adapt to multiple affective states (specifically, user disengagement and uncertainty) has the potential to significantly improve performance even in the presence of noise due to automatic detection, when compared with only adapting to one affective state or ignoring affect entirely.

2 Related Work

Our focus in this paper is on first using machine learning to develop a detector of user disengagement for spoken dialogue systems, and then evaluating its usefulness as fully as possible prior to its implementation and deployment with real users.

Disengaged users are highly undesirable in human-computer interaction because they increase the potential for user dissatisfaction and task failure; thus over the past decade there has already been substantial prior work focused on detecting user disengagement and the closely related states of boredom, motivation and lack of interest (e.g., (Schuller et al., 2010; Wang and Hirschberg, 2011; Jeon et al., 2010; Schuller et al., 2009a; Bohus and Horvitz, 2009; Martalo et al., 2008; Porayska-Pomsta et al., 2008; Kapoor and Picard, 2005; Sidner and Lee, 2003; Forbes-Riley and Litman, 2011b)).

Within this work, specific affect definitions vary slightly with the intention of being coherent within the application and domain and being relevant to the specific adaptation goal (Martalo et al., 2008). However, affective systems researchers generally agree that disengaged users show little involvement in the interaction, and often display facial, gestural and linguistic signals such as gaze avoidance, finger tapping, humming, sarcasm, et cetera.

The features used to detect disengagement also vary depending on system domain and application. For example, Sidner & Lee (2003) are interested in modeling more natural and collaborative human-robot interactions during basic conversations. They define an algorithm for the engagement process that involves appropriate eye gaze and turn-taking. Martalo et al. (2008) study how user engagement influences dialogue patterns during interactions with an embodied agent that gives advice about healthy dieting. They model engagement using manually coded dialogue acts based on the SWBDL-DAMSL scheme (Stolcke et al., 2000). Bohus and Horvitz (2009) study systems that attract and engage users for dynamic, multi-party dialogues in open-world settings. They model user intentions to engage the system with cues from facial sensors and the dialogue. Within recent spoken dialogue research, acoustic-prosodic, lexical and contextual features have been found to be effective detectors of disengagement (Schuller et al., 2010; Wang and Hirschberg, 2011; Jeon et al., 2010); we will briefly compare our own results with these in Section 5.

While all of the above-mentioned research has presented intrinsic evaluations of their disengagement modeling efforts that indicate a reasonable degree of accuracy as compared to a gold standard (e.g., manual coding), only a few have yet demonstrated that the model's detected values are useful in practice and/or are a reasonable substitute for the gold standard with respect to some practical objective (e.g., a relationship to performance). In particular, two studies (Bohus and Horvitz, 2009; Schuller et al., 2009a) have gone directly from intrinsic evaluation of (dis)engagement models to performing user studies with the implemented model, thereby bypassing other less expensive and less labor-intensive means of extrinsic evaluation to quantify their model's usefulness–and potentially indicate its need to be further improved–before deployment with real users. Neither study reports statistically significant improvements in system performance as a result of detecting user (dis)engagement.

Finally, while substantial spoken dialogue and affective systems research has shown that users display a range of affective states while interacting with a system (e.g. (Schuller et al., 2009b; Conati and Maclaren, 2009; Batliner et al., 2008; Devillers and Vidrascu, 2006; Lee and Narayanan, 2005; Shafran et al., 2003; Ang et al., 2002)), to date only a few affective systems have been built that detect and adapt to multiple user affective states (e.g., (D'Mello et al., 2010; Aist et al., 2002; Tsukahara and Ward, 2001)), and most of these have been deployed with crucial natural language processing components "wizarded" by a hidden human agent (e.g., who performs speech recognition or affect annotation on the user turns); moreover, none have yet shown significant improvements in system performance as a result of adapting to multiple user affective states.

