RECOGNIZING EMOTIONS FROM STUDENT SPEECH IN TUTORING DIALOGUES

Diane Litman[*]

University of Pittsburgh
Department of Computer Science
Learning Research and Development Ctr.
Pittsburgh PA, 15260, USA

Kate Forbes

University of Pittsburgh
Learning Research and Development Ctr.
Pittsburgh PA, 15260, USA

ABSTRACT

We investigate the automatic classification of student emotional states in a corpus of human-human spoken tutoring dialogues. We first annotated student turns in this corpus for negative, neutral and positive emotions. We then automatically extracted acoustic and prosodic features from the student speech, and compared the results of a variety of machine learning algorithms that use 8 different feature sets to predict the annotated emotions. Our best results have an accuracy of 80.53% and show 26.28% relative improvement over a baseline. These results suggest that the intelligent tutoring spoken dialogue system we are developing can be enhanced to automatically predict and adapt to student emotional states.

1. INTRODUCTION

In this paper, we investigate the automatic classification of student emotional states in human-human spoken tutoring dialogues. Motivation for this investigation comes from the discrepancy between the performance of human tutors and current machine tutors. While human tutors can respond both to the content of student speech and to the emotions they perceive to be underlying it, most intelligent tutoring dialogue systems cannot detect student emotional states, and furthermore, are text-based [1], which may limit their success at emotion prediction. Building intelligent tutoring *spoken* dialogue systems thus has great potential benefit. Speech is the most natural and easy to use form of natural language interaction, and studies have shown considerable benefits of spoken human tutoring [2]. Furthermore, connections between learning and emotion are well-documented [3], and it has been suggested that the success of computer tutors could be increased by recognizing and responding to student emotion, e.g. reinforcing positive states, while rectifying negative states [4, 5]. We are currently building an intelligent tutoring spoken dialogue system with the goal of enhancing it to automatically predict and adapt to student emotional states. The larger question motivating this paper is thus whether, and how, student emotional states can be automatically predicted by our intelligent spoken dialogue tutoring system.

Speech supplies a rich source of acoustic and prosodic information about the speaker's current emotional state. Research in the area of spoken dialogue systems has already shown that acoustic and prosodic features can be extracted from the speech signal and used to develop predictive models of user mental states (c.f. [6, 7, 8, 9, 10, 11, 12]). Some work uses speech read by actors or native speakers as training data (c.f. [6, 8, 12]), but such prototypical emotional speech does not necessarily reflect naturally-occurring speech [13], such as found in tutoring dialogues. Other work uses naturally-occurring speech from a variety of corpora (c.f. [7, 9, 10, 11]), but little work to date addresses emotion detection in computer-based educational settings such as tutoring.

Our methodology builds on and generalizes the results of this prior work while applying them to the new domain of naturally occurring tutoring dialogues. We first annotated student turns in our human-human tutoring corpus for emotion. We then automatically extracted acoustic and prosodic features from these turns, and performed a variety of machine learning experiments using different feature combinations to predict our emotion categorizations. Like [6], we compare the performance of a variety of modern machine learning algorithms using the Weka software package [14]; we also present a variety of evaluation metrics for our results.

Although much of the past work in this area predicts only two emotional classes (e.g. negative/non-negative) [7, 10, 11], our preliminary experiments produced the best predictions using a three-way distinction between negative, neutral, and positive emotional classes. Like [11], we focus here on features that can be computed fully automatically from the student speech and will be available to our intelligent tutoring dialogue system in real-time. We show that by using these features alone or in combination with features identifying specific subjects and tutoring sessions, we can significantly improve a baseline performance for emotion prediction. Our best results show an accuracy of 80.53% and a relative improvement of 26.28% over the baseline error rate. These results suggest that our spoken dialogue tutoring system can be enhanced to automatically predict and adapt to student emotional states.

Section 2 describes ITSPOKE, our intelligent tutoring spoken dialogue system. Section 3 describes machine learning experiments in automatic emotion recognition. We first discuss our emotion annotation in a human-human spoken tutoring corpus that is parallel to the corpus that will be produced by ITSPOKE. We then discuss how acoustic and prosodic features available in real-time to ITSPOKE are computed from these dialogues. We finally compare the performance of various machine learning algorithms using our annotations and different feature combinations. Section 4 discusses further directions we are pursuing.

