Prosodic Cues to Disengagement and Uncertainty in Physics Tutorial Dialogues

Diane Litman[1,2]*, Heather Friedberg*[1]*, Kate Forbes-Riley*[2]

[1]Computer Science Department, University of Pittsburgh, Pittsburgh, PA, US
[2]Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, US
dlitman@pitt.edu, hfriedberg@gmail.com, forbesk@pitt.edu

Abstract

This paper focuses on the analysis and prediction of student disengagement and uncertainty, using a corpus of dialogues collected with a spoken tutorial dialogue system in the STEM domain of qualitative physics. We first compare and contrast the prosodic characteristics of dialogue turns exhibiting disengagement or not, and those exhibiting uncertainty or not. We then compare the utility of using multiple prosodic features to predict both disengagement and uncertainty.

Index Terms: spoken dialogue systems, educational applications, emotion detection, prosody

1. Introduction

Tutorial *dialogue* systems are being developed for a variety of STEM[1] domains (e.g., biology, computer science, electricity and electronics, physics, and thermodynamics), as one method for closing the performance gap between human and computer tutors. Dialogue is the natural interaction modality for human tutors; however, to date its communicative power has not yet been fully flexed in computer tutors. While most tutorial dialogue systems respond based only on the correctness of a student answer, it has been hypothesized that performance could be improved by also responding to student affective states [1, 2, 3, 4]. There has been considerable research on affect detection in naturally occurring spoken dialogue, but most of this work has focused on states typically seen during customer care and information-seeking applications (e.g. annoyance and frustration [5]). Less work has addressed the detection of affective states more commonly seen during tutoring interactions (e.g. boredom, confusion, delight, flow, frustration, and surprise [2]). In contrast, while research in intelligent tutoring systems has attempted to detect such pedagogically relevant states, only a few computer tutors are spoken tutorial dialogue systems [1, 2, 3, 4]. Thus, the prosodic detection of learner affect has not been well studied to understand whether such cues generalize across domains.

Our research focuses on the speech-based detection of student uncertainty and disengagement during physics tutorial dialogue. Since both states negatively correlate with student learning and user satisfaction in our prior studies [6], we hypothesize that detecting and responding to these states will improve system performance. These states are also of interest to the broader speech community. Uncertainty detection has recently been studied in the context of voice search [7], while disengagement is related to the recent "Level of Interest" Interspeech Challenge [8]. In this paper, we in particular focus on comparing and contrasting the role of prosody in both characterizing and predicting disengaged versus engaged dialogue turns, and uncertain versus certain turns. Our primary interest is comparing model content, rather than optimizing detection performance.

[1]Science, Technology, Engineering and Mathematics

2. System and corpus

Our corpus consists of dialogues between users and ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) [1], a speech-enhanced and otherwise modified version of the Why2-Atlas text-based qualitative physics tutor. Fig. 1 illustrates one of the problems tutored by ITSPOKE (one per dialogue), along with an associated dialogue excerpt (annotations explained in Section 3).

Problem 1: Suppose a man is in a freefalling elevator that has nothing touching it (you should ignore air resistance). The man is holding his keys motionless right in front of his face. He then lets go. He doesn't toss them up or throw them down; he just releases his grip on them. What will be the position of the keys relative to the man's face as time passes?

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?

Student$_1$: same same same (DISE**,** CER)

...

ITSPOKE$_{12}$: What are the forces exerted on the man after he releases his keys?

Student$_{12}$: gravity??? (ENG**,** UNC)

Figure 1: *Dialogue excerpt with binary disengagement (*DISE*/*ENG*) and uncertainty (*UNC*/*CER*) annotations.*

Our corpus was collected in an experiment evaluating the utility of detecting and adapting to student uncertainty

Table 1: *Turn-level annotations (N=7216).*

Student Turn Label	Total	Percent
Disengaged (DISE)	1170	16.21%
Uncertain (UNC)	1483	20.55%
Uncertain+Disengaged	373	5.17%

in ITSPOKE [1]. Subjects were college students with no college-level physics, who were native speakers of English. The corpus contains 432 dialogues (6 per student) and 7216 turns from 72 students, 47 female and 25 male. Average pretest and posttest scores on multiple-choice physics tests that students took before and after their set of ITSPOKE dialogues were 51.0% and 73.1%, respectively, indicating that students improved their physics knowledge after ITSPOKE's tutoring.