3 ITSPOKE: Spoken Dialogue Tutor

We develop and evaluate our disengagement detector using a corpus of spoken dialogues from a 2008 controlled experiment evaluating our uncertainty-adaptive spoken dialogue tutoring system, ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialog system) (Forbes-Riley and Litman, 2011a).[2]

ITSPOKE tutors 5 Newtonian physics problems (one per dialogue), using a Tutor Question - Student Answer - Tutor Response format. After each tutor question, the student speech is digitized from head-mounted microphone input and sent to the Sphinx2 recognizer, which yields an automatic transcript (Huang et al., 1993). This answer's (in)correctness is then automatically classified based on this transcript, using the TuTalk semantic analyzer (Jordan et al., 2007), and the answer's (un)certainty is automatically classified by inputting features of the speech signal, the automatic transcript, and the dialogue context into a logistic regression model. We will discuss these features further in Section 5. All natural language processing components were trained using prior ITSPOKE corpora. The appropriate tutor response is determined based on the answer's automatically labeled (in)correctness and (un)certainty and then sent to the Cepstral text-to-speech system[3], whose audio output is played through the student headphones and is also displayed on a web-based interface.

The experimental procedure was as follows: college students with no college-level physics (1) read a short physics text, (2) took a pretest, (3) worked 5 "training" problems with ITSPOKE, where each user received a varying level of uncertainty adaptation based on condition, (4) took a user satisfaction survey, (5) took a posttest isomorphic to the pretest, and (6) worked a "test" problem with ITSPOKE that was isomorphic to the 5th training problem, where no user received any uncertainty adaptation.

The resulting corpus contains 432 dialogues (6 per student) and 7216 turns from 72 students, 47 female and 25 male. All turns are used in the disengagement detection experiments described next. However, only the *training* problem dialogues (360, 5 per student, 6044 student turns) are used for the performance analyses in Sections 6-7, because the final *test* problem was given after the instruments measuring performance (survey and posttest).

Our survey and tests are the same as those used in multiple prior ITSPOKE experiments (c.f., (Forbes-Riley and Litman, 2011a)). The pretest and posttest each contain 26 multiple choice questions querying knowledge of the topics covered in the dialogues. Average pretest and posttest scores in the corpus were 51.0% and 73.1% (out of 100%) with standard deviations of 14.5% and 13.8%, respectively. The user satisfaction survey contains 16 statements rated on a 5-point Likert scale. Average total sur-

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

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

vey score was 60.9 (out of 80), with a standard deviation of 8.5. While the statements themselves are listed elsewhere (Forbes-Riley and Litman, 2009), 9 statements concern the tutoring domain (e.g., The tutor was effective/precise/useful), 7 of which were taken from (Baylor et al., 2003) and 2 of which were created for our system. 3 statements concern user uncertainty levels and were created for our system. 4 statements concern the spoken dialogue interaction (e.g., It was easy to understand the tutor's speech) and were taken from (Walker et al., 2002). Our survey has also been incorporated into other recent work exploring user satisfaction in spoken dialogue computer tutors (Dzikovska et al., 2011). In Section 6 we discuss how user scores on these instruments are used to measure system performance. See (Forbes-Riley and Litman, 2011a) for further details of ITSPOKE and the 2008 experiment.

Following the experiment, the entire corpus was manually labeled for (in)correctness (correct, incorrect), (un)certainty (CER, UNC) and (dis)engagement (ENG, DISE) by one trained annotator. Table 1 shows the distribution of the labeled turns in the 2008 ITSPOKE corpus. In prior ITSPOKE corpora, our annotator displayed interannotator agreement of 0.85 and 0.62 Kappa on correctness and uncertainty, respectively (Forbes-Riley and Litman, 2011a). For the disengagement label, a reliability analysis was performed over several annotation rounds on subsets of the 2008 ITSPOKE corpus by this and a second trained annotator, yielding 0.55 Kappa (this analysis is described in detail elsewhere (Forbes-Riley et al., 2011)). Our Kappas indicate that user uncertainty and disengagement can both be annotated with moderate reliability in our dataset, on par with prior emotion annotation work (c.f., (Pon-Barry and Shieber, 2011)). Note however that the best way to label users' internal affective state(s) is still an open question. Many system researchers (including ourselves) rely on trained labelers (e.g., (Pon-Barry et al., 2006; Porayska-Pomsta et al., 2008)) while others use self-reports (e.g., (Conati and Maclaren, 2009; Gratch et al., 2009; McQuiggan et al., 2008)). Both methods are problematic; for example both can be rendered inaccurate when users mask their true feelings. Two studies that have compared self-reports, peer labelers, trained labelers, and combinations of labelers (Afzal and Robinson, 2011; D'Mello et al., 2008) both illustrate the common finding that human annotators display low to moderate interannotator reliability for affect annotation, and both studies show that trained labelers yield the highest reliability on this task. Despite the lack of high interannotator reliability, responding to affect detected by trained human labels has still been shown to improve system performance (see Section 1).