2. THE ITSPOKE SYSTEM AND CORPUS

We are developing a spoken dialogue system, called ITSPOKE (*I*ntelligent *T*utoring *SPOKE*n dialogue system), which uses as

[*]This research is supported by the NSF Grant No. 9720359 to the Center for Interdisciplinary Research on Constructive Learning Environments (CIRCLE) at the University of Pittsburgh and Carnegie-Mellon University.

its "back-end" the *text-based* Why2-Atlas dialogue tutoring system [15]. In Why2-Atlas, a student types an essay answering a qualitative physics problem and a computer tutor then engages him/her in typed dialogue to provide feedback, correct misconceptions, and elicit more complete explanations, after which the student revises his/her essay, thereby ending the tutoring or causing another round of tutoring/essay revision. In ITSPOKE, we replace the typed dialogue with spoken input and output. We have interfaced the Sphinx2 speech recognizer [16] with stochastic language models trained from example user utterances, and the Festival speech synthesizer [17] for text-to-speech, to the Why2-Atlas back-end. The rest of the natural language processing components, e.g. the sentence-level syntactic and semantic analysis modules [18], and a finite-state dialogue manager [19], are provided by a toolkit that is part of the Why2-Atlas back-end. The student speech is digitized from microphone input, while the tutor's synthesized speech is played to the student using a speaker and/or headphone. We have adapted the knowledge sources needed by the spoken language components; e.g. we have developed a set of dialogue-dependent language models using 4551 student utterances from a Why2-Atlas 2002 human-computer typed corpus and continue to enhance them using student utterances from our parallel human-human spoken corpus (described below). An evaluation comparing ITSPOKE to human-human and Why2-Atlas (typed) tutoring will begin in Fall, 2003.

Our human-human spoken corpus contains spoken dialogues collected via a web interface supplemented with a high quality audio link, where a human tutor performs the same task as ITSPOKE. Our subjects are University of Pittsburgh students who have taken no college level physics and are native speakers of American English. Our experimental procedure, taking roughly 7 hours/student over 1-2 sessions, is as follows: students 1) take a pretest measuring their physics knowledge, 2) read a small document of background material, 3) use the web and voice interface to work through up to 10 training problems with the human tutor (via essay revision as described above), and 4) take a post-test similar to the pretest. We have to date collected 114 dialogues (2187.17 minutes of speech from 5 females and 8 males) and transcribed 90 of them. An average dialogue contains 47.49 student turns (264.18 student words); a comparison of these and other dialogue features between our human-human spoken corpus and a parallel typed corpus is found in [20]. A corpus example is shown in Figure 1, containing the typed problem, the student's original typed essay, and an annotated (Section 3) excerpt from the subsequent spoken dialogue, beginning with the tutor's sixth turn (some punctuation is added for clarity).

PROBLEM: The sun pulls on the earth with the force of gravity and causes the earth to move in orbit around the sun. Does the earth pull equally on the sun? Defend your answer.

ESSAY: No. Both objects apply a force on the other. The sun applies more force because it is causing the earth to have more acceleration.

. . . dialogue excerpt at 4.3 minutes into session . . .

TUTOR$_6$: The only thing asked is about the force, whether the force of earth pulls equally on the sun or not. That's the only question.

STUDENT$_7$: Well I think it does but I don't know why. I do-don't I- do they move in the same direction? I do-don't
(EMOTION = NEGATIVE)

TUTOR$_7$: You see, against- you see, they don't have to move. If a force acts on a body it does not mean that uh uh I mean it will -

STUDENT$_8$: If two forces um apply- if two forces react on each other then the force is equal. It's the Newtons third law.
(EMOTION = POSITIVE)

TUTOR$_8$: Um you see the uh- actually in this case the motion is there but it is a little more complicated motion. This is orbital motion.

