Detecting Certainness in Spoken Tutorial Dialogues

Jackson Liscombe, Julia Hirschberg, Jennifer J. Venditti

Spoken Language Processing Group
Department of Computer Science
Columbia University
New York City, NY, USA
{jaxin,julia,jjv}@cs.columbia.edu

Abstract

What role does affect play in spoken tutorial systems and is it automatically detectable? We investigated the classification of student certainness in a corpus collected for ITSPOKE, a speech-enabled Intelligent Tutorial System (ITS). Our study suggests that tutors respond to indications of student uncertainty differently from student certainty. Results of machine learning experiments indicate that acoustic-prosodic features can distinguish student certainness from other student states. A combination of acoustic-prosodic features extracted at two levels of intonational analysis — breath groups and turns — achieves 76.42% classification accuracy, a 15.8% relative improvement over baseline performance. Our results suggest that student certainness can be automatically detected and utilized to create better spoke dialog ITSs.

1. Introduction

As Intelligent Tutoring Systems (ITSs) move from text-based interactive systems to spoken dialogue systems, new avenues of exploration emerge by virtue of the rich meta-linguistic information encapsulated in human speech; among them emotion. While researchers have been studying emotion as it is manifested in isolated, acted speech for some time, interest in detecting emotion in conversational speech has emerged only in the past few years as a response to the needs of real-world systems. Emotion detection is considered an important task in a variety of applications, such as customer care centers [1, 2], task planning systems [3, 4], as well as ITSs [5]. The expression of user emotion in these contexts — emotions such as anger, frustration, or confusion — conveys important information that, if detected, could be used to improve user satisfaction.

In this paper, we examine manifestations of student *certainness* as it is expressed within the context of a spoken dialogue ITS. We investigate a spoken corpus of human-human tutorial dialogues, described in Section 2, in which student turns are annotated with certainness labels (Section 3). A few important questions to consider when looking at student certainness in spoken dialogue ITSs are whether or not human tutors use such information when tutoring students and whether or not detection of certainness aids in student learning. To address these questions, we describe in Section 4 how tutor behavior differs based on whether the student is perceived to be 'certain' or 'uncertain' about their previous statement. These turns have also been segmented into *breath groups* by procedures described in Section 5.2. We present results of automatic classification of certainness using acoustic-prosodic information, calculated both at the turn and breath group level. Section 5 enumerates the acoustic-prosodic features we extracted from the corpus dialogues at both levels, while Section 6 compares certainness classification results using different feature set partitioning. In Section 7 we discuss the implications this study has on future research in the detection of certainness – and emotion in general – in spoken dialogue ITSs.

... 9.7 min. into dialogue ...

TUTOR:	So when you apply a force what is the result of application of force on a body?
STUDENT:	The force is transferred to the container. (UNCERTAIN)
TUTOR:	No. Force does not get transferred to anything.
STUDENT:	Uh-huh. (NEUTRAL)
TUTOR:	Force is exerted and what does the force produce?
STUDENT:	Movement of the container. (UNCERTAIN)
TUTOR:	No. acceleration.
STUDENT:	Acceleration. OK. (CERTAIN)

Figure 1: A transcribed excerpt from our corpus of human-human spoken tutorial dialogues (with certainness annotation of student turns in parentheses).

2. Corpus Description

Our corpus is comprised of human-human spoken dialogues collected for the development of ITSPOKE, an intelligent tutoring spoken dialogue system in the physics domain [6]. In total, 141 dialogues from 17 subjects (7 female, 10 male) were used for our study. A dialogue consists of audio recordings of a tutoring session between a student and a tutor. Each student is first asked to type an essay in response to a physics question. The tutor and student then discuss the student's answer until the tutor determines that the student has successfully mastered the material. The student and tutor were each recorded with different microphones and each channel was manually transcribed and segmented into turns. While both the student and tutor were in the same room together, they were separated by a partition in such a way that they could not see each other. In total, our corpus contains 6778 student turns (about 400 turns per subject), each averaging 2.3 seconds in length. An excerpt of a dialogue from the corpus is shown in Figure 1.

3. Certainness Annotation

All student turns containing human speech in our corpus (6778 in total) were labeled for certainness. In particular, a student turn was annotated with one of the following labels: *uncertain, certain, neutral, mixed*. A *neutral* turn is one that is perceived to be neither *certain* nor *uncertain*, whereas a *mixed* turn is one that appears to convey both. Student turns were annotated based on the perception of the labeler. The distribution of the labels is: 64.2% *neutral*, 18.4% *certain*, 13.6% *uncertain*, 3.8% *mixed*. In this study we exclude student turns that were labeled *mixed*.

Inter-labeler agreement for this annotation was calculated using Cohen's Kappa statistic [7] on a subset of the data consisting of 505 student turns labeled for certainness by three different labelers. The average Kappa score among the three labelers was 0.52. This score is consonant with labeling agreement in spoken dialogue emotion classification [2, 3, 5]. The labels used in this study are those from a single labeler.