Table 1: 2008 ITSPOKE Corpus Description (N=7216)

Turn Label	Total	Percent
Disengaged	1170	16.21%
Correct	5330	73.86%
Uncertain	1483	20.55%
Uncertain+Disengaged	373	5.17%

4 Automatically Detecting User Disengagement (DISE) in ITSPOKE

As noted in Section 1, we have developed a user disengagement detector to incorporate into our existing uncertainty-adaptive spoken dialogue system. The result will be a state of the art system that adapts to multiple affective states during the dialogue.

4.1 Binary DISE Label

Our disengagement annotation scheme (Forbes-Riley et al., 2011) was derived from empirical observations in our data but draws on prior work, including work mentioned in Section 2, appraisal theory-based emotion models (e.g., Conati and Maclaren (2009))[4], and prior approaches to annotating disengagement or related states in tutoring (Lehman et al., 2008; Porayska-Pomsta et al., 2008).

Briefly, our **overall Disengagement label (DISE)** is used for turns expressing moderate to strong disengagement towards the interaction, i.e., responses given without much effort or without caring about appropriateness. Responses might also be accompanied by signs of inattention, boredom, or irritation. Clear examples include answers spoken quickly in leaden monotone, with sarcastic or playful tones, or with off-task sounds such as rhythmic tapping or

[4] Appraisal theorists distinguish emotional behaviors from their underlying causes, arguing that emotions result from an evaluation of a context.

electronics usage.[5] Note that our DISE label is defined independently of the tutoring domain and thus should generalize across spoken dialogue systems.

Figure 1 illustrates the DISE, (in)correctness, and (un)certainty labels across 3 tutor/student turn pairs. $\mathbf{U}_1$ is labeled DISE and UNC because the student gave up immediately and with irritation when too much prior knowledge was required. $\mathbf{U}_2$ is labeled DISE and UNC because the student avoided giving a specific numerical value, offering instead a vague (and obviously incorrect) answer. $\mathbf{U}_3$ is labeled DISE and CER because the student sang the correct answer, indicating a lack of interest in the larger purpose of the material being discussed.[6]

$\mathbf{T}_1$: What is the definition of Newton's Second Law?

$\mathbf{U}_1$: I have no idea <*sigh*>. **(DISE,** ***incorrect*****, UNC)**

...

$\mathbf{T}_2$: What's the numerical value of the man's acceleration? Please specify the units too.

$\mathbf{U}_2$: The speed of the elevator. Meters per second. **(DISE,** ***incorrect*****, UNC)**

...

$\mathbf{T}_3$: What are the forces acting on the keys after the man releases them?

$\mathbf{U}_3$: graaa-vi-tyyyyy <*sings the answer*> **(DISE,** ***correct*****, CER)**

Figure 1: Corpus Example Illustrating the User Turn Labels ((Dis)Engagement, (In)Correctness, (Un)Certainty)

4.2 DISE Detection Method

Machine learning classification was done at the turn level using WEKA software[7] and 10-fold cross validation. A J48 decision tree was chosen because of its easily read output and the fact that previous experiments with our data showed little variance between different machine learning algorithms (Drummond and Litman, 2011). We also use a cost matrix, which heavily penalizes classifying a true DISE instance as false, because our class distributions are highly skewed (16.21% DISE turns) and the cost matrix successfully mitigated the skew's effect in our prior work, where the uncertainty distribution is also skewed (20.55% UNC turns) (Drummond and Litman, 2011).

To train our DISE model, we first extracted the set of speech and dialogue features shown in Figure 2 from the user turns in our corpus. As shown, the acoustic-prosodic features represent duration, pausing, pitch, and energy, and were normalized by the first user turn, as well as totaled and averaged over each dialogue. The lexical and dialogue features consist of the current dialogue name (i.e., one of the six physics problems) and turn number, the current ITSPOKE question's name (e.g.,T_3 in Figure 1 has a unique identifier) and depth in the discourse structure (e.g., an ITSPOKE remediation question after an incorrect user answer would be at one greater depth than the prior question), a word occurrence vector for the automatically recognized text of the user turn, an automatic (in)correctness label, and lastly, the number of user turns since the last correct turn ("incorrect runs"). We also included two user-based features, gender and pretest score.