STUDENT$_9$: mm-hm *(EMOTION = NEUTRAL)*

Fig. 1. Annotated Excerpt from Human-Human Spoken Corpus

3. PREDICTING EMOTIONAL SPEECH

3.1. Emotion Inter-Annotation

Human emotional states can only be identified indirectly, e.g. via what is said and/or how it is said. However, such evidence is not always obvious, unambiguous, or consistent across speakers. Our objective is thus to manually annotate the student turns in our human-human tutoring dialogues for *perceived expressions of emotion*. In our annotation schema, expressions of emotion are viewed along a linear scale, shown and defined as follows:

negative $\longleftarrow$ **neutral** $\longrightarrow$ **positive**

Negative: a strong expression of emotion that can be detrimental for learning, e.g. *uncertain, confused, bored, frustrated*. Because a syntactic question by definition expresses uncertainty, a student turn containing only a question is by default labeled negative. An example of a *negative* student turn is **student**$_7$ in Figure 1. Evidence[1] of a negative emotion in this case comes from the lexical expressions of uncertainty, e.g. the phrase "I don't know why", the syntactic question, and the disfluencies, as well as acoustic and prosodic features, including pausing and a wide variation in pitch.

Positive: a strong expression of emotion that can be beneficial for learning, e.g. *confident, interested, engaged, encouraged*. An example of a *positive* student turn is **student**$_8$ in Figure 1. Evidence of a positive emotion in this case comes from lexical expressions of certainty, e.g. "It's the...", as well as acoustic and prosodic features, including loud speech and a fast tempo.

Neutral: no *strong* expression of emotion, including weak (negative or positive) or contrasting (negative and positive) expressions, as well as no expression. Because groundings serve mainly to encourage another speaker to continue speaking, a student turn containing only a grounding is by default labeled neutral. An example of a *neutral* student turn is **student**$_9$ in Figure 1. Neutrality is assigned by default in this case, but acoustic and prosodic features, including moderate loudness, tempo, and inflection, give further evidence for the *neutral* label (rather than over-riding it).

There are a number of differences between this annotation schema and those in prior work. In contrast to [11], for example, our classifications are context-relative (e.g. relative to the other turns in the dialogue), and task-relative (e.g. relative to tutoring), because like [7], we are interested in detecting emotional changes across our tutoring dialogues. For example, if a student has been answering a tutor's questions with apparent ease, but then responds

[1] As determined by *post*-annotation discussion (see below).

to the next question slowly and says "Um, now I'm confused", this turn would likely be labeled "negative". During a heated argument between two speakers, however, this *same* turn would likely be labeled "neutral". Although [10] also employs a relative classification, their schema (see [13]) explicitly associates specific acoustic, prosodic, lexical and dialogue features with emotional utterances. To avoid restricting or otherwise influencing the annotator's intuitive understanding of emotion expression, and because such features are not used consistently or unambiguously across speakers, we allowed annotators to be guided by their intuition rather than a set of expected features. Post-annotation discussion was used to elicit the particular features of each turn that led the annotator to select the chosen label. Finally, as noted in Section 1, [11, 10] annotate only two emotion classes (e.g. emotional/non-emotional), while [7] annotates 6 emotion classes but only uses negative/other in their machine learning experiments. Our annotation and our machine learning experiments employ a three-way distinction between negative, neutral, and positive classes.

For use in our machine learning experiments, we randomly selected 5 transcribed dialogues from our human-human tutoring corpus, totaling 263 student turns from 4 different students (2 male, 2 female). First, turn boundaries were manually annotated (based on consensus labelings from two annotators) when: 1) the speaker stopped speaking and the other party in the dialogue began to speak, 2) the speaker asked a question and stopped speaking to wait for an answer, 3) the other party in the dialogue interrupted the speaker and the speaker paused to allow the other party to speak. Each turn was then manually annotated by two annotators as *negative*, *neutral* or *positive*, as described above. The two annotators agreed on the annotations of 215/263 turns, achieving 81.75% agreement (Kappa = 0.624). This inter-annotator agreement is expected given the difficulty of the task, and is on par with prior studies. [7], for example, achieved inter-annotator agreement of 71% (Kappa 0.47), while [11] averaged around 70% inter-annotator agreement. Although when we conflated our *positive* and *neutral* classes, our own inter-annotator agreement rose to 92.40% (Kappa = 0.761), preliminary machine learning experiments gave better results when learning all three emotion classes.

