Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources

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

Diane J. Litman
University of Pittsburgh
Department of Computer Science
Learning Research and Development Center
Pittsburgh PA, 15260, USA
litman@cs.pitt.edu

Abstract

We examine the utility of multiple types of turn-level and contextual linguistic features for automatically predicting student emotions in human-human spoken tutoring dialogues. We first annotate student turns in our corpus for negative, neutral and positive emotions. We then automatically extract features representing acoustic-prosodic and other linguistic information from the speech signal and associated transcriptions. We compare the results of machine learning experiments using different feature sets to predict the annotated emotions. Our best performing feature set contains both acoustic-prosodic and other types of linguistic features, extracted from both the current turn and a context of previous student turns, and yields a prediction accuracy of 84.75%, which is a 44% relative improvement in error reduction over a baseline. Our results suggest that the intelligent tutoring spoken dialogue system we are developing can be enhanced to automatically predict and adapt to student emotions.

1 Introduction

This paper investigates the automatic classification of student emotional states using acoustic-prosodic, non-acoustic-prosodic, and contextual information, in a corpus of human-human spoken tutoring dialogues. Motivation for this work comes from the discrepancy between the performance of human tutors and current machine tutors. In recent years, the development of computational tutorial *dialogue* systems has become more prevalent (Aleven and Rose, 2003), as one method of attempting to close the performance gap between human and computer tutors. It has been hypothesized that the success of such computer dialogue tutors could be further increased by modeling and adapting to student emotion; for example (Aist et al., 2002) have shown that adding human-provided emotional scaffolding to an automated reading tutor increases student persistence. We are building an intelligent tutoring spoken dialogue system with the goal of using spoken and natural language processing capabilities to automatically predict and adapt to student emotions. Here we present results of an empirical study demonstrating the feasibility of modeling student emotion in a corresponding corpus of human-human spoken tutoring dialogues.

Research in emotional speech has already shown that acoustic and prosodic features can be extracted from the speech signal and used to develop predictive models of emotion. Much of this research has used databases of speech read by actors or native speakers as training data (often with semantically neutral content) (Oudeyer, 2002; Polzin and Waibel, 1998; Liscombe et al., 2003). However, such prototypical emotional speech does not necessarily reflect natural speech (Batliner et al., 2003), such as found in tutoring dialogues. When actors are asked to read the same sentence with different emotions, they are restricted to conveying emotion using only acoustic and prosodic features. In natural interactions, however, speakers can convey emotions using other types of features, and can also combine acoustic-prosodic and other feature types. As a result of this mismatch, recent work motivated by spoken dialogue applications has started to use naturally-occurring speech to train emotion predictors (Litman et al., 2001; Lee et al., 2001; Ang et al., 2002; Lee et al., 2002; Batliner et al., 2003; Devillers et al., 2003; Shafran et al., 2003), but often predicts emotions using only acoustic-prosodic features that would be automatically available to a dialogue system in real-time. With noisier data and fewer features, it is not surprising that acoustic-prosodic features alone have been found to be of less predictive utility in these studies, leading spoken dialogue researchers to supplement such features with features based on other sources of information (e.g., lexical, syntactic, discourse).

Our methodology builds on and generalizes the results of this prior work in spoken dialogue emotion prediction, by introducing new linguistic and contextual features, and exploring emotion prediction in the domain of naturally occurring tutoring dialogues. We first annotate student turns in our human-human tutoring corpus for emotion. We then automatically extract acoustic-prosodic and other types of linguistic features from the student utterances in our corpus, and from their local and global dialogue contexts. We perform a variety of machine learning experiments using different feature combinations to predict our emotion categorizations. Our experiments show that 1) by using either acoustic-prosodic or other types of features alone, prediction accuracy is significantly improved compared to a baseline classifier for emotion prediction, 2) the addition of features identifying specific subjects and tutoring sessions only sometimes improves performance, and 3) prediction accuracy can typically be improved by combining features across multiple knowledge sources, and/or by adding contextual information. Our best learned model achieves a prediction accuracy of 84.75%, which is a relative improvement of 44% over the baseline error. Our results provide an empirical basis for enhancing the corresponding spoken dialogue tutoring system we are developing to automatically predict and ultimately to adapt to a student model that includes emotional states.