- **Acoustic-Prosodic Features**

  temporal features: turn duration, prior pause duration, turn-internal silence

  fundamental frequency (f0) and energy (RMS) features: maximum, minimum, mean, std. deviation

  running totals and averages for all features

- **Lexical and Dialogue Features**

  dialogue name and turn number

  question name and question depth

  ITSPOKE-recognized lexical items in turn

  ITSPOKE-labeled turn (in)correctness

  incorrect runs

- **User Identifier Features**:

  gender and pretest score

Figure 2: Features Used to Detect Disengagement (DISE) for each User Turn

[5] Affective systems research has found total disengagement rare in laboratory settings (Lehman et al., 2008; Martalo et al., 2008). As in that research, we equate the DISE label with no or low engagement. Since total disengagement is common in real-world unobserved human-computer interactions (deleting unsatisfactory software being an extreme example) it remains an open question as to how well laboratory findings generalize.

[6] Our original scheme distinguished six DISE subtypes that trained annotators distinguished with a reliability of .43 Kappa (Forbes-Riley et al., 2011). However, pilot experiments indicated that our models cannot accurately distinguish them, thus our DISE detector focuses on the DISE label.

[7] http://www.cs.waikato.ac.nz/ml/weka/

Table 2: Results of 10-fold Cross-Validation Experiment with J48 Decision Tree Algorithm Detecting the Binary DISE Label in the 2008 ITSPOKE Corpus (N=7216 user turns)

Algorithm	Accuracy	UA Precision	UA Recall	UA Fmeasure	CC	MLE
Decision Tree	83.1%	68.9%	68.7%	68.8%	0.52	0.25
Majority Label	83.8%	41.9%	50.0%	45.6%	–	0.27

Note that although our feature set was drawn primarily from our prior uncertainty detection experiments (Forbes-Riley and Litman, 2011a; Drummond and Litman, 2011), we have also experimented with other features, including state-of-the-art acoustic-prosodic features used in the last Interspeech Challenges (Schuller et al., 2010; Schuller et al., 2009b) and made freely available in the openSMILE Toolkit (Florian et al., 2010). To date, however, these features have only decreased the cross-validation performance of our models.[8] While some of our features are tutoring-specific, these have similar counterparts in other applications (i.e., answer (in)correctness corresponds to a more general notion of "response appropriateness" in other domains, while pretest score corresponds to the general notion of domain expertise). Moreover, all of our features are fully automatic and available in real-time, so that the model can be directly implemented and deployed. To that end, we now describe the results of our intrinsic and extrinsic evaluations of our DISE model, aimed at determining whether it is ready to be evaluated with real users.

5 Intrinsic Evaluation: Cross-Validation

Table 2 shows the averaged results of the cross-validation with the J48 decision tree algorithm. In addition to accuracy, we use Unweighted Average (UA) Precision[9], Recall, and F-measure because they are the standard measures used to evaluate current affect recognition technology, particularly for unbalanced two-class problems (Schuller et al., 2009b). In addition, we use the cross correlation (CC) measure and mean linear error (MLE) because these metrics were used in recent work for evaluating disengagement (level of interest) detectors for the Interspeech 2010 challenge (Schuller et al., 2010; Wang and Hirschberg, 2011; Jeon et al., 2010)).[10] Note however that the Interspeech 2010 task differs from ours not only in the corpus and features, but also in the learning task: they used regression to detect a continuous level of interest ranging from 0 to 1, while we detect a binary class. Thus comparison between our results and those are only suggestive rather than conclusive.

