

# Exploiting Word-level Features for Emotion Prediction

*Greg Nicholas, Mihai Rotaru, Diane J. Litman*

Department of Computer Science  
University of Pittsburgh, 210 S. Bouquet, Pittsburgh, PA, 15260, USA  
{gdn2, mrotaru, litman}@cs.pitt.edu

## ABSTRACT

In this paper we study two techniques for combining word-level features for emotion prediction. Prior research has primarily focused on the use of turn-level features as predictors. Recently, the utility of word-level features has been highlighted but only tested on relatively small human-computer corpora. We extend over previous work by investigating the strengths and weaknesses of two different techniques for using word-level features and by using a larger corpus of human-computer dialogue. Our results confirm that the word-level pitch features fare better than the turn-level ones regardless of the combination technique. In addition, we find that each word combination technique has different strengths and weaknesses in terms of precision and recall.

Index Terms— Speech analysis, Feature extraction, Pattern classification, Natural language interfaces

## 1. INTRODUCTION

In the past few years there has been a surge of interest in emotion-handling in spoken dialogue systems. Detecting and reacting to user emotions is considered to be an important direction for improving spoken dialogue systems in domains like call centers [1-3] and intelligent tutoring systems [4, 5].

Previous work on emotion prediction uses features derived from a variety of sources: prosody, acoustic information, lexical information, and meta-dialogue information. An important factor when computing these features is deciding the level of granularity within the turn to extract them from. The majority of previous work computes features at the turn level [1-3], but recent work [6, 7] has shown that computing features at sub-turn levels can be beneficial for emotion prediction. For example, [6] shows that adding acoustic-prosodic features computed at the breath group level can further improve emotion detection;

[7] directly compares the utility of turn-level and word-level pitch features and shows that the latter are more informative. The intuition behind using sub-turn features is that they offer a better approximation of the acoustic-prosodic profile and that emotion might not be expressed over the entire turn. Following [7], our work focuses on word-level as the level of turn granularity.

Using word-level features introduces an additional complication: since the goal is to predict the emotional classification of an entire turn, how can a variable number of word-level features be combined to produce a turn-level emotion prediction? Previous work uses two techniques. The first is a model that predicts each individual word's emotional classification and then combines these predictions to classify the turn [7, 8]. The second model uses a single set of features derived from a predefined subset of sub-turn units [6] (in our case, words).

The goal of our work is to investigate which kinds of features can produce the best emotion classifier. In this study we perform an in-depth comparison of the utility of turn-level features and of word-level features using the two combination techniques mentioned above.

Our work extends upon previous work on several dimensions. While previous work has looked at only one sub-turn combination technique [6-8], here we experiment with two techniques and study the advantages/disadvantages of each technique. We extend the turn-versus-word-level pitch features comparison from [7] by looking at a second word combination technique and by looking at a different emotion prediction task. While [6] applies its sub-turn combination technique at the breath group level and on a human-human corpus, here we investigate the utility of a variation of their technique at the word level and on a human-computer corpus. Finally, our analysis is performed on a human-computer corpus considerably larger than previous work (i.e. 45 times bigger than the one in [7]).

Our results show that using pitch features at the word level via both word combination techniques outperforms turn-level pitch features. Our analysis of the two word combination techniques shows that they both produce

models with similarly high recall but precision that is lower and more variable.

## 2. CORPUS AND ANNOTATION

The corpus analyzed in this paper consists of 357 experimentally obtained spoken tutoring dialogues between 82 students and our system ITSPOKE, a speech-enabled version of the text-based WHY2-ATLAS conceptual physics tutoring system [10]. As part of the interaction, ITSPOKE engages the student in spoken dialogue (using speech-based input and output) to correct misconceptions from previously typed essays and/or to elicit more complete explanations. For speech generation, our system uses the Cepstral text-to-speech system and for recognition, the Sphinx2 speech recognizer with stochastic language models. To compute word level features, in our offline experiments each turn was automatically segmented into words by running Sphinx2 in forced-alignment mode using the human transcript. For real-time applications, the word segmentation is available as a byproduct of the automated speech recognition.

Each student turn from our corpus (9,854 turns) was annotated on an Uncertain/non-Uncertain (UnU) dimension: turns where the student displayed uncertainty in his/her response are labeled as uncertain, while all others are labeled as non-uncertain. This particular emotional classification was used because it is widely believed that human tutors respond to student uncertainty. To date, one annotator has annotated our entire corpus. In a preliminary inter-annotator study, a second annotator labeled a subset of the corpus (2,334 turns) resulting in a 90% agreement (Kappa = 0.68).

Table 1 – Comparison between corpora from current and previous studies

	Previous study	Current study
Number of turns	220	9854
Number of words	511	27548
Average words per turn	2.32	2.80
Emotional classification	Emotional/ Non-emotional	Uncertain/ Non-uncertain
Class distribution	129/91 (E/nE)	2189/7665 (U/nU)
Baseline	58.64%	77.79%

Table 1 highlights the differences between the corpora from this and the previous study [7].