As in [11], our machine learning experiments use only those 215 student turns where annotators agreed on an emotion label. Of these turns, 158 were *neutral*, 42 were *negative*, and 15 were *positive*. 13 *neutral* turns were removed, however, as they contained only non-speech noise, yielding 202 agreed student turns.

3.2. Extracting Features from the Speech Signal

Like [11], we focus in this paper on acoustic and prosodic features of individual turns that can be computed automatically from the speech signal and that will be available in real-time to ITSPOKE. For each of the 202 agreed student turns, we automatically computed the 33 acoustic and prosodic features itemized in Figure 2, which prior studies cited above have shown to be useful for predicting emotions in other domains. Except for turn duration and tempo, which were calculated via the hand-segmented turn boundaries (but will be computed automatically in ITSPOKE via the output of the speech recognizer), the acoustic and prosodic features were calculated automatically from each turn in isolation. f0 and RMS values, representing measures of pitch excursion and loudness, respectively, were computed using Entropic Research Laboratory's pitch tracker, *get_f0*, with no post-correction. Speaking rate was calculated using the Festival synthesizer OALD dictionary as syllables per second in the turn, and amount of silence was defined as the proportion of zero frames in the turn, e.g. the proportion of time that the student was silent. We also recorded for each turn the 3 "identifier" features shown last in Figure 2. Prior studies have shown that "subject" and "gender" features can play an important role in emotion recognition, because different genders and different speakers can convey emotions differently. "subject ID" and "problem ID" are uniquely important in our tutoring domain, because in contrast to e.g. call centers, where every caller is distinct, students will use our system repeatedly, and problems are repeated across students. Thus such features can be used to recognize individual speaker emotions and/or specific contexts.

- 4 raw fundamental frequency (f0) features: maximum, minimum, mean, standard deviation
- 4 raw energy (RMS) features: maximum, minimum, mean, standard deviation
- 3 raw temporal features: total turn duration, speaking rate, amount of silence in the turn
- The above 11 features normalized to the first turn (n1)
- The above 11 features normalized to the prior turn (n2)
- 3 identifier features: subject ID, subject gender, problem ID

Fig. 2. 36 Features per Student Turn

We then created the 8 feature sets itemized in Figure 3 to study the effects that various feature combinations had on predicting the emotion labels in our data. As shown, we compare feature sets containing only "raw" acoustic and prosodic features with feature sets containing each of the "n1" and "n2" normalized acoustic and prosodic features, and we also compare feature sets containing all the acoustic and prosodic features (raw and normalized). Note further that we compare "...speech" and "...subj" feature sets; these compare how well our emotion data would be learned with only acoustic, prosodic and temporal features (either raw or normalized or both), versus adding in our individualized identifier features.

- rawspeech: 11 raw f0, RMS, and temporal features
- rawsubj: 11 raw f0, RMS, and temporal features + 3 identifier features
- n1speech: 11 n1 f0, RMS and temporal features
- n1subj: 11 n1 f0, RMS and temporal features + 3 identifier features
- n2speech: 11 n2 f0, RMS and temporal features
- n2subj: 11 n2 f0, RMS and temporal features + 3 identifier features
- allspeech: 33 f0, RMS, and temporal features (raw, n1, n2)
- allsubj: all 36 features

Fig. 3. 8 Feature Sets for Machine Learning Experiments

3.3. Using Machine Learning to Predict Emotions

We next performed machine learning experiments with our feature sets and our emotion-annotated data, using the Weka machine learning software [14] as in [6]. This software allows us to com-

pare the predictions of some of the most modern machine learning algorithms with respect to a variety of evaluation metrics.

