

# Using Word-level Pitch Features to Better Predict Student Emotions during Spoken Tutoring Dialogues

Mihai Rotaru, Diane J. Litman

Department of Computer Science  
University of Pittsburgh, 210 S. Bouquet, Pittsburgh, PA, 15260, USA  
{mrotaru, litman}@cs.pitt.edu

## Abstract

In this paper, we advocate for the usage of word-level pitch features for detecting user emotional states during spoken tutoring dialogues. Prior research has primarily focused on the use of turn-level features as predictors. We compute pitch features at the word level and resolve the problem of combining multiple features per turn using a *word-level emotion model*. Even under a very simple word-level emotion model, our results show an improvement in prediction using word-level features over using turn-level features. We find that the advantage of word-level features lies in a better prediction of longer turns.

## 1. Introduction

We investigate the utility of using pitch features applied at the word level for the task of predicting student emotions in two corpora of spoken tutoring dialogues. Motivation for this work comes from the performance gap between human tutors and current machine tutors; typically students tutored by human tutors learn more than students tutored by computer tutors. One of the methods currently being explored as a way of closing this gap is to incorporate affective reasoning into current computer tutoring systems, including dialogue-based tutoring systems, e.g. [1, 2].

Previous spoken dialogue research in other domains has shown that turn-level prosodic, lexical, dialogue, and other features can be used to predict user emotional states [3-5]. To better approximate the prosodic information [6] uses word-level features and successfully applies them to a different emotion detection task. To our knowledge, there is no previous work that directly compares the impact of using features at the sub-turn rather than the turn level for emotion prediction. In this paper we are performing a first comparison of the two levels for the task of detecting student emotional states.

There are many choices for sub-turn units (breath groups, intonational phrases, syntactic chunks, words, syllables). We will use words as our sub-turn units because it is straightforward to do the segmentation and because these units have been used successfully by other researchers for similar tasks [6]. Moreover, in a real-time dialogue system, the segmentation is available as a byproduct of the automatic speech recognition.

To simplify our word versus turn-level feature comparison, we will focus *only* on pitch features. Pitch describes how high or low (frequency-wise) speech is rendered (the melody of the rendering). For example, in English, the sentence ‘This is great’ uttered as an exclamation usually expresses a positive emotion, while the same lexical construct uttered with an alternative pitch contour often

expresses a negative emotion. Changes in the speaking style are directly reflected in the shape of the pitch contour.

Our hypothesis is that using word-level features will be better for emotion prediction than using turn-level features. The intuition behind this hypothesis is that, at least for pitch information, computing the pitch features at the word level will give a better approximation of the pitch contour which in turn will help us do better in emotion prediction. For example, in Figure 1, the pitch contour shape for ‘‘This is great’’ uttered as an exclamation is better approximated by three linear regression lines at the word level than by a single regression line at the turn level. Moreover, emotion might not be expressed over an entire turn (especially for long turns) but on certain parts of a student turn; for this reason computing the features at the turn level might mitigate the effect of ‘‘emotional’’ parts of the turn. Returning to our previous ‘This is great’ example, in general, the word ‘great’ bears the highest change in prosody between the two styles of rendering the sentence. The small change in prosody for the first part of the sentence will mitigate the effect of ‘great’ if pitch features are computed at the turn level.

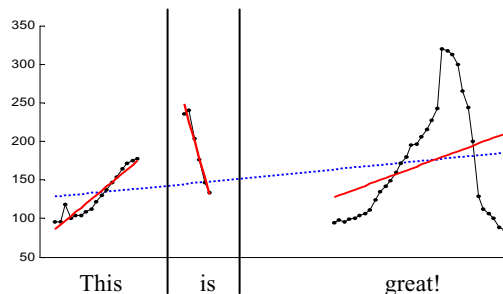


Figure 1: Pitch contour approximation using linear regression (turn level and word level).