## 3. FEATURES

Previous work [1-8] has highlighted the utility of a variety of information sources (e.g acoustic-prosodic, lexical, dialogue context) for emotion prediction. As in [7]

and to simplify our comparison, here we focus primarily on the pitch information. The pitch contour is approximated by nine features: Minimum, Maximum, Mean, Standard Deviation, Onset, Offset, Linear regression coefficient, Linear regression error, and Quadratic regression coefficient. For a more thorough description of these features, see [7]. For the turn-level features, the entire pitch contour is used; for the word-level features, only the pitch contour of the word is used.

Predicting emotions using the turn-level features is straightforward. First, the learner is given features computed at the turn-level and the corresponding class for training. The resulting classifier can then predict the turn’s emotional classification using turn-level features. Classification using word-level features is more complicated and we describe it in the following section.

## 4. TECHNIQUES FOR COMBINING WORD-LEVEL FEATURES

As one technique for combining word-level features, we use the original feature extraction scheme developed in our previous work [7]: a word-level emotion model (**WLEM**). In the training phase, the learner is given features computed at the word-level. Since our annotation is only available at the turn-level, each instance of word-level features is labeled with the class of the entire parent turn, regardless of the emotional properties of the word segment itself. In the testing phase, the class of each word segment is predicted from word-level features and these predictions are combined using majority voting to produce the class for the entire turn.

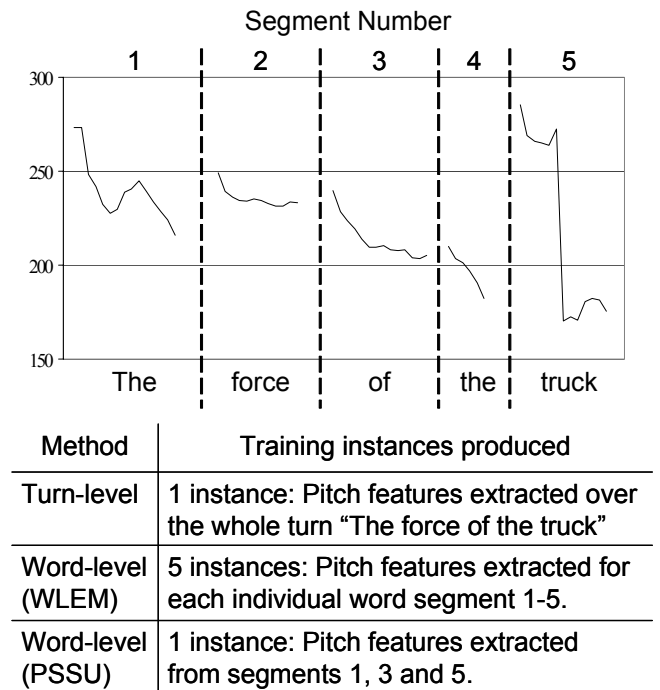
In addition, the current study explores an alternative word-level combination technique inspired from [6]. In their study, a training instance is composed of three sub-turn-level feature sets derived from the first, last and longest sub-turn. Their choice of sub-turn unit is the breath group. Here, we apply the same technique at the word level but use the middle word instead of the longest word. Since words are much smaller than breath groups, using the “middle” word ensures we capture what happens in the middle of the turn. This implementation will be called the **PSSU** (predefined subset of sub-turn units) method.

We choose to apply PSSU at the word level as opposed to the breath group level (as in [6]) due to differences in corpora type: [6] uses human-human dialogues while here we use human-computer dialogues. These differences manifest in shorter student turns in the human-computer case: 2.80 words/turn on average in our corpus (recall Table 1) as opposed to 6.11 words/turn on a subset of the corpus used in [6] (see [7] for statistics). Also, as a result of the system-initiative interaction style used in our system, our student turns are less conversational, resulting in fewer pauses that are required in differentiating breath groups. Thus, in [6] there was a greater opportunity (2.5 breath-

groups per turn on average) than in our study to utilize the multiple breath groups that PSSU is designed for. Nonetheless, in future work we would like to investigate the utility of breath group-level features on our corpus.

Please note that the WLEM method produces one instance for each word, while the PSSU method uses one instance for each turn. For an example of the different feature sets each technique produces, see the example in Figure 1.

Figure 1 – Computing turn and word-level pitch features for the student turn “The force of the truck.”



## 5. EXPERIMENTAL METHOD

To test which configuration of features lead to the most robust prediction model, we use our features to build and test a boosted decision tree provided by the Weka toolkit [9]. For all of the classification tasks, 10 trials of 10-fold cross validation experiments are performed on our corpus.

## 6. RESULTS

First, we will look at the overall performance of the three conditions. To gain more insight, we will discuss the differences between the conditions by investigating their recall and precision rates. As our baseline, we used the majority baseline: always predicting non-uncertain (77.79% accuracy).

Our previous study [7] showed that using word-level features with the WLEM method on an HC corpus yields improved prediction accuracy over turn-level features.

Table 2 displays the mean accuracy performance and standard error over these 10 trials. As the table shows, word-level features do indeed outperform turn-level ones on our larger corpus.