As shown in Table 2, we also compare our results with those of majority class (ENG) labeling of the same turns. Since (7216-1170)/7216 user turns in the corpus are engaged (recall Table 1), always selecting the majority class (ENG) label for these turns thus yields 83.8% accuracy (with 0% precision and recall for DISE, and 83.8% precision and 100% recall for ENG). While our DISE model does not outperform majority class labeling with respect to accuracy, this is not surprising given the steep skew in class distribution, and our learned model significantly outperforms the baseline with respect to all the other measures ($p<.001$).[11]

Our CC and MLE results are on par with the best results from the state-of-the-art systems competing in the 2010 Interspeech Challenge, where the task was to detect level of interest. In particular, the winner obtained a CC of 0.428 (higher numbers are better) and an MLE of 0.146 (lower numbers are better) (Jeon et al., 2010), while a subsequent study yielded a CC of 0.480 and an MLE of 0.131 on the same corpus (Wang and Hirschberg, 2011). Our results are also on par with the best results of the other prior research on detecting disengagement discussed in Section 2 that detects a small number of disengagement classes and reports accuracy and/or recall and precision. For example, (Martalo et al., 2008) report average precision of 75% and recall

[8] We also tried using our automatic UNC label as a feature in our DISE model, but our results weren't significantly improved.

[9] simply ((Precision(DISE) + Precision(ENG))/2)

[10] Pearson product-moment correlation coefficient (CC) is a measure of the linear dependence that is widely used in regression settings. MLE is a regression performance measure for the mean absolute error between an estimator and the true value.

[11] CC is undefined for majority class labeling.

of 74% (detecting three levels of disengagement), while (Kapoor and Picard, 2005) report an accuracy of 86% for detecting binary (dis)interest.

Our final DISE model was produced by running the J48 algorithm over our entire corpus. The resulting decision tree contains 141 nodes and 75 leaves. Inspection of the tree reveals that all of the feature types in Figure 2 (acoustic-prosodic, lexical/dialogue, user identifier) are used as decision nodes in the tree, although not all variations on these types were used. The upper-level nodes of the tree are usually considered to be more informative features as compared to lower-level nodes, since they are queried for more leaves. The upper level of the DISE model consists entirely of temporal, lexical, pitch and energy features as well as question name and depth and incorrect runs, while features such as gender, turn number, and dialogue name appear only near the leaves, and pretest score and turn (in)correctness don't appear at all. The amount of pausing prior to the start of the user turn is the most important feature for determining disengagement, with pauses shorter than a quarter second being labeled DISE, suggesting that fast answers are a strong signal of disengagement in our system. Users who answer quickly may do so without taking the time to think it through; the more engaged user, in contrast, takes more time to prepare an answer.

Three lexical items from the student turns, "friction", "light", and "greater", are the next most important features in the tree, suggesting that particular concepts and question types can be typically associated with user disengagement in a system. For example, open-ended system questions may lead users to disengage due to frustration from not knowing when their answer is complete. One common case in ITSPOKE involves asking users to name all the forces on an object; some users don't know how many to list, so they start listing random forces, such as "friction." On the other hand, multiple choice questions can also lead users to disengage; they begin with a reasonable chance of being correct and thus don't take the time to think through their answer. One common case in ITSPOKE involves asking users to determine which of two objects has the greater or lesser force, acceleration, and velocity.

While our feature set is highly generalizable to other domains, it is an empirical question as to whether the feature values we found maximally effective for predicting disengagement also generalize to other domains. Intuition is often unreliable, and it has been widely shown in affect prediction that the answer can depend on domain, dataset, and learning algorithm employed. Moreover, there are many types of spoken dialogue systems with different styles and no single type can represent the entire field. That said, it is also important to note that there are lessons to be learned from the features selected for one particular domain, in terms of the take-home message for other domains. For example, the fact that ”prior pause” is selected as a strong signal of disengagement in ITSPOKE dialogues may indicate that the feature itself (regardless of its selected value) could be transferred to different domains, alone or in the demonstrated combinations with the other selected features.

6 Extrinsic Evaluation: Correlation

Next we use extrinsic evaluation to confirm that our final DISE model is both useful and a reasonable substitute for our gold standard manual DISE labels. With respect to showing the utility of detecting DISE, we use a correlational analysis to show that the gold standard (manual) DISE values are significantly predictive of two different measures of system performance.[12] With respect to showing the adequacy of our current level of detection performance for the learned DISE model, we demonstrate that after replacing the manual DISE labels with the automatic DISE labels when running our correlations, the automatic labels are related to performance in the same way as the gold standard labels.