In the version of ITSPOKE used to collect our corpus, student speech is digitized from head-mounted microphone input and sent to the Sphinx recognizer. The recognizer's transcript as well as prosodic features (see Section 4) extracted from the speech are then sent to a finite state dialogue platform. The recognition output is classified as incorrect or not via semantic analysis, while student (un)certainty is classified by inputting prosodic, lexical, and contextual features into a logistic regression model. Finally, the tutor response is determined based on the answer's (in)correctness and (un)certainty and then sent to the Cepstral text-to-speech system, as well as displayed on a web-based interface.

3. Annotation of UNC/CER and DISE/ENG

Each student turn in the corpus was manually annotated by one trained annotator. This annotator displayed inter-annotator agreement of 0.62 Kappa for annotating uncertainty in prior ITSPOKE corpora [1], and 0.55 Kappa for annotating disengagement in the current corpus [6]. Our Kappas indicate that uncertainty and disengagement can be annotated with moderate reliability, on par with prior emotion annotation work (c.f., Kappas of .45 for uncertainty [9] and .66 for level of interest [10]). These annotations serve as the dependent variables in Sections 5-6.

Example (dis)engagement and (un)certainty annotations are shown in Fig. 1. Student turns expressing uncertainty or confusion about the physics topic are annotated as uncertain (UNC), while turns expressing moderate or strong disengagement towards the interaction (i.e. responses given without much effort) are annotated as disengaged (DISE). All other turns are annotated as certain (CER) and engaged (ENG), respectively.

Table 1 shows the class distributions of the annotated turns in our corpus. Students are either disengaged and/or uncertain in 32% of their turns ($\frac{1170+1483-373}{7216}$), although they are only in both states 5% of the time.

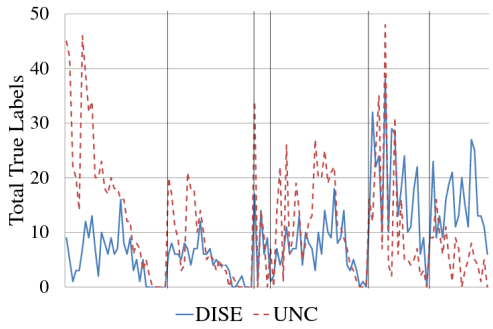

Figure 2: *Annotation distribution over tutoring.*

Fig. 2, which plots the total number of DISE and UNC annotations over the course of all student tutoring, shows that the temporal patterns of the two states differ both within and across the six physics dialogues (separated by the vertical lines). UNC is highest at the beginning of each dialogue then tapers off around the middle, suggesting that students are most uncertain when asked questions about new material. In contrast, DISE is more consistent within each problem but increases as the number of dialogues increases (the mean time to tutor all six problems is 40.6 minutes), suggesting that students disengage as they are forced to stay in the experimental setting longer.

In sum, while Table 1 shows that student uncertainty and disengagement are common in ITSPOKE dialogues, Fig. 2 suggests that different features and models will be needed to best characterize the two states.

4. Extraction of acoustic-prosodic features

The following features were automatically extracted from the speech file for each student turn:

- **Temporal:** turn duration, prior pause duration
- **Pitch (f0):** max, min, mean, std. deviation
- **Energy (RMS):** max, min, mean, std. deviation

Turn duration and prior pause duration (in seconds) were computed via the automatically labeled start and end turn boundaries provided by the speech recognizer during the experiment. F0 and RMS values, representing measures of pitch and loudness, respectively, were computed using openSMILE [11] after the experiment[2].

These automatically computed features serve as the independent variables in Sections 5-6, and are motivated by previous studies of emotion prediction in spontaneous dialogues by ourselves and others (as reviewed in [1]). While we have experimented with adding other features used in recent Interspeech Challenges [8] that can be

[2]During the experiment, Entropic Research Laboratory's pitch tracker was used to detect uncertainty. ITSPOKE's prosodic analysis component (recall Section 2) has now been updated to use openSMILE.

computed in real-time using openSMILE [11], to date this has only decreased the cross-validated performance of the predictive models discussed in Section 6 (perhaps due to the relatively small size of our corpus).