4. Tutor Responses to Student Certainness

In addition to the certainness annotation described in Section 3, our corpus has also been labeled with dialogue acts indicating the pragmatic effect of turns and tailored for the tutoring domain. Tutoring dialogue acts include the following: question types a tutor might ask a student (ShortAnsQ, LongAnsQ, DeepAnsQ), directives (RD), restatements or rewordings of student answers (Rst), tutor hints (Hint), tutor answers in the face of student failure (Bot), novel information (Exp), review of past arguments (Rcp), and direct positive and negative feedback (Pos, Neg). For detailed description of these tutorial dialogue acts see [8].

One of the most obvious questions concerning the perception of student certainness in tutorial domains is, do tutors respond to it? Do they change their behavior if they detect that a student is uncertain despite the fact that what they may have said is factually correct; in other words, a lucky guess? In order to explore this in our corpus we examined all the dialogue acts that directly follow *certain* and *uncertain* student turns. Table 1 lists the frequency of use of each tutor dialogue act given that the preceding student turn was *certain* or *uncertain*. Based on this evidence, we can make the following observation: The tutor uses the following techniques more frequently when a student turn is perceived to be *uncertain*: solving the problem explicitly (Bot), providing direct negative feedback (Neg), and recapping past discussion (Rcp). In addition, the tutor restates the students answer (Rst) less frequently in the face of *certain* student turns. Finally, the tutor more frequently utilizes deep reasoning type questions (DeepAnsQ) when the student is *certain*; whereas, he asks surface level questions (LongAnsQ, ShortAnsQ) more often when the student seems *uncertain*. This evidence seems to suggest that tutors do indeed respond differently to students given the perceived certainness in their utterances.

Dialogue Act	*Certain*	*Uncertain*
Bot	4.5%	6.9%
DeepAnsQ	5.6%	3.1%
LongAnsQ	1.1%	3.0%
Neg	6.7%	9.6%
Pos	23.7%	22.2%
Rcp	1.1%	2.5%
RD	1.1%	3.2%
Rst	27.1%	14.4%
SC	3.4%	6.1%
ShortAnsQ	9.3%	13.0%

Table 1: Frequency of tutor dialogue acts immediately following *certain* or *uncertain* student turns.

5. Feature Extraction

In order to characterize the acoustic-prosodic characteristics of student certainness, several sets of features were automatically extracted from the data. Feature extraction was performed both at the turn level and at the breath group level using Praat, a program for speech analysis and synthesis [9]. Feature values are represented as zscores normalized by speaker.

5.1. Turn features

In order to generalize the prosodic aspects of each student turn in its entirety, 57 acoustic-prosodic features were extracted over each student turn in the data. Turn level features were divided into two feature sets: (1) those extracted from the current turn only (**t_cur**) and (2) contextual features expressed as the relationship between the current student turn and select turns in the dialogue history (**t_cxt**).

The **t_cur** feature set includes 15 acoustic-prosodic features. The features in this set comprise:

- (5) mean absolute slope, minimum, maximum, mean, and standard deviation statistics of fundamental frequency ($f0$)
- (4) minimum, maximum, mean, and standard deviation statistics of intensity (RMS)
- (1) ratio of voiced frames to total frames in the speech signal as an approximation of speaking rate
- (4) relative position in the turn where minimum $f0$, maximum $f0$, minimum RMS, and maximum RMS occur
- (1) turn duration

The **t_cxt** feature set contains 42 features. These capture contextual information provided by the dialogue history by tracking how the student's prosody changes over time. The intuition behind features in this set is that changing acoustic-prosodic measurements may be an indication of changes in emotional state. Features in this set include features comparing the rate of change between 10 of the **t_cur** features: mean absolute slope, minimum, maximum, mean, and standard deviation of $f0$; minimum, maximum, mean, and standard deviation of RMS; and ratio of voiced frames to total frames. Ten **t_cxt** features are expressed as the rate of change between these **t_cur** features of the current student turn and those **t_cur** features of the previous student turn. Similarly, 10 additional features record the rate of change between these **t_cur** features of the current student turn and those of the first student turn in the dialogue.

Similarly, 20 features are included in the **t_cxt** feature set to monitor whether the values of these same 10 **t_cur** features have been monotonically changing over the previous 3 turns and, if so, the amount of change represented. Additional features include a record of the number of student turns preceding the current turn as well as the count of overall dialogue turns preceding the current student turn.

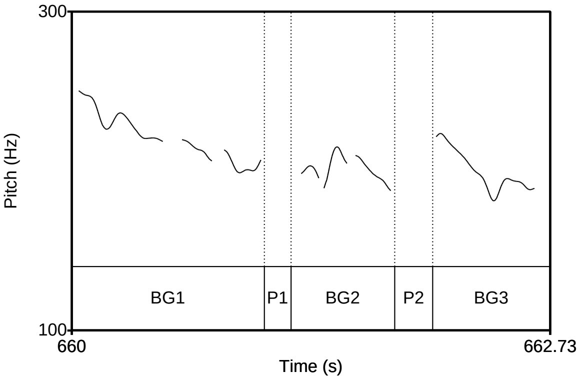

Figure 2: The output of semi-automatic breath group segmentation of a student turn plotted alongside pitch. ”BG” indicates that the segment is a breath group, while ”P” indicates a pause.