We will investigate this hypothesis using various corpus-learner combinations. As our results will show, even under a very simple word-level emotion model, using word-level features proves to be better than turn level features. Moreover, our results indicate that the advantage of word-level features lies in a better prediction of longer turns.

## 2. Corpora

We have developed an annotation scheme for annotating emotions and attitudes in the tutoring domain, and have previously applied it to corpora of both human-human (HH) and human-computer (HC) spoken tutoring dialogues [2]. In the HC setting, students interact with our speech enabled tutoring system (ITSPOKE) using head-mounted microphone input and speech output. In the HH setting, our system is replaced by a human tutor performing the same tasks. In our

annotation scheme, each student turn is labeled for both strong and weak perceived expressions of emotion (confused, bored, irritated, uncertain, sad, confident, enthusiastic, etc.).

In our previous work [2], various combinations of our emotion classes were studied to learn about the ability to predict different types of emotional distinctions. In this paper we will use the emotional/non-emotional (*EnE*) distinction: turns where the student exhibited any emotion are labeled as emotional; all other turns are labeled as non-emotional. Our ongoing research on the HH corpus suggests that the *EnE* classification will be useful for triggering system adaptation to student emotions.

The agreed turns are the turns labeled with the same emotion class by our two annotators (the last author and another member of our group). Following [2, 3], this first study will use only the *agreed* subsets of our corpora because they offer the clearest cases of emotional turns. Given our promising results, replicating our experiments on the consensus labeling is among our priorities for the future (for the consensus labeling, the two original annotators revisited each originally disagreed case and, through discussion, sought a consensus label). For full details regarding our system and the annotation scheme see [2].

### 3. Features

Conveying the intended meaning of a sentence involves not only appropriate word selection but also the appropriate way of uttering the words. Prosodic features are often computed to quantify this rendering aspect. Pronunciation aspects can be captured using various information sources such as pitch, duration and amplitude. In this paper we will focus only on the pitch information because changes in speaking style are directly reflected in the shape of the pitch contour. Moreover word-level pitch features offer a better approximation of the pitch contour shape than turn-level features (recall Figure 1). If pitch contour shape is indeed useful for emotion prediction, then a better approximation of the contour might result in an improvement in prediction. Since the advantage of word-level features is not that clear for the other prosodic information sources, we elected not to use them in this study. Nonetheless, we believe they are important and we plan to incorporate them in our future work.

We will approximate the pitch contour (fundamental frequency or F0) using nine features: Minimum, Maximum, Mean, Standard Deviation, Onset, Offset, Linear regression coefficient, Linear regression error and Quadratic regression coefficient. The first four are commonly used by researchers for various tasks (negative emotion detection [5]; predicting user corrections [7]) and were also employed in previous studies on our corpora [2]. These four pitch features give us a very coarse approximation of the pitch contour for an entire turn.

Inspired by [6], we will use the following new features that offer a better approximation of the pitch contour: onset (the first F0 value), offset (the last F0 value), regression coefficient and regression error. Linear regression is performed to approximate the pitch contour shape. The regression coefficient estimates the direction of the pitch contour (rising or dropping) and can be used to distinguish, for example, questions and statements (at least for English). The regression error offers a better approximation than the standard deviation of the spread of the pitch contour relative

to the pitch contour direction (the regression line). Since highly emotional speech is believed to have a large variation in pitch contour, this value may be a good indicator of emotional speech. Moreover, to better approximate the shape of the pitch contour (at least at levels smaller than the entire turn – convex or concave) we also use the second order coefficient of the quadratic interpolation. This value relates to the Tilt model [8] and approximates the intonation used.

### 4. Turn and word-level prediction tasks

Recall that our goal is to investigate whether using features at a word-level will help in emotion prediction. To investigate our hypothesis, we extract the pitch features described in Section 3 at the turn and word level. For the word level features, for each word, instead of using the entire pitch contour, only the segment corresponding to the word in question is used in computing the features. Also, to account for word order and word position in the turn, we create two additional positional features for each word: the number of words before and after that word.