2 The Dialogue System and Corpus

We are currently building a spoken dialogue tutorial system called ITSPOKE (*I*ntelligent *T*utoring *SPOKE*n dialogue system) (Litman and Silliman, 2004), with the goal of automatically predicting and adapting to student emotions. ITSPOKE uses as its "back-end" the *text-based* Why2-Atlas dialogue tutoring system (VanLehn et al., 2002). In ITSPOKE, a student types an essay answering a qualitative physics problem. ITSPOKE then engages the student in spoken dialogue to correct misconceptions and elicit more complete explanations, after which the student revises the essay, thereby ending the tutoring or causing another round of tutoring/essay revision. Student speech is digitized from microphone input and sent to the Sphinx2 recognizer. The most probable "transcription" output by Sphinx2 is then sent to the Why2-Atlas natural language back-end for syntactic, semantic and dialogue analysis. Finally, the text response produced by the back-end is sent to the Cepstral text-to-speech system, then played to the student through a headphone. ITSPOKE has been pilot-tested and a formal evaluation with students is in progress.

Our human-human corpus contains spoken dialogues collected via a web interface supplemented with an audio link, where a human tutor performs the same task as ITSPOKE. Our subjects are university students who have taken no college 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 problems with the human tutor (via essay revision as described above), and 4) take a post-test similar to the pretest.[1] Our corpus contains 149 dialogues from 17 students. An average dialogue contains 45.3 student turns (242.2 words) and 44.1 tutor turns (1096.2 words). A corpus example is shown in Figure 1, containing the problem, the student's original essay, and an annotated (Section 3) excerpt from the subsequent spoken dialogue (some punctuation is added for clarity).

PROBLEM (TYPED): If a car is able to accelerate at 2 m/s^2, what acceleration can it attain if it is towing another car of equal mass?

ESSAY (TYPED): The maximum acceleration a car can reach when towing a car behind it of equal mass will be halved. Therefore, the maximum acceleration will be $1m/s^2$.

DIALOGUE (SPOKEN): . . . *9.1 min. into session* . . .

TUTOR$_1$: Uh let us talk of one car first.

STUDENT$_1$: ok. *(EMOTION = NEUTRAL)*

TUTOR$_2$: If there is a car, what is it that exerts force on the car such that it accelerates forward?

STUDENT$_2$: The engine *(EMOTION = POSITIVE)*

TUTOR$_3$: Uh well engine is part of the car, so how can it exert force on itself?

STUDENT$_3$: um... *(EMOTION = NEGATIVE)*

Figure 1: Excerpt from Human-Human Spoken Corpus

3 Annotating Student Emotion

In our spoken dialogue tutoring corpus, student emotional states can only be identified indirectly – via what is said and/or how it is said. We have developed an annotation scheme for hand labeling the student turns in our corpus with respect to three types of perceived emotions (Litman and Forbes-Riley, 2004):

Negative: a strong expression of emotion such as *confused, bored, frustrated, uncertain*. Because a syntactic question by definition expresses uncertainty, a turn containing only a question is by default labeled negative. An example negative turn is **student**$_3$ in Figure 1. Evidence of a negative emotion comes from the lexical item "um",

[1] The human-human corpus corresponds to the human-computer corpus that will result from ITSPOKE's evaluation, in that both corpora are collected using the same experimental method, student pool, pre- and post-test, and physics problems.

as well as acoustic and prosodic features, e.g., prior and post-utterance pausing and low pitch, energy and tempo.

Positive: a strong expression of emotion such as *confident, interested, encouraged*. An example is **student**$_2$ in Figure 1, with its lexical expression of certainty, "The engine", and acoustic and prosodic features of louder speech and faster 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 is **student**$_1$ in Figure 1. In this case, acoustic and prosodic features such as moderate loudness and tempo give evidence for the *neutral* label (rather than overriding it).

The features mentioned in the examples above were elicited during *post*-annotation discussion, for expository use in this paper. To avoid influencing the annotator's intuitive understanding of emotion expression, and because such features are not used consistently or unambiguously across speakers, our manual contains examples of labeled dialogue excerpts (as in Figure 1) with links to corresponding audio files, rather than a description of particular features associated with particular labels.