Thus for both our automatically detected DISE labels (auto) and our gold standard DISE labels (manual), we first computed the total number of occurrences for each student, and then computed a bivariate Pearson's correlation between this total and two different metrics of performance: learning gain (LG) and user satisfaction (US). In the tutoring domain, learning is the primary performance metric and as is common in this domain we compute it as normalized learning gain ((posttest score-pretest score)/(1-

[12] Spoken dialogue research has shown that redesigning a system in light of such correlational analysis can indeed yield performance improvements (Rotaru and Litman, 2009).

Table 3: Correlations between Disengagement and both Satisfaction and Learning in ITSPOKE Corpus (N=72 users)

Measure	Mean (SD)	User Satisfaction		Learning Gain	
		R	p	R	p
Total Manual DISE	12.3 (7.3)	-0.25	0.031	-0.35	0.002
Total Automatic DISE	12.6 (7.4)	-0.26	0.029	-0.31	0.009

pretest score)). In spoken dialogue systems, user satisfaction is the primary performance metric and as is common in this domain we compute it by totaling over the user satisfaction survey scores.[13]

Table 3 shows first the mean and standard deviation for the DISE label over all students, the Pearson's Correlation coefficient (R) and its significance (p). As shown, both our manual and automatic DISE labels are significantly related to performance, regardless of whether we measure it as user satisfaction or learning gain.[14] Moreover, in both cases the correlations are nearly identical between the manual and automatic labels. These results indicate that the detected DISE values are a useful substitute for the gold standard, and suggest that redesigning ITSPOKE to recognize and respond to DISE can significantly improve system performance.

7 Extrinsic Evaluation: Affective State Multiple Regression

Because we are adding our disengagement detector to a spoken dialogue system that already detects and adapts to user uncertainty, we argue that it is also necessary to evaluate whether greater performance benefits are likely to be obtained by adapting to a second state. In other words, given how difficult it is to effectively detect and adapt to one user affective state, is performance likely to improve by detecting and adapting to multiple affective states?

To answer this question, we performed a multiple linear regression analysis aimed at quantifying the relative usefulness of the automatically detected disengagement and uncertainty labels when predicting our system performance metrics. We ran four stepwise linear regressions. The first regression predicted learning gain, and gave the model two possible inputs: the total number of automatic DISE labels and UNC labels per user. We then ran the same regression again, this time predicting user satisfaction. For comparison, we ran the same two regressions using the manual DISE and UNC labels.

As the trained regression models in Figure 3 show, when predicting learning gain, selecting both automatically detected affective state metrics as inputs significantly increases the model's predictive power as compared to only selecting one.[15] The (standardized) feature weights indicate relative predictive power in accounting for the variance in learning gain. As shown, both automatic affect metrics have the same weight in the final model. This result suggests that adapting to our automatically detected disengagement and uncertainty labels can further improve learning over and above adapting to uncertainty alone. Although the final model's predictive power is low (R^2=0.15), our interest here is only in investigating whether the two affective states are more useful in combination than in isolation for predicting performance. In similar types of stepwise regressions on prior ITSPOKE corpora, we've shown that more complete models of system performance incorporating many predictors of learning (i.e. affective states in conjunction with other dialogue features) can yield R^2 values of over .5 (Forbes-Riley et al., 2008).[16]

[13] Identical results were obtained by using an average instead of a total, and only slightly weaker results were obtained when normalizing the DISE totals as the percentages of total turns.

[14] We previously found a related correlation between different DISE and learning measures, during the analysis of our DISE annotation scheme (Forbes-Riley and Litman, 2011b). In particular, we showed a significant partial correlation between the percentage of manual DISE labels and posttest controlled for pretest score.

[15] Using the stepwise method, Automatic DISE was the first feature selected, and Automatic UNC the second. However, note that a model consisting of only the Automatic UNC metric also yields significantly worse predictive power than selecting both affective state metrics. Further note that almost identical models were produced using percentages rather than totals.

[16] R^2 is the standard reported metric for linear regressions. However, for consistency with Table 3, note that the two models in Figure 3 yield R values of -.31 and -.38, respectively.