However, using word level features introduces two major problems given our turn-level annotation scheme. First, we do not know which of the words in an emotional turn are the words where the emotion is expressed. The only thing we know is that the sequence of words results in a certain emotional class. This will impact our training procedure, as discussed below. Second, assuming that we can predict an emotional class for each word, we still need to combine the sequence of predicted emotional classes into a single class (to label the turn as a whole, as in our annotations).

We will use the following simplified *word-level emotion model*. In the training phase, each word is labeled with the turn class and a model for predicting the word emotion is built using all the words from all turns in our training corpus (i.e. we predict word labels). In the test phase, for each turn, we predict the class of each word in the turn and then combine the word classes using majority voting (ties broken randomly). That is, the most frequent emotional class among the turn’s words will be the turn’s emotional class.

Here is an example from our HC corpus. In the training phase, the student turn “They are the same” will produce *four* training instances, one for each word. Pitch features will be computed for each individual word. In our corpus, this turn was labeled as emotional, thus all four instances will have the emotional class. This training data (which is larger than the training data for turn-level features since many turns have at least two words) is used by the classifier to learn a model. During the test phase, whenever we need to predict the class for a turn, for example the turn “It will change”, we will produce an instance for each word in the turn (three instances in our example) and use the learned model to classify them. Finally, the turn class will be the class that labeled the highest number of words in the turn.

### 5. Results

We will test our hypothesis on four combinations of two contrasting corpora and two contrasting learners. Table 1 highlights some of the differences between our two corpora (the HH and HC corpora). The HC corpus is smaller in size and has shorter turns than the HH one. Conceivably, the HC turns contain less emotional content making prediction more

difficult. Our two corpora also have different class distributions. Our previous turn-level studies [2] showed that the two corpora also differ in the types of features that offer the best performance for emotion prediction. The best accuracy values and the best accuracy on prosodic features are given for reference only. Our results and previous work results can not be compared directly (some differences in the corpus size, we use more pitch features while the previous work uses less pitch features but more prosodic and contextual features).

As another way to investigate the generality of our results, we use two contrasting learners from the Weka toolkit [9]: a nearest neighbor classifier (IB1) and boosted decision trees (ADA). IB1 is a lazy learner while ADA is an abstraction-based learner. In our previous turn-level studies [2], ADA yielded the most robust performance across feature sets and corpora. [10] have found that memory-based learning and abstraction-based learning algorithms can produce significantly different performance depending on several factors such as the language learning task, the number of features, and the type of features.

Table 1: HH and HC corpora properties

	HH	HC
Number of turns (# of turn-level instances)	319	220
Number of words (# of word-level instances)	1310	511
Class distribution (E/nE)	148/171	129/91
Average turn length in words	6.11	2.42
Best accuracy (previous work)	88.86%	66.36%
prosodic features only	84.71%	59.18%

Tables 2 and 3 show, for each corpus, the mean accuracy (% correct) and the standard error (SE) for our two learners using turn-level and word-level features, computed across 10 runs of 10-fold cross-validation. For comparison, the majority class baseline is shown in each caption. To determine whether the difference between word-level and turn-level performance is statistically significant we use confidence intervals. The “REL” column reports the results of our comparison. A “<\*” means that word-level features significantly outperformed turn-level features ( $p < 0.05$ ), while a “<” means a trend ( $p < 0.10$ ).