Our work differs from prior emotion annotations of spontaneous spoken dialogues in several ways. Although much past work predicts only two classes (e.g., negative/non-negative) (Batliner et al., 2003; Ang et al., 2002; Lee et al., 2001), our experiments produced the best predictions using our three-way distinction. In contrast to (Lee et al., 2001), our classifications are context-relative (relative to other turns in the dialogue), and task-relative (relative to tutoring), because like (Ang et al., 2002), we are interested in detecting emotional changes across our dialogues. Although (Batliner et al., 2003) also employ a relative classification, they explicitly associate specific features with emotional utterances.

To analyze the reliability of our annotation scheme, we randomly selected 10 transcribed dialogues from our human-human tutoring corpus, yielding a dataset of 453 student turns. (Turn boundaries were manually annotated prior to emotion annotation by a paid transcriber.) The 453 turns were separately annotated by two different annotators as *negative, neutral* or *positive*, following the emotion annotation instructions described above. The two annotators agreed on the annotations of 385/453 turns, achieving 84.99% agreement, with Kappa = 0.68.[2] This inter-annotator agreement exceeds that of prior studies of emotion annotation in naturally occurring speech (e.g., agreement of 71% and Kappa of 0.47 in (Ang et al., 2002), and Kappa ranging between 0.32 and 0.42 in (Shafran et al., 2003)). As in (Lee et al., 2001), the machine learning experiments described below use only those 385 student turns where the two annotators agreed on an emotion label. Of these turns, 90 were *negative*, 280 were *neutral*, and 15 were *positive*.

[2] $Kappa = \frac{P(A)-P(E)}{1-P(E)}$ (Carletta, 1996). P(A) is the proportion of times the annotators agree, and P(E) is the proportion of agreement expected by chance.

4 Feature Extraction

For each of the 385 agreed student turns described above, we next extracted the set of features itemized in Figure 2. These features are used in our machine learning experiments (Section 5), and were motivated by previous studies of emotion prediction as well as by our own intuitions.

Acoustic-Prosodic Features

- 4 normalized fundamental frequency (f0) features: maximum, minimum, mean, standard deviation
- 4 normalized energy (RMS) features: maximum, minimum, mean, standard deviation
- 4 normalized temporal features: total turn duration, duration of pause prior to turn, speaking rate, amount of silence in turn

Non-Acoustic-Prosodic Features

- lexical items in turn
- 6 automatic features: turn begin time, turn end time, isTemporalBarge-in, isTemporalOverlap, #words in turn, #syllables in turn
- 6 manual features: #false starts in turn, isPriorTutorQuestion, isQuestion, isSemanticBarge-in, #canonical expressions in turn, isGrounding

Identifier Features: subject, subject gender, problem

Figure 2: Features Per Student Turn

Following other studies of spontaneous dialogues (Ang et al., 2002; Lee et al., 2001; Batliner et al., 2003; Shafran et al., 2003), our acoustic-prosodic features represent knowledge of pitch, energy, duration, tempo and pausing. F0 and RMS values, representing measures of pitch and loudness, respectively, are computed using Entropic Research Laboratory's pitch tracker, *get_f0*, with no post-correction. Turn Duration and Prior Pause Duration are calculated via the turn boundaries added during the transcription process. Speaking Rate is calculated as syllables (from an online dictionary) per second in the turn, and Amount of Silence is approximated as the proportion of zero f0 frames for the turn, i.e., the proportion of time the student was silent. In a pilot study of our corpus, we extracted raw values of these acoustic-prosodic features,

then normalized (divided) each feature by the same feature's value for the first student turn in the dialogue, and by the value for the immediately prior student turn. We found that features normalized by first turn were the best predictors of emotion (Litman and Forbes, 2003).

While acoustic-prosodic features address how something is said, features representing what is said are also important. Lexical information has been shown to improve speech-based emotion prediction in other domains (Litman et al., 2001; Lee et al., 2002; Ang et al., 2002; Batliner et al., 2003; Devillers et al., 2003; Shafran et al., 2003), so our first non-acoustic-prosodic feature represents the transcription[3] of each student turn as a word occurrence vector (indicating the lexical items that are present in the turn).