We selected four machine learning algorithms for comparison. First, we chose a C4.5 decision tree learner, called "J48" in Weka; by default this algorithm produces a pruned decision tree with a 0.25 confidence threshold, where leaves correspond to a predicted emotion label if at least two instances of that path are found in the training data. An advantage of decision tree algorithms is that they allow automatic feature selection and their tree output provides an intuitive way to gain insight into the data. Second, we chose a nearest-neighbor classifier, called "IB1" in Weka; this algorithm uses a distance measure to predict as the class of a test instance the class of the first closest training instance that is found. Also available in Weka is the "IBK" classifier, where the number of nearest neighbors considered (k) can be specified manually or determined automatically using leave-one-out cross-validation. We experimented with alternative values of k but found k=1 to achieve the best results. Third, we chose a "boosting" algorithm, called "AdaBoostM1" in Weka. Boosting algorithms generally enable the accuracy of a "weak" learning algorithm to be improved by repeatedly applying that algorithm to different distributions or weightings of training examples, each time generating a new weak prediction rule, and eventually combining all weak prediction rules into a single prediction [21]. Following [6], we selected "J48" as AdaBoost's weak learning algorithm. Finally, we chose a standard baseline algorithm, called "ZeroR" in Weka; this algorithm simply predicts the majority class ("neutral") in the training data, and thus is used as a performance benchmark for the other algorithms.

Table 1 shows the accuracy (percent correct) of these algorithms on each of our 8 feature sets, as compared to the accuracy of AdaBoost. Significances of accuracy differences are automatically computed in Weka using a two-tailed t-test and a 0.05 confidence threshold across 10 runs of 10-fold cross-validation. "*" indicates that the algorithm performed statistically worse than AdaBoost on that feature set. "v" indicates that an algorithm performed statistically better than AdaBoost on that feature set. Lack of either symbol indicates that there was no statistical difference between the performance of the algorithm as compared to AdaBoost.

Feature Set	AdaBoost	J48	IB1	ZeroR
rawspeech	74.85	71.43 *	75.98	71.80 *
rawsubj	77.62	75.49	80.53 v	71.80 *
n1speech	79.21	75.94 *	70.61 *	71.80 *
n1subj	78.61	74.01 *	72.84 *	71.80 *
n2speech	70.88	72.09	64.99 *	71.80 v
n2subj	72.36	71.11	68.01 *	71.80
allspeech	77.44	75.00 *	74.57	71.80 *
allsubj	77.96	75.01 *	79.21	71.80 *

Table 1. Percent Correct, 0.05 Confidence (two-tailed)

As shown, the single best accuracy of 80.53% is achieved by the nearest-neighbor (IB1) algorithm on the "rawsubj" feature set. We see that this is significantly better than the 77.62% correct achieved by the boosted algorithm (AdaBoost), and a separate comparison showed that it is also significantly better than the 75.49% correct achieved by the decision tree algorithm (J48). However, we also see that AdaBoost still performed significantly better than the baseline (ZeroR) on the "rawsubj" feature set, and a separate comparison showed that the performance of J48 was statistically the same as AdaBoost on this feature set.

We also see that overall, AdaBoost produces the most robust performance; on every other feature set AdaBoost performed as well as or better than both J48 and IB1. In particular, AdaBoost significantly outperforms IB1 on the four normalized feature sets, and AdaBoost significantly outperforms J48 on five feature sets: "rawspeech", the "n1" sets, and the "all" sets. AdaBoost's best performance is achieved on the "n1" feature sets; in fact AdaBoost significantly outperforms all the other algorithms on these two feature sets. Moreover, AdaBoost significantly outperforms the baseline on all feature sets other than the "n2" feature sets. Only on the "n2speech" feature set does AdaBoost perform significantly worse than the baseline; however, no algorithm outperformed the baseline on this feature set. In fact, we see that IB1 performed significantly worse than even AdaBoost on this feature set, while a separate comparison showed that J48 performed statistically the same as the baseline on this feature set.

AdaBoost's improvement for each feature set, relative to the baseline error of 28.2%, is shown in Table 2. Except for the "n2" feature sets, where improvement is not statistically significant, AdaBoost's relative improvement averages near 20%.