Table 2: Results for the HH corpus (Baseline: 53.61%)

	Turn-level		REL	Word-level	
	Mean Acc	SE		Mean Acc	SE
ADA	74.22	0.59	<*	78.67	0.39
IB1	67.98	0.35	<*	69.60	0.44

We will first discuss our results on the HH corpus (Table 2). All learner-feature type combinations perform statistically better than the majority class baseline. For both learners we find that word-level features statistically outperform turn-level features proving our hypothesis true on this corpus. In the case of ADA, we see a 4.4% absolute improvement from computing pitch features at the word level which corresponds to a 17% relative improvement over turn-level features. IB1 performs worse than ADA, but word-level features still outperform turn-level features with this learner (1.6% absolute improvement). On par with our previous work, we

observe that the pitch features alone, either at turn or word level, perform worse than the best accuracy reported on this corpus (recall Table 1, last row). The pitch features alone also perform worse than the previously reported performance on prosodic features, suggesting the importance of other prosodic information sources (amplitude, duration) for emotion prediction in this corpus.

To better understand the difference between turn-level and word-level feature sets, we also analyzed the performance as a function of turn length (Figure 2). Given our small dataset (319 turns), we divided the turns in our corpus in four categories: single (turns with only one word), short (turns with 2 to 4 words), medium (turns with 5 to 10 words) and long (turns with more than 10 words). The distribution in the HH corpus is: single 48%, short 25%, medium 17% and long 10%. Next, we used the predictions from the 10 x 10 cross validation experiments and computed the average accuracy for each category for all learner-feature type combinations.

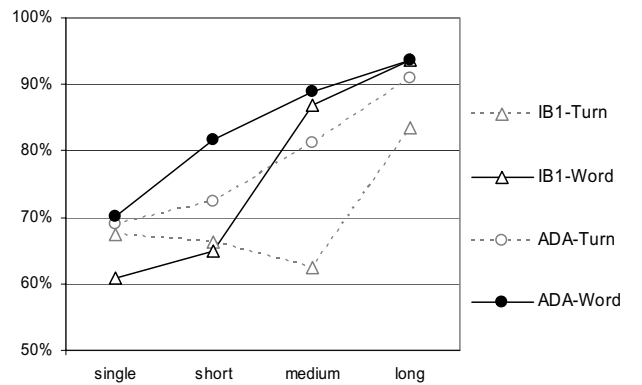


Figure 2: Accuracy as function of turn length (HH corpus)

We observe that as turn length increases, the word-level features outperform turn-level features. For the “single” category (turn with only one word) we can observe that turn-level features have better or similar accuracy with the word-level feature. We hypothesize that this difference in performance is due to learner’s sensitivity to the noise introduced by our simple word-level emotion model (all words from emotional turns are labeled as emotional). As a result of this simple model, the training set for emotional words is contaminated with words that are actually neutral, resulting in a lower accuracy for word-level features. But as turn length increases, the word-level features catch up and outperform turn-level features. This supports our hypothesis that word-level pitch features give a more accurate account of the pitch information, at least for our emotion prediction task. Surprisingly, for the “long” category, the gap between word-level and turn-level performance decreases; our model seems to have a bigger impact on short and medium turns. Interestingly, as turn length increases, the two predictor’s performance on word level features becomes similar.

Table 3 reports our results for the HC corpus. Again, all learner-feature type combinations performed better than the baseline but the improvement over the baseline is much smaller than for the HH corpus. This is on par with our previous work [2] which showed that emotion annotation and prediction in HC corpus is harder than in the HH corpus (recall Table 1, last row). Word-level features perform better than turn level features for both learners but the difference is

significant only for IB1 while for ADA it is only a trend (note ADA's high standard errors). Even if emotion prediction in the HC corpus is a hard task, our results show that improvement can be obtained by switching to word-level pitch features. Similar to the HH corpus, the advantage of word-level features lies in a better prediction of longer turn.

Table 3: Results for the HC corpus (Baseline: 58.64%)

	Turn-level		REL	Word-level	
	Mean Acc	SE		Mean Acc	SE
ADA	64.59	0.88	< <sup>t</sup>	67.61	0.81
IB1	64.21	0.57	<*	68.41	0.59

In contrast with the HH corpus, in the HC corpus word-level pitch features perform better than both previous results with prosodic features and the best reported accuracy on this corpus (though these results can not be compared directly). Since the turn-level features also perform well, we believe this improvement is due to the extended pitch features set as well as to the word-level approach.