5. Descriptive analysis and results

We first looked for distinguishing prosodic characteristics of DISE versus ENG student turns, and UNC versus CER turns. We hypothesized that prosodic differences would exist for both annotation types, but that the differences between UNC/CER turns could be different than those between DISE/ENG turns. To examine the prosodic differences between uncertain versus certain turns in a speaker independent manner, for each student and for each feature, we calculated the mean value of that student's UNC turns, and the mean value of his/her CER turns. Then, for each feature, we created vectors of these 72 student means for UNC and CER turns and performed paired t-tests on the vectors. A similar analysis was performed to compare DISE and ENG turns.

Table 2: *Comparisons of* ENG *vs.* DISE *and of* CER *vs.* UNC *turns, by acoustic-prosodic features.*

Feature	Mean Diff ENG - DISE	Mean Diff CER - UNC
turn duration	.076	-.032
prior pause	-1.661*	-3.077*
max f0	10.910*	9.971*
min f0	1.152	1.254
mean f0	4.755*	4.907*
stddev f0	2.889*	5.183*
max RMS	.005	.011*
min RMS	<.001*	<.001*
mean RMS	.001	.002*
stddev RMS	.001*	.003*

* denotes significant difference ($p < .05$) in mean values

Table 2 shows the mean difference between labels by annotation type. We first focus on the statistically significant prosodic commonalities in distinguishing UNC/CER turns and in distinguishing DISE/ENG turns. With respect to temporal features, the second row of the table shows that when students are disengaged (DISE), they take significantly longer to respond to the tutor ("prior pause") than when they are engaged (ENG). This row also shows that students take longer to respond when they are uncertain (UNC) about the physics content of their response, compared to their CER responses. With respect to pitch, three features have lower values when student turns are annotated as either DISE or UNC (thus yielding positive differences in the table). Disengaged turns and uncertain turns are uttered with lower maximum ("max f0") and mean ("mean f0") pitch values than engaged turns or certain turns, and pitch is more constant (i.e., "stdev f0" is lower) throughout. Energy is also lower ("min RMS") and more constant ("stdev RMS") in both disengaged and uncertain turns, as compared to their counterparts[3].

In contrast, the two remaining energy features only significantly differ between uncertain versus certain turns. While uncertain turns are spoken more softly than CER turns ("max RMS" and "mean RMS"), disengaged turns are neither softer nor louder than engaged turns.

6. Predictive results

Given the significant differences just identified, we next examine whether our features also have predictive utility, particularly when used in combination with each other. Decision tree models for predicting both (un)certainty and (dis)engagement were trained using the WEKA machine learning toolkit[4]. A decision tree representation was chosen due to our ability to compute feature usage as described below, and the fact that our previous experiments predicting uncertainty showed little variance between learning algorithms [1]. Because of the skewed annotation distributions shown in Table 1 (16.21% DISE, 20.55% UNC), a cost matrix was used during learning to penalize classifying a true DISE or true UNC instance as false. To mitigate issues of speaker-dependence, we normalized values of all features by first turn value[5].

Table 3 summarizes how the acoustic-prosodic features are used to predict uncertainty and disengagement, based on the structure of the learned decision trees. Following [5], feature usage is reported as the percentage of decisions for which the feature type is queried; features higher in the tree thus have higher usage than features lower in the tree[6]. As can be seen, the feature usage of both learned models show a similar pattern with respect to the three broad types of acoustic-prosodic knowledge. Temporal features are most highly queried, followed by pitch, followed by energy.

At the level of specific features, however, there are differences. First, only the disengagement model includes both types of temporal features. Prior pause duration is included in both models, and is in fact the root of both trees; the importance of prior pause is consistent with the descriptive results in Table 2. In contrast, turn duration (which was not discriminative in isolation for either state) is now used by the disengagement model in conjunction with other features.

With respect to pitch and energy, the uncertainty model uses both types of features more heavily than the disengagement model. In addition, their relative utility differs across the models. As with the temporal features,

[3]Additional t-tests determined that of the differences asterisked in both columns of Table 2, the "prior pause," "stdev f0" and "stdev RMS" differences were significantly different across the two columns.