Learning Gain = -.31 * Total Automatic DISE *(R^2=.09, p=.009)*

Learning Gain = -.24 * Total Automatic DISE - .24 * Total Automatic UNC *(R^2=.15, p=.004)*

Figure 3: Performance Model's Predictive Power Increases Significantly with Multiple Affective Features

Interestingly, for the regression models of learning gain that used manual affect metrics, only the DISE metric was selected as an input. This indicates that the automatic affective state labels are useful in combination for predicting performance in a way that is not reflected in their gold standard counterparts. Detecting multiple affective states might thus be one way to compensate for the noise that is introduced in a fully-automated affective spoken dialogue system.

Similarly, only the DISE metric was selected for inclusion in the regression model of user satisfaction, regardless of whether manual or automatic labels were used. A separate correlation analysis showed that user uncertainty is not significantly correlated with user satisfaction in our system, though we previously found that multiple uncertainty-related metrics do significantly correlate with learning (Litman and Forbes-Riley, 2009).

8 Summary and Current Directions

In this paper we used extrinsic evaluations to provide evidence for the utility of a new system design involving the complex task of user affect detection, prior to undertaking an expensive and time-consuming evaluation of an affect-adaptive system with real users. In particular, we first presented a novel model for automatically detecting user disengagement in spoken dialogue systems. We showed through intrinsic evaluations (i.e., cross-validation experiments using gold-standard labels) that the model yields results on par with prior work. We then showed crucially through novel extrinsic evaluation that the resulting automatically detected disengagement labels correlate with two primary performance metrics (user satisfaction and learning) in the same way as gold standard (manual) labels. This suggests that adapting to the automatic disengagement labels has the potential to significantly improve performance even in the presence of noise from the automatic labeling. Finally, further extrinsic analyses using multiple regression suggest that adapting to our automatic disengagement labels can improve learning (though not user satisfaction) over and above the improvement achieved by only adapting to automatically detected user uncertainty.

We have already developed and implemented an adaptation for user disengagement in ITSPOKE. The disengagement adaptation draws on empirical analyses of our data and effective responses to user disengagement presented in prior work (c.f., (Forbes-Riley and Litman, 2011b)), We are currently evaluating our disengagement adaptation in the "ideal" environment of a Wizard of Oz experiment, where user disengagement, uncertainty, and correctness are labeled by a hidden human during user interactions with ITSPOKE.

Based on the evaluations here, we believe our disengagement model is ready for implementation in ITSPOKE. We will then evaluate the resulting spoken dialogue system for detecting and adapting to multiple affective states in an upcoming controlled experiment with real users.

Acknowledgments

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

References

S. Afzal and P. Robinson. 2011. Natural affect data: Collection and annotation. In Sidney D'Mello and Rafael Calvo, editors, *Affect and Learning Technologies*. Springer.

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. Intelligent Tutoring Systems Conference (ITS) Workshop on Empirical Methods for Tutorial Dialogue Systems*, pages 16–28, San Sebastian, Spain.

A. Alwan, Y. Bai, M. Black, L. Caseyz, M. Gerosa, M. Heritagez, M. Iseliy, M. Jonesz, A. Kazemzadeh, S. Lee, S. Narayanan, P. Pricex, J. Tepperman, and

S. Wangy. 2007. A system for technology based assessment of language and literacy in young children: the role of multiple information sources. In *Proceedings of the 9th IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 26–30, Chania, Greece, October.

J. Ang, R. Dhillon, A. Krupski, E.Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In J. H. L. Hansen and B. Pellom, editors, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 2037–2039, Denver, USA.

A. Batliner, S. Steidl, C. Hacker, and E. Noth. 2008. Private emotions vs. social interaction - a data-driven approach towards analysing emotion in speech. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 18:175–206.

A. L. Baylor, J. Ryu, and E. Shen. 2003. The effect of pedagogical agent voice and animation on learning, motivation, and perceived persona. In *Proceedings of the ED-MEDIA Conference*, Honolulu, Hawaii, June.

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*, 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.

L. Devillers and L. Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing (ICSLP*, pages 801–804, Pittsburgh, PA, September.

S. D'Mello, S. Craig, A. Witherspoon, B. McDaniel, and A. Graesser. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 18:45–80.

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, June.

J. Drummond and D. Litman. 2011. Examining the impacts of dialogue content and system automation on affect models in a spoken tutorial dialogue system. In *Proc. 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 312–318, Portland, Oregon, June.