## 6. Related work

Recognizing the importance of pitch for detecting emotion in dialogues, previous research has extensively used it to improve emotion prediction. But the majority of previous work has employed turn-level features. In general, the pitch contour was coarsely approximated using only four features (minimum, maximum, mean and standard deviation) [2, 5, 7]. In contrast, [6] uses word-level features extending dramatically the features extracted from the pitch information (recall Section 3). Although their assumptions are similar to ours (all words in an emotional turn are emotional), they do not identify the existence of a word-level emotion model. While our word-level emotion model is a very simplistic one, more complicated models can be imagined that could lead to better performance (e.g. HMMs). Moreover, [6] does not study the impact of word-level features which is precisely the scope of our paper.

[11] proposes a method to circumvent the word-level problem by creating new turn-level features out of sub-turn level features. Pitch features extracted from sub-turn levels are combined by fitting appropriate parametric distributions based on the assumption that these features are generated by certain parametric distributions. The parameters of these distributions are then used as turn-level features, in this way bypassing the sub-turn level problems.

Previous work [2, 3, 5] has shown that the lexical information is very important for emotion prediction. But the approach from [11] has problem with using lexical features. In contrast, our word level emotion model can be easily extended to include lexical information. All we have to do is add to the word-level pitch feature set a new feature. Its value will be the word itself. In this way, the new feature set will capture not only how the word was pronounced but also the word itself. Experimenting with the lexical extension of our model is one direction we are currently pursuing.

## 7. Conclusions and future work

In this paper we have been advocating for the usage of word-level pitch features for emotion prediction in speech-based tutoring corpora. We described the problems that we face

when using word level features and addressed them via a word-level emotion model. Even under a very simple word-level emotion model, our results indicate that word-level pitch features outperform turn-level pitch features. Our investigation of the performance as function of turn length indicates that word-level pitch features handle longer turns better than turn level features. The fact that our results hold for our combinations of contrasting corpora and learners supports the generality of our conclusions.

In our future work we plan to experiment with more refined word-level emotion models. We plan to learn a prosodic model for non-emotional words (based on the assumption that all words from a non-emotional turn are non-emotional) and use it to better identify emotional words in an emotional turn. Then, based on the training set we can learn how to predict the emotional class of a turn from the classes of its words (instead of using majority voting). Filtering irrelevant words (e.g. stop words) might offer further improvements. We also plan to extend our analysis to other emotion classifications (e.g. positive, negative; see [2]) and corpora from other domains. Finally, integrating lexical and other prosodic information sources (amplitude and duration) is among our priorities.

## 8. Acknowledgements

This research is supported by NSF Grant No. 0328431. We thank the ITSPOKE and NLP groups, Dan Bohus and Emil Talpes for their helpful suggestions.

## 9. References

- [1] G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard, "Experimentally augmenting an intelligent tutoring system with human-supplied capabilities," *Intelligent Tutoring Systems*, 2002.
- [2] D. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," *Assoc. for Computational Linguistics (ACL)*, 2004.
- [3] J. Ang, R. Dhillon, A. Krupski, A. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," *ICSLP*, 2002.
- [4] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion Detection in Task-Oriented Spoken Dialogs," *IEEE Int. Conference on Multimedia & Expo*, 2003.
- [5] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," *ICSLP*, 2002.
- [6] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," *Speech Communication*, vol. 40 (1-2), 2003.
- [7] J. Hirschberg, D. Litman, and M. Swerts, "Identifying User Corrections Automatically in Spoken Dialogue Systems," *North American Chapter of the Association for Computational Linguistics*, 2001.
- [8] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107, 2000.
- [9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java implementations*: Morgan Kaufmann, 1999.
- [10] M. Rotaru and D. Litman, "Exceptionality and Natural Language Learning," *Computational Natural Language Learning*, 2003.
- [11] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," *ICSLP*, 1998.