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

[5]We included both the raw and normalized values of "prior pause" because both were found to be useful in prior studies.

[6]The percentages sum to 99 rather than 100 due to rounding.

Table 3: *Feature usage in learned models.*

Features	Uncertainty	Disengagement
Temporal	**50%**	**72%**
turn duration	0%	23%
prior pause	50%	49%
Pitch	**34%**	**16%**
max f0	9%	4%
min f0	19%	4%
mean f0	0%	8%
stdev f0	6%	0%
Energy	**15%**	**11%**
max RMS	0%	0%
min RMS	6%	1%
mean RMS	9%	4%
stdev RMS	0%	6%

the prosodic differences identified in Table 2 sometimes differ from the predictive features identified in Table 3, where multiple features are used in combination. For example, "stdev f0" is a significant difference in Table 2 for both states, but is only used by the uncertainty model. Conversely, "min f0" is not a significant difference in Table 2, but is included in both predictive models.

Finally, we quantitatively evaluated our two learned models using 10-fold cross validation. The unweighted average precision and recall are 63% and 61% for uncertainty, and 61% and 56% for disengagement, respectively[7]. While these results already significantly improve over majority class baselines (unweighted precision/recall of 40%/50% for predicting all turns as certain, and 42%/50% for predicting all turns as engaged), recall that the focus of this paper is only on acoustic-prosodic feature comparisons. The absolute performance results of our predictive models can in fact be increased when lexical, semantic, discourse, and other types of features are included (e.g., 69% unweighted precision and recall for the disengagement model [6], and Kappas of .5 and .4 for disengagement and uncertainty, respectively).

7. Conclusions

Our results indicate that turns annotated as disengaged or uncertain do prosodically differ from turns annotated as engaged or certain, respectively; however, the features on which turns differ do vary slightly according to student state ((dis)engagement, (un)certainty). Disengaged turns have longer response times, lower pitch values, and less pitch and energy variation than engaged turns. Uncertain turns differ from certain turns in these same ways, and in addition are not as loud as certain turns. Our results also indicate the utility of acoustic-prosodic features in predictive models. Again, however, the best combination of features differ across our two binary prediction tasks (disengaged vs. engaged, and uncertain vs. certain). While temporal features are the most important type of knowledge in both predictive models, the temporal features play a more central role when predicting `DISE/ENG` as compared to when predicting `UNC/CER`. Conversely, pitch features are more prominent when predicting `UNC/CER`.

We plan to replicate the prosodic analyses presented here on two publicly available corpora with related annotations ("Level of Interest" [8], uncertainty [9]), to explore whether our findings from tutoring generalize to other types of dialogue systems. We are also now implementing our best predictive models in ITSPOKE (using prosody supplemented with other feature types), to evaluate the utility of detecting and adapting to both uncertainty and disengagement in a controlled experiment.

[7]Unweighted metrics are the standards for evaluating affect recognition, particularly for unbalanced class distributions [8].

8. Acknowledgments

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

9. References

[1] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, no. 9–10, pp. 1115–1136, 2011.

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

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

[4] W. Chen and J. Mostow, "A tale of two tasks: Detecting children's off-task speech in a reading tutor," in *Proc. Interspeech*, Florence, Italy, 2011.

[5] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. ICSLP*, J. H. L. Hansen and B. Pellom, Eds., Denver, USA, 2002, pp. 2037–2039.

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

[7] T. Paek and Y.-C. Ju, "Accommodating explicit user expressions of uncertainty in voice search or something like that," in *Proc. Interspeech*, Brisbane, Australia, September 2008, pp. 1165–1168.

[8] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," in *Proc. Interspeech*, Chiba, Japan, Sept. 2010.

[9] H. Pon-Barry and S. Shieber, "Recognizing uncertainty in speech," *EURASIP Jrnl. on Advances in Signal Processing*, 2011.

[10] B. Schuller, R. Mller, F. Eyben, J. Gast, B. Hrnler, M. Wllmer, G. Rigoll, A. Hthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760 – 1774, 2009.

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