Feature Set	%Error	Rel.Imp.
rawspeech	25.15%	10.82%
rawsubj	22.38%	20.64%
n1speech	20.79%	26.28%
n1subj	21.39%	24.15%
n2speech	29.12%	–
n2subj	27.64%	–
allspeech	22.56%	20.00%
allsubj	22.04%	21.84%

Table 2. Relative Improvement of AdaBoost over Baseline

Other important evaluation metrics to consider include recall, precision, and F-Measure ((2*recall*precision)/(recall+precision)). Table 3 shows AdaBoost's performance on each feature set with respect to these metrics; in Weka, baseline (ZeroR) precision, recall and F-measure are all 0%. Though not shown, for recall and precision, AdaBoost performs statistically better than or equal to both IB1 and J48 on every feature set. For F-measure, AdaBoost performs statistically better than or equal to J48 on every feature set, and there is only one feature set ("n1subj") where IB1 performs significantly better. Overall, AdaBoost produces better precision than recall, but of the three metrics, precision is the most important in the intelligent tutoring domain, because it is better if ITSPOKE predicts that a turn is "neutral" than if ITSPOKE wrongly predicts a turn is positive or negative and reacts accordingly.

Feature Set	Precision	Recall	F-Measure
rawspeech	0.44	0.41	0.40
rawsubj	0.52	0.44	0.45
n1speech	0.54	0.45	0.48
n1subj	0.56	0.45	0.48
n2speech	0.40	0.31	0.33
n2subj	0.39	0.28	0.31
allspeech	0.49	0.43	0.44
allsubj	0.53	0.42	0.44

Table 3. Precision, Recall and F-Measure of AdaBoost

As discussed above, we use AdaBoost to "boost" the J48 de-

cision tree algorithm. Although the output of AdaBoost does not include a decision tree, to get an intuition about how our features are used to predict emotion classes in our domain, we ran the J48 algorithm on the "n1speech" feature set, using leave-one-out cross validation (LOO). We chose "n1speech" because it yields J48's best performance using 10x10 cross-validation, and contains only acoustic and prosodic features, thereby allowing the results to generalize to different speakers and problems. Although J48's accuracy on "n1speech" using 10x10 cross-validation was statistically worse than AdaBoost, a separate comparison showed it was significantly better than both IB1 and the baseline. In Figure 4 we present the final decision tree for J48 (LOO) and "n1speech". In this structure, a colon introduces the emotion label assigned to a leaf. The structure is read by following the (indented) paths through the feature nodes until reaching a leaf.

```
n1turnDur ≤ 2.476193
|   n1tempo ≤ 0.648936
|   |   n1turnDur ≤ 1.15: negative
|   |   n1turnDur > 1.15
|   |   |   n1maxf0 ≤ 0.904914: neutral
|   |   |   n1maxf0 > 0.904914: negative
|   n1tempo > 0.648936
|   |   n1tempo ≤ 2.558065: neutral
|   |   n1tempo > 2.558065
|   |   |   n1turnDur ≤ 0.360655
|   |   |   |   n1tempo ≤ 111.176611: neutral
|   |   |   |   n1tempo > 111.176611
|   |   |   |   |   n1turnDur ≤ 0.020106: neutral
|   |   |   |   |   n1turnDur > 0.020106: negative
|   |   |   n1turnDur > 0.360655: negative
n1turnDur > 2.476193
|   n1tempo ≤ 0.954545
|   |   n1meanRMS ≤ 0.221876: neutral
|   |   n1meanRMS > 0.221876
|   |   |   n1stdRMS ≤ 6.223242: negative
|   |   |   n1stdRMS > 6.223242
|   |   |   |   n1minf0 ≤ 0.286531
|   |   |   |   |   n1meanf0 ≤ 0.686056: neutral
|   |   |   |   |   n1meanf0 > 0.686056: negative
|   |   |   |   n1minf0 > 0.286531
|   |   |   |   |   n1turnDur ≤ 3.952382
|   |   |   |   |   |   n1tempo ≤ 0.600001: positive
|   |   |   |   |   |   n1tempo > 0.600001: neutral
|   |   |   |   |   n1turnDur > 3.952382
|   |   |   |   |   |   n1minf0 ≤ 0.38471: positive
|   |   |   |   |   |   n1minf0 > 0.38471
|   |   |   |   |   |   |   n1tempo ≤ 0.350877: positive
|   |   |   |   |   |   |   n1tempo > 0.350877
|   |   |   |   |   |   |   |   n1maxf0 ≤ 0.734688: positive
|   |   |   |   |   |   |   |   n1maxf0 > 0.734688: negative
|   n1tempo > 0.954545
|   |   n1%Silence ≤ 2.026144: neutral
|   |   n1%Silence > 2.026144
|   |   |   n1meanf0 ≤ 0.885155: negative
|   |   |   n1meanf0 > 0.885155: neutral
```