M. Dzikovska, J. Moore, N. Steinhauser, and G. Campbell. 2011. Exploring user satisfaction in a tutorial dialogue system. In *Proc. 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 162–172, Portland, Oregon, June.

E. Florian, M. Wollmer, and B. Schuller. 2010. The Munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, Florence, Italy.

K. Forbes-Riley and D. Litman. 2009. A user modeling-based performance analysis of a wizarded uncertainty-adaptive dialogue system corpus. In *Proc. Interspeech*, Brighton, UK, September.

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. When does disengagement correlate with learning in spoken dialog computer tutoring? In *Proceedings 15th International Conference on Artificial Intelligence in Education (AIED)*, Auckland, NZ, June.

K. Forbes-Riley, M. Rotaru, and D. Litman. 2008. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*, 18(1-2):11–43.

K. Forbes-Riley, D. Litman, and H. Friedberg. 2011. Annotating disengagement for spoken dialogue computer tutoring. In Sidney D'Mello and Rafael Calvo, editors, *Affect and Learning Technologies*. Springer.

Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic. 2009. Assessing the validity of appraisal-based models of emotion. In *Proceedings of ACII*, Amsterdam, Netherlands.

X. D. Huang, F. Alleva, H. W. Hon, M. Y. Hwang, K. F. Lee, and R. Rosenfeld. 1993. The SphinxII speech recognition system: An Overview. *Computer, Speech and Language*.

J. H. Jeon, R. Xia, and Y. Liu. 2010. Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In *INTERSPEECH'10*, pages 2802–2805.

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.

A. Kapoor and R. W. Picard. 2005. Multimodal affect recognition in learning environments. In *13th Annual ACM International Conference on Multimedia*, pages 677–682, Singapore.

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

C. M. Lee and S. Narayanan. 2005. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2), March.

B. Lehman, M. Matthews, S. D'Mello, and N. Person. 2008. What are you feeling? Investigating student affective states during expert human tutoring sessions. In *Intelligent Tutoring Systems Conference (ITS)*, pages 50–59, Montreal, Canada, June.

D. Litman and K. Forbes-Riley. 2009. Spoken tutorial dialogue and the feeling of another's knowing. In *Proceedings 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, London, UK, September.

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

A. Martalo, N. Novielli, and F. de Rosis. 2008. Attitude display in dialogue patterns. In *Proc. AISB 2008 Symposium on Affective Language in Human and Machine*, pages 1–8, Aberdeen, Scotland, April.

S. McQuiggan, B. Mott, and J. Lester. 2008. Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction (UMUAI)*, 18(1-2):81–123, February.

T. Paek and Y.-C. Ju. 2008. Accommodating explicit user expressions of uncertainty in voice search or something like that. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 08)*, pages 1165–1168, Brisbane, Australia, September.

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

H. Pon-Barry, K. Schultz, E. Owen Bratt, B. Clark, and S. Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, 16:171–194.

K. Porayska-Pomsta, M. Mavrikis, and H. Pain. 2008. Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 18:125–173.

H. Prendinger and M. Ishizuka. 2005. The Empathetic Companion: A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence*, 19(3):267–285.

M. Rotaru and D. Litman. 2009. Discourse structure and performance analysis: Beyond the correlation. In *Proceedings 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, London, UK.

B. Schuller, R. Muller, F. Eyben, J. Gast, B. Hrnler, M. Wollmer, G. Rigoll, A. Hthker, and H. Konosu. 2009a. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 27:1760–1774.

B. Schuller, S. Steidl, and A. Batliner. 2009b. The Interspeech 2009 Emotion Challenge. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*, 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 *Proceedings of the 11th Annual Conference of the International Speech Communication Assocation (Interspeech)*, pages 2794–2797, Chiba, Japan, September.

I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 31–36, St. Thomas, US Virgin Islands.

C. Sidner and C. Lee. 2003. An architecture for engagement in collaborative conversations between a robot and a human. Technical Report TR2003-12, MERL.

A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3).

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. 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*.

L. Vidrascu and L. Devillers. 2005. Detection of real-life emotions in dialogs recorded in a call center. In *Proceedings of INTERSPEECH*, Lisbon, Portugal.

M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard. 2002. DARPA communicator: Cross-system results for the 2001 evaluation. In *Proc. ICSLP*.

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. 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (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.