The next set of non-acoustic-prosodic features are also automatically derivable from the transcribed dialogue. Turn begin and end times[4] are retrieved from turn boundaries, as are the decisions as to whether a turn is a temporal barge-in (i.e., the turn began before the prior tutor turn ended) or a temporal overlap (i.e., the turn began and ended within a tutor turn). These features were motivated by the use of turn position as a feature for emotion prediction in (Ang et al., 2002), and the fact that measures of dialogue interactivity have been shown to correlate with learning gains in tutoring (Core et al., 2003). The number of words and syllables in a turn provide alternative ways to quantify turn duration (Litman et al., 2001).

The last set of 6 non-acoustic-prosodic features represent additional syntactic, semantic, and dialogue information that had already been manually annotated in our transcriptions, and thus was available for use as predictors; as future research progresses, this information might one day be computed automatically. Our transcriber labels false starts (e.g., I do-don't), syntactic questions, and semantic barge-ins. Semantic barge-ins occur when a student turn interrupts a tutor turn at a word or pause boundary. Unlike temporal barge-ins, semantic barge-ins do not overlap temporally with tutor turns. Our transcriber also labels certain canonical expressions that occur frequently in our tutoring dialogues and function as hedges or groundings. Examples include "uh", "mm-hm", "ok", etc. (Evens, 2002) have argued that hedges can indicate emotional speech (e.g., "uncertainty"). However, many of the same expressions also function as groundings, which generally correspond to neutral turns in our dialogues. We distinguish groundings as turns that consist only of a labeled canonical expression and are not preceded by (i.e., not answering) a tutor question.[5]

Finally, we recorded 3 "identifier" features for each turn. Prior studies (Oudeyer, 2002; Lee et al., 2002) have shown that "subject" and "gender" can play an important role in emotion recognition, because different genders and/or speakers can convey emotions differently. "subject" and "problem" 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.

5 Emotion Prediction using Learning

We next performed machine learning experiments using the feature sets in Figure 3, to study the effects that various feature combinations had on predicting emotion. We compare our normalized acoustic-prosodic feature set (speech) with 3 non-acoustic-prosodic feature sets, which we will refer to as "text-based" sets: one containing only the lexical items in the turn (lexical), another containing the lexical items and the automatic features (autotext), and a third containing all 13 features (alltext). We further compare each of these 4 feature sets with an identical set supplemented with our 3 identifier features (+ident sets).

- **speech**: 12 normalized acoustic-prosodic features
- **lexical**: lexical items in turn
- **autotext**: lexical + 6 automatic features
- **alltext**: lexical + 6 automatic + 6 manual features
- **+ident**: each of the above sets + 3 identifier features

Figure 3: Feature Sets for Machine Learning

We use the Weka machine learning software (Witten and Frank, 1999) to automatically learn our emotion prediction models. In earlier work (Litman and Forbes, 2003), we used Weka to compare a nearest-neighbor classifier, a decision tree learner, and a "boosting" algorithm. We found that the boosting algorithm, called "AdaBoost" (Freund and Schapire, 1996), consistently yielded the most robust performance across feature sets and evaluation metrics; in this paper we thus focus on AdaBoost's performance. Boosting algorithms generally enable the accuracy of a "weak" learning algorithm to be improved by repeatedly applying it to different distributions of training examples (Freund and Schapire, 1996). Following (Oudeyer, 2002), we select the decision tree learner as AdaBoost's weak learning algorithm.

To investigate how well our emotion data can be learned with only speech-based or text-based features, Table 1 shows the mean accuracy (percent correct) and

[3] In our human-computer data, all features computed from transcriptions will be computed from ITSPOKE's logs (e.g., the best speech recognition hypothesis).

[4] These are computed relative to the beginning of the dialogue, e.g., the begin time of **tutor**$_1$ in Figure 1 is 9.1 minutes.