Fig. 4. Decision Tree for J48 (LOO) on "n1speech"

We see that the tree uses all 3 of our normalized temporal features: turn duration (n1turnDur), speaking rate (n1tempo), and amount of silence (n1%Silence), to predict our emotion classes; temporal features are also found to be predictive of user mental states in [7, 9, 10]. Of our 4 normalized pitch features, however, only minimum, maximum and mean f0 values (n1minf0, n1maxf0, n1meanf0) are used to predict our emotion classes; as in [10, 11], f0 standard deviation is not used to predict our classes. Of our 4 normalized energy features, only RMS mean and standard deviation (n1meanRMS, n1stdRMS) are used to predict our emotion classes; unlike in [9, 10, 11], RMS maximums and minimums are not used. Based on this tree, longer turns with slower tempos, higher RMS mean and standard deviation, median f0 minimums, and lower f0 maximums will be predicted positive. The tree will predict negative for shorter turns (but not the shortest turns, e.g. groundings) with either slower tempos and higher f0 maximums or with faster tempos. For longer turns with slower tempos and higher RMS means, negative emotions are predicted when RMS standard deviation is lower or when it is higher and the turn has either a lower f0 minimum and higher f0 mean or a higher f0 minimum and maximum. In longer turns with fast tempos and high % silence, a low f0 mean also predicts negative.

The accuracy (percent correct) of this decision tree is 77.23%. Table 4 shows its performance with respect to precision, recall and F-measure on each of the three emotion classes.

Precision	Recall	F-Measure	Class
0.535	0.548	0.541	negative
0.878	0.897	0.887	neutral
0.273	0.200	0.231	positive

Table 4. J48's (LOO) Precision, Recall and F-Measure

A confusion matrix summarizing J48's (LOO) performance on "n1speech" is shown in Table 5. The matrix shows how many instances of each class have been assigned to each class, where rows correspond to annotator-assigned labels and columns correspond to predicted labels. For example, 23 negatives were correctly predicted, while 13 negatives were incorrectly predicted to be neutral and 6 negatives were incorrectly predicted to be positive.

	negative	neutral	positive
negative	23	13	6
neutral	13	130	2
positive	7	5	3

Table 5. Confusion Matrix for J48 (LOO) on "n1speech"

4. CONCLUSIONS AND CURRENT DIRECTIONS

We have addressed the use of machine learning techniques for automatically predicting student emotional states in intelligent tutoring spoken dialogues. Our methodology extends the results of prior research into spoken dialogue systems by applying them to a new domain. Our emotion annotation schema distinguishes negative, neutral and positive emotions, and our inter-annotator agreement is on par with prior emotion annotation in other types of corpora. From our annotated student turns we automatically extracted 33 acoustic and prosodic features that will be available in real-time to ITSPOKE, and added 3 identifier features for student, gender, and problem. We compared the results of a variety of machine learning algorithms in terms of a variety of evaluation metrics; our best results suggest that ITSPOKE can be enhanced to automatically predict student emotional states.

We are exploring the use of other emotion annotation schemas. [22, 23, 8] for example, address how more complex emotional categorizations encompassing multiple emotional dimensions (e.g. "valence" and "activation") can be incorporated into frameworks for automatic prediction. One limiting factor of more complex schemas, however, is the difficulty of keeping inter-annotator agreement high; another is the difficulty of making good predictions about fine-grained emotion categorizations. We are also experimenting with annotating orthogonal tutoring sub-domains. For example, although we've found that most expressions of emotion pertain to the physics material being learned, they can also pertain to the tutoring process itself (e.g. attitudes toward the tutor/being tutored). We are also addressing the incorporation of "borderline" cases in our own schema. For example, we have already found that incorporating "weak" negative and positive labels into the corresponding "strong" category increases inter-annotator agreement but decreases machine learning performance.