Table 2 – Mean accuracy (with standard error)  
Baseline: 77.79%

Turn-level	Word-level (WLEM)	Word-level (PSSU)
81.97 (0.09)	82.53 (0.07)	84.11 (0.05)

One interesting finding is that using the PSSU method for obtaining word-level features increases performance even more than the previously used WLEM method. As Table 2 shows, PSSU word-level features give a 2.14% absolute improvement in performance over turn-level features. In fact, this improvement is almost four times larger than the absolute improvement of WLEM word-level features over turn-level features (0.56%).

Since the majority of the corpus is made up of non-uncertain turns (77.79% of the corpus), uncertain turns are more difficult to predict. Overall, the classifier tends to over-predict uncertainty, as shown by Figure 2’s display of recall and precision rates for uncertain turns. In particular, different ratios of recall and precision occur among the three different techniques.

Using WLEM word-level features seems to result in the best recall performance – correctly labeling more uncertain turns than the other two methods. However, this recall performance is countered by its extremely poor precision. Thus, when using this approach to word-level feature extraction, the classifier tends to greatly over-generalize what constitutes an uncertain turn. This poor precision brings its f-measure performance down to 0.472 – the lowest of the three techniques.

Figure 2 – Recall, precision, and f-measure in predicting uncertain turns



To contrast, using PSSU word-level features yields a better recall/precision ratio. Although this implementation does not quite reach the recall levels of the WLEM method, it causes much less over-generalization by the classifier, and as a result has much higher precision. Like with WLEM features, the recall is still greater than the precision, but the ratio between the two is much closer to 1. This boosts its f-measure to 0.616, clearly making it the more reliable of the two word-level feature extraction methods.

Experiments that use turn-level features have a recall/precision ratio that is similar to the PSSU features. However, their recall, precision, and f-measure values are only about 90% as high as PSSU's values. Thus, in regards to f-measure, word-level features are once again shown to outperform turn-level features.

We hypothesize that the WLEM method is outperformed by the PSSU method because of WLEM simplifying assumptions: all words in an uncertain turn are assumed to be uncertain and the word-level predictions are combined using a simple majority voting scheme.

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, we have explored two techniques for using word-level features to improve emotion prediction over the turn-level features used in earlier work [1-3]. Our results confirm that the word-level pitch features fare better than the turn-level ones regardless of the combination technique.

Our findings suggest that word-level pitch features are more informative than turn-level pitch features, at least for emotion prediction. In our previous work we have shown that this is true for a different emotion prediction task (emotional versus non-emotional) on a small subset of our corpus. Here, we show that the same observation holds for the uncertain versus non-uncertain distinction and on a much larger corpus (220 turns versus 9,854). In addition, we show that particular attention should be paid to the word-level feature combination technique with PSSU being more robust than WLEM.

In our future work we plan on using other techniques to combine word level features. For instance, using the assumption that all words in a non-emotional turn are non-emotional, we can develop a model that could better identify emotional words than our current method of labeling every word by its parent turn's classification. Using this information, it may be possible to use a more sophisticated technique to classify a turn based on its words instead of our simple majority voting (with WLEM) or our predefined subset (with PSSU) techniques. In addition, we plan on implementing an enlarged prosodic feature set that includes amplitude and duration, as well as supplementing these features with lexical information. Combining the turn-level features and sub-turn level features obtained via the PSSU technique as in [6] is another future direction of study.

Finally, we plan to explore other techniques for selecting the words to include in the PSSU method (such as using the longest word instead of the middle word or ensuring that those chosen are content words relevant to the physics domain), which we believe will help us improve the utility of this technique even further.

## 8. ACKNOWLEDGEMENTS

This research is supported by NSF Grant No. 0328431. We thank Rebecca Hwa for helpful suggestions, and Jackson Liscombe and Julia Hirschberg at Columbia University for annotations.

## 9. REFERENCES

- [1] J. Ang, R. Dhillon, A. Krupski, A. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," Proceedings of ICSLP, 2002.
- [2] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion Detection in Task-Oriented Spoken Dialogs," Proceedings of IEEE Int. Conference on Multimedia & Expo, 2003.
- [3] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," Proceedings of ICSLP, 2002.
- [4] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters, "Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems," International Journal of Artificial Intelligence in Education, vol. 16, pp. 171-194, 2006.
- [5] D. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," Proceedings of Assoc. for Computational Linguistics (ACL), 2004.
- [6] J. Liscombe, J. Hirschberg, and J. J. Venditti, "Detecting Certainty in Spoken Tutorial Dialogues," Proceedings of Interspeech, 2005.
- [7] M. Rotaru and D. Litman, "Using Word-level Pitch Features to Better Predict Student Emotions during Spoken Tutoring Dialogues," Proceedings of Interspeech, 2005.
- [8] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to Find Trouble in Communication," Speech Communication, vol. 40 (1-2), pp.117-143, 2003.
- [9] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java implementation. San Francisco, CA: Morgan Kaufmann, 1999.
- [10] K. VanLehn, P. W. Jordan, C. P. Rose, D. Bhembe, M. Bottner, 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," Proceedings of the Intelligent Tutoring Systems Conference, 2002.