[5] This definition is consistent but incomplete, e.g., repeats can also function as groundings, but are not currently included.

standard error (SE)[6] of AdaBoost on the 8 feature sets from Figure 3, computed across 10 runs of 10-fold cross-validation.[7] Although not shown in this and later tables, all of the feature sets examined in this paper predict emotion significantly better than a standard majority class baseline algorithm (always predict "neutral", which yields an accuracy of 72.74%). For Table 1, AdaBoost's improvement for each feature set, relative to this baseline error of 27.26%, averages 24.40%, and ranges between 12.69% ("speech-ident") and 43.87% ("alltext+ident").[8]

Feature Set	-ident	SE	+ident	SE
speech	76.20	0.55	77.41	0.52
lexical	78.31	0.44	79.55	0.27
autotext	80.38	0.43	81.19	0.35
alltext	83.19	0.30	**84.70**	0.20

Table 1: %Correct on Speech vs. Text (cross-val.)

As shown in Table 1, the best accuracy of 84.70% is achieved on the "alltext+ident" feature set. This accuracy is significantly better than the accuracy of the seven other feature sets,[9] although the difference between the "+/-ident" versions was not significant for any other pair besides "alltext". In addition, the results of five of the six text-based feature sets are significantly better than the results of both acoustic-prosodic feature sets ("speech +/- ident"). Only the text-only feature set ("lexical-ident") did not perform statistically better than "speech+ident" (although it did perform statistically better than "speech-ident"). These results show that while acoustic-prosodic features can be used to predict emotion significantly better than a majority class baseline, using only non-acoustic-prosodic features consistently produces even significantly better results. Furthermore, the more text-based features the better, i.e., supplementing lexical items with additional features consistently yields further accuracy increases. While adding in the subject- and problem- specific "+ident" features improves the accuracy of all the "-ident" feature sets, the improvement is only significant for the highest-performing set ("alltext").

The next question we addressed concerns whether *combinations* of acoustic-prosodic and other types of features can further improve AdaBoost's predictive accuracy. We investigated AdaBoost's performance on the set of 6 feature sets formed by combining the "speech" acoustic-prosodic set with each text-based set, both with and without identifier features, as shown in Table 2.

Feature Set	-ident	SE	+ident	SE
lexical+speech	79.26	0.46	79.09	0.36
autotext+speech	79.64	0.47	79.36	0.48
alltext+speech	83.69	0.36	**84.26**	0.26

Table 2: %Correct on Speech+Text (cross-val.)

AdaBoost's best accuracy of 84.26% is achieved on the "alltext+speech+ident" combined feature set. This result is significantly better than the % correct achieved on the four "autotext" and "lexical" combined feature sets, but is not significantly better than the "alltext+speech-ident" feature set. Furthermore, there was no significant difference between the results of the "autotext" and "lexical" combined feature sets, nor between the "-ident" and "+ident" versions for the 6 combined feature sets.

Comparing the results of these combined (speech+text) feature sets with the speech versus text results in Table 1, we find that for autotext+speech-ident and all +ident feature sets, the combined feature set slightly decreases predictive accuracy when compared to the corresponding text-only feature set. However, there is no significant difference between the best results in each table (alltext+speech+ident vs. alltext+ident).

Emotion Class	Precision	Recall	F-Measure
negative	0.71	0.60	0.65
neutral	0.86	0.92	0.89
positive	0.50	0.27	0.35

Table 3: Other Metrics on "alltext+speech+ident" (LOO)

In addition to accuracy, other important evaluation metrics include recall, precision, and F-Measure ($\frac{2*recall*precision}{recall+precision}$). Table 3 shows AdaBoost's performance with respect to these metrics across emotion classes for the "alltext+speech+ident" feature set, using leave-one-out cross validation (LOO). AdaBoost accuracy here is 82.08%. As shown, AdaBoost yields the best performance for the neutral (majority) class, and has better performance for negatives than for positives. We also found positives to be the most difficult emotion to annotate. Overall, however, AdaBoost performs significantly better than the baseline, whose precision, recall and F-measure for negatives and positives is 0, and for neutrals is 0.727, 1, and 0.842, respectively.