We are also expanding our machine-learning investigations. We are extending our set of features to incorporate lexical and dialogue features. [24, 7, 9, 10] have shown that such features can contribute to emotion recognition, and in a pilot study [25] we showed that these features alone predicted emotional states significantly better than the baseline. In addition, we are pursuing further comparisons of alternative machine-learning techniques. Finally, we are in the process of increasing our data set. With more data we can try methods such as down-sampling to better balance the data in each emotion class. When ITSPOKE begins testing in Fall, 2003, we will address the same questions for our human-computer tutoring dialogues that we've addressed here for our parallel corpus of human-human tutoring dialogues.

5. REFERENCES

[1] C. P. Rose and V. Aleven, "Proc. of the ITS 2002 workshop on empirical methods for tutorial dialogue systems," Tech. Rep., San Sebastian, Spain, June 2002.

[2] R. Hausmann and M. Chi, "Can a computer interface support self-explaining?," *The International Journal of Cognitive Technology*, vol. 7, no. 1, 2002.

[3] G. Coles, "Literacy, emotions, and the brain," Reading Online, March 1999, 1999.

[4] M. Evens, S. Brandle, R. Chang, R. Freedman, M. Glass, Y. H. Lee, L. S. Shim, C. W. Woo, Y. Zhang, Y. Zhou, J. A. Michaeland, and A. A. Rovick, "Circsim-tutor: An intelligent tutoring system using natural language dialogue," in *Proceedings of the Twelfth Midwest AI and Cognitive Science Conference, MAICS 2001*, Oxford, OH, 2001, pp. 16–23.

[5] G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard, "Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens," in *Proc. of ITS*, 2002.

[6] P-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human Computer Interaction*, in press.

[7] 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. of ICSLP*, 2002.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, Jan. 2001.

[9] D. Litman, J. Hirschberg, and M. Swerts, "Predicting user reactions to system error," in *Proc.of ACL*, 2001.

[10] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions: Actors, wizards, and human beings," in *ISCA Workshop on Speech and Emotion*, 2000.

[11] C.M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *ASRU*, 2000.

[12] T. S. Polzin and A. H. Waibel, "Detecting emotions in speech," in *Cooperative Multimodal Communication*, 1998.

[13] K. Fischer, "Annotating emotional language data," Verbmobil Report 236, December 1999.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java implementations*, Morgan Kaufmann, San Francisco, 1999.

[15] K. VanLehn, P. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson, "The architecture of Why2-Atlas: A coach for qualitative physics essay writing," in *Proc. of ITS*, 2002.

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

[17] A. Black and P. Taylor, "Festival speech synthesis system: system documentation (1.1.1)," Human Communication Research Centre Technical Report 83, U. Edinburgh, 1997.

[18] C. P. Rosé, D. Bhembe, A. Roque, and K. VanLehn, "An efficient incremental architecture for robust interpretation," in *Proc. Human Languages Technology Conference*, 2002.

[19] C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein, "Interactive conceptual tutoring in Atlas-Andes," in *Proc. of A.I. in Education*, 2001, pp. 256–266.

[20] Carolyn Penstein Rosé, Diane Litman, Dumisizwe Bhembe, Kate Forbes, Scott Silliman, Ramesh Srivastava, and Kurt VanLehn, "A comparison of tutor and student behavior in speech versus text based tutoring," in *HLT/NAACL Workshop: Building Educational Applications Using NLP*, 2003.

[21] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," in *Proc. of the International Conference on Machine Learning*, 1996.

[22] J. Liscombe, J. Venditti, and J.Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *Proc. of EuroSpeech*, 2003.

[23] H. Holzapfel, C. Fuegen, M. Denecke, and A. Waibel, "Integrating emotional cues into a framework for dialogue management," in *Proc. of ICMI*, 2002.

[24] C.M. Lee, S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. of ICSLP*, 2002.

[25] D. Litman, K. Forbes, and S. Silliman, "Towards emotion prediction in spoken tutoring dialogues," in *Proc. of HLT/NAACL*, 2003.