6 Adding Context-Level Features

Research in other domains (Litman et al., 2001; Batliner et al., 2003) has shown that features representing the di-

[6] We compute the SE from the std. deviation (std(x)/sqrt(n), where n=10 (runs)), which is automatically computed in Weka.

[7] For each cross-validation, the training and test data are drawn from turns produced by the same set of speakers. We also ran cross-validations training on n-1 subjects and testing on the remaining subject, but found our results to be the same.

[8] Relative improvement over the baseline error for feature set x = $\frac{error(baseline)-error(x)}{error(baseline)}$, where error(x) is 100 minus the %correct(x) value shown in Table 1.

[9] For any feature set, the mean +/- 2*SE = the 95% confidence interval. If the confidence intervals for two feature sets are non-overlapping, then their mean accuracies are significantly different with 95% confidence.

alogue context can sometimes improve the accuracy of predicting negative user states, compared to the use of features computed from only the turn to be predicted. Thus, we investigated the impact of supplementing our turn-level features in Figure 2 with the features in Figure 4, representing local and global[10] aspects of the prior dialogue, respectively.

- **Local Features**: feature values for the two student turns preceding the student turn to be predicted
- **Global Features**: running averages and totals for each feature, over all student turns preceding the turn to be predicted

Figure 4: Contextual Features for Machine Learning

We next performed machine learning experiments using our two original speech-based feature sets ("speech +/- ident"), and four of our text-based feature sets ("autotext" and "alltext" +/- ident), each separately supplemented with local, global, and local+global features. Table 4 presents the results of these experiments.

Feature Set	-ident	SE	+ident	SE
speech+loc	76.90	0.45	76.95	0.40
speech+glob	77.77	0.52	78.02	0.33
speech+loc+glob	77.00	0.46	76.88	0.47
autotext+loc	78.06	0.33	78.24	0.45
autotext+glob	79.35	0.18	80.39	0.43
autotext+loc+glob	77.67	0.54	77.74	0.48
alltext+loc	80.33	0.46	80.99	0.40
alltext+glob	**83.85**	0.37	83.74	0.55
alltext+loc+glob	81.02	0.35	81.23	0.58

Table 4: %Correct, Speech vs. Text, +context (cross-val.)

AdaBoost's best accuracy of 83.85% is achieved on the "alltext+glob-ident" combined feature set. This result is not significantly better than the % correct achieved on its "+ident" counterpart, but both of these results are significantly better than the % correct achieved on all other 16 feature sets. Moreover, all of the results for both the "alltext" and "autotext" feature sets were significantly better than the results for all of the "speech" feature sets. Although the "alltext+loc" feature sets were not significantly better than the best autotext feature sets (autotext+glob), they were better than the remaining "autotext" feature sets, and the "alltext+loc+glob" feature sets were better than all of the autotext feature sets. For all feature sets, the difference between the "-ident" and "+ident" versions was not significant. In sum, we see again that the more text-based features the better: adding text-based features again consistently improves results significantly. We also see that global features perform better than local features, and while global+local perform better than local features, global features alone consistently yield the best performance.

Comparing these results with the results in Tables 1 and 2, we find that while overall the performance of contextual non-combined feature sets shows a small performance increase over most non-contextual combined or non-combined feature sets, there is again a slight decrease in performance across the best results in each table. However, there is no significant difference between these best results (alltext+glob-ident vs. alltext+speech+ident vs. alltext+ident).

Table 5 shows the results of combining speech-based and text-based contextual feature sets. We investigated AdaBoost's performance on the 12 feature sets formed by combining the "speech" acoustic-prosodic set with our "autotext" and "alltext" text-based feature sets, both with and without identifier features, and each separately supplemented with local, global, and local+global features.

Feature Set	-iden	SE	+iden	SE
auto+speech+lo	78.23	0.39	77.30	0.52
auto+speech+gl	79.33	0.22	78.84	0.39
auto+speech+lo+gl	78.26	0.20	78.01	0.43
all+speech+lo	82.44	0.31	82.15	0.56
all+speech+gl	**84.75**	0.32	84.35	0.20
all+speech+lo+gl	81.43	0.28	81.04	0.43

Table 5: %Correct on Text+Speech+Context (cross-val.)

AdaBoost's best accuracy of 84.75% is achieved on the "alltext+speech+glob-ident" combined feature set. This result is not significantly better than the % correct achieved on its "+ident" counterpart, but both results are significantly better than the % correct achieved on all 10 other feature sets. In fact, all the "alltext" results are significantly better than all the "autotext" results. Again for all feature sets, the difference between the "-ident" and "+ident" versions was not significant. In sum, adding text-based features again consistently improves results significantly, and global features alone consistently yield the best performance. Although the best result across all experiments is that of "alltext + speech + glob - ident", there is no significant difference between the best results here and those in our three other experimental conditions.

A summary figure of our best results for text (alltext) and speech alone, then combined with each other and with our best result for context (global), is shown in Figure 5, for the "+/- ident" conditions; baseline performance is also shown. As shown, the accuracy of the "-ident" condition monotonically increases as features

[10]Running totals are only computed for numeric features if the result is interpretable, e.g., for turn duration, but not for tempo. Running averages for text-based features additionally include a "# turns so far" feature and a "# essays so far" feature.

are added or replaced in the right-to-left order shown. The "+ident" condition initially increases, then decreases with the addition of "global" or "speech" features to the "alltext" feature set, but then slightly increases again when these feature sets are combined. With less features "+ident" typically outperforms "-ident", although this switches when "alltext" and "global" features are combined (with and without "speech").

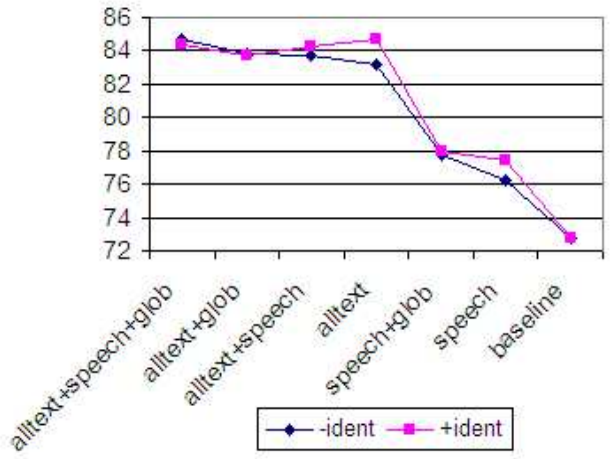

Figure 5: Comparison of %Correct for Best Results

7 Feature Usage in Machine Learning

As discussed above, we use AdaBoost to "boost" a decision tree algorithm. Although the Weka 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 basic decision tree algorithm on our highest-performing feature set, "alltext+speech+glob-ident". Table 6 shows the feature types used in this feature set, and the feature usages of each based on the structure of the tree. Following (Ang et al., 2002), feature usage is reported as the percentage of decisions for which the feature type is queried. As shown, the turn-based (non-context) text-based features are the most highly queried, with lexical items and manual features queried most, followed by the temporal (speech-based) features. Manual text-based global features are queried far more than other global features.

8 Conclusions and Current Directions

We have examined the utility of different features for automatically predicting student emotions in a corpus of tutorial spoken dialogues. Our emotion annotation schema distinguishes negative, neutral and positive emotions, with inter-annotator agreement and Kappa values that exceed those obtained for other types of spoken dialogues. From our annotated student turns we extracted a

Features	Turn	Global	Total
Speech-Based	**14.29%**	**1.97%**	**16.26%**
Temporal	12.81%	0.99%	13.79%
Energy	1.48%	0.99%	2.46%
Pitch	0%	0%	0%
Text-Based	**67.98**	**15.76**	**83.74%**
Lexical	41.87%	-	41.87%
Automatic	8.37%	0.99%	9.36%
Manual	17.73%	14.78%	32.51%

Table 6: Feature Usage for "alltext+speech+glob-ident"

variety of acoustic and prosodic, text-based, and contextual features. We used machine learning to examine the impact of different feature sets (with and without identifier features) on prediction accuracy. Our results show that while acoustic-prosodic features outperform a baseline, non-acoustic-prosodic features, and combinations of both types of features, perform even better. Adding certain types of contextual features and identifier features also often improves performance. Our best performing feature set, which contains speech and text-based features extracted from the current and previous student turns, yields an accuracy of 84.75% and a 44% relative improvement in error reduction over a baseline. Our experiments suggest that ITSPOKE can be enhanced to automatically predict student emotions.

We are currently exploring the use of other emotion annotation schemas for emotion prediction, such as those that incorporate categorizations encompassing multiple dimensions (Craggs, 2004; Cowie et al., 2001) and those that examine emotions at smaller units of granularity than turns (Batliner et al., 2003). With respect to predicting emotions, we plan to explore additional features found to be useful in other studies of spoken dialogue (e.g., language model, speaking style, dialog act, part-of-speech, repetition, emotionally salient keywords, word-level prosody (Batliner et al., 2003; Lee et al., 2002; Ang et al., 2002)) and in text-based applications (Qu et al., 2004). We are also exploring methods of combining information other than by feature level combination, such as data fusion across multiple classifiers (Lee et al., 2002; Batliner et al., 2003). For evaluation, we would like to see whether the ordering preferences among feature sets (as in Figure 5) are the same when recall, precision, and F-measure are plotted instead of accuracy. Furthermore, we are investigating whether greater tutor response to emotions correlates with greater student learning. Finally, when ITSPOKE's evaluation is completed, we will address the same questions for our human-computer dialogues that we have addressed here for our corresponding human-human dialogues.

Acknowledgments

This research is supported by NSF Grants Nos. 9720359 and No. 0328431. We thank Kurt VanLehn and the Why2-Atlas team, and Scott Silliman of ITSPOKE, for system development and data collection. We also thank Pamela Jordan and Mihai Rotaru for helpful suggestions.

References

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. of Intelligent Tutoring Systems, 6th International Conf. (ITS)*.

V. Aleven and C. P. Rose, editors. 2003. *Proc. of the AIED 2003 Workshop on Tutorial Dialogue Systems: with a view toward the classroom*.

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 *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 2037–2040.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. How to find trouble in communication. *Speech Communication*, 40:117–143.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

M. G. Core, J. D. Moore, and C. Zinn. 2003. The role of initiative in tutorial dialogue. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 67–74.

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

R. Craggs. 2004. A two dimensional annotation scheme for emotion in dialogue. In *Proc. of AAAI Spring Symposium: Exploring Attitude and Affect in Text*.

L. Devillers, L. Lamel, and I. Vasilescu. 2003. Emotion detection in task-oriented spoken dialogs. In *Proc. of the IEEE International Conference on Multimedia & Expo (ICME)*.

M. Evens. 2002. New questions for Circsim-Tutor. Presentation at the 2002 Symposium on Natural Language Tutoring, University of Pittsburgh.

Y. Freund and R.E. Schapire. 1996. Experiments with a new boosting algorithm. In *Proc. of 19th International Conf. on Machine Learning (ICML)*, pages 148–156.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2001. Recognition of negative emotions from the speech signal. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2002. Combining acoustic and language information for emotion recognition. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*.

J. Liscombe, J. Venditti, and J.Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proc. of EuroSpeech*.

D. Litman and K. Forbes-Riley. 2004. Annotating student emotional states in spoken tutoring dialogues. In *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue*.

D. Litman and K. Forbes. 2003. Recognizing emotion from student speech in tutoring dialogues. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

D. Litman and S. Silliman. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Proc. of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL) (Companion Proceedings)*.

D. Litman, J. Hirschberg, and M. Swerts. 2001. Predicting user reactions to system error. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL)*, pages 362–369.

P-Y. Oudeyer. 2002. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Studies*, 59(1-2):157–183.

T. Polzin and A. Waibel. 1998. Detecting emotions in speech. In *Proc. of Cooperative Multimodal Communication*.

Y. Qu, J. G. Shanahan, and J. Wiebe, editors. 2004. *AAAI Working Notes of the Spring Symposium: Exploring Attitude and Affect in Text*, Stanford, CA.

I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In *Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

K. VanLehn, P. W. Jordan, C. P. 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. of Intelligent Tutoring Systems, 6th International Conference (ITS)*.

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