5.2. Breath group features

Student turns in our corpus can be up to a minute in length. We hypothesize that extracting acoustic-prosodic measurements over such long intervals may be less meaningful than using smaller, more prosodically coherent segmentation; namely, intonational phrases. While the automatic labeling of intonational phrase units is itself a difficult problem [10, 11], we adapted a procedure to segment audio data into breath groups[1], a segment of speech between two pauses [12], which inspection shows to roughly approximate intonational phrases. The procedure isolates contiguous segments of speech bounded by silence with a minimum length of 200 milliseconds using intensity values above a specified threshold. To predict breath groups for all the data, a statistic equal to the 75th quantile of intensity measurements over all non-student turns was calculated for each dialogue. This measurement can be thought of as the background noise estimation for each dialogue and relies on turn segmentation of the data, which in our corpus was done manually. Figure 2 shows a student turn that has been segmented into breath groups using this procedure.

To evaluate the effectiveness of the breath group identification algorithm, three dialogues, each with different subjects, were hand-annotated for breath groups. We used a program called JBeeferman, an implementation of the P_k evaluation metric for linear segmentation [13]. The average error between the three hand-labeled dialogues and those segmented with the breath group segmentation algorithm was 0.04. A score of 0.00 indicates perfect alignment. As a baseline we inserted the same number of segmentations for each dialogue at random locations in the three sample dialogues. The average error for this alignment was 0.47. We feel that this indicates our alignment is sufficiently reliable.

Semi-automatic segmentation of breath groups was calculated over all the data in the manner described above. On average, we observe 2.5 breath groups per student turn with an average length of 0.6 seconds each.

Fifteen (15) features were extracted from each student breath group in an analogous manner to how they were extracted from each student turn in Section 5.1. In fact, the features here include the same features as the **t_cur** feature set, the only difference being that they were calculated over individual breath groups instead of entire turns.

However, because we do not have certainness annotation at the level of breath groups, it would be impossible for us, at this stage, to evaluate the effectiveness of using breath groups alone in the classification of student certainness. Instead, we chose to use breath groups in a slightly different manner – as additional features at the turn level. Though, this is not without its own complications seeing as how student turns contain variable number of breath groups. Our solution was to create a feature set **bg_cur** that calculates the 15 features mentioned above for the first, last, and longest breath groups in each student turn. Thus, **bg_cur** comprises 45 features. Note that in some cases the first breath group may also be the last, or the first or last group may also be the longest. In these cases **bg_cur** will contain redundant information. For example, the student turn presented in Figure 2 would have identical values for the first and longest breath groups features in its calculation of **bg_cur**.

6. Classification Experiments

For the experiments conducted in this study we used the WEKA machine learning software package. The decision tree learner C4.5 was boosted using AdaBoost, a learning strategy that iteratively builds weak models (in this case, using C4.5 decision trees) and combines them into a significantly better model used to predict the classification of unseen data. This particular machine learning approach was used because it was found in [5] to perform well on a similar classification task using a subset of the data presented in Section 2.

Our data were randomly split into a training set consisting of 90% of the data (6100 student turns) and test set made up of the remaining 10% (687 student turns). All results presented here represent performance accuracy of AdaBoost models trained on the training set and applied to the test set. Furthermore, classification decisions are always made between the following certainness classes: *certain*, *uncertain*, and *neutral*.

Figure 2 presents classification accuracy using different combinations of the feature sets described in Section 5. For our baseline performance of 66% we used majority class classification (*neutral*). From Table 2 we can see that the **t_cur** feature set, consisting of acoustic-prosodic features calculated over the current student turn only, performs well. A performance of 74.73% is a relative improvement of 13.2% over baseline. Furthermore, while turn level contextual features (**t_cxt**) do not perform so well on their own (68.30% is only 3.5% relative improvement over baseline performance), when combined with the **t_cur** feature set the two feature sets outperform either of the two individually. Additionally, the performance of **t_cur + t_cxt** is 15.1% above baseline performance. This level of improvement over baseline performance is consistent with that found by [5] using a similar but more general classification scheme (*positive* vs. *negative* vs *other*) and turn-level acoustic-prosodic features, on a subset of the data used for this study.

Interestingly, acoustic-prosodic measurements of the first, last, and longest breath groups within a student turn (**bg_cur**) perform better than the same measurements taken over the turn in its entirety (**t_cur**). In fact, **bg_cur** exhibits performance accuracy of 75.96%, which represents a 15.1% improvement over the baseline and the same classification accuracy observed when combining both isolated and contextual turn level features. We observe the best classification accuracy (76.42%) when all the feature sets are combined.

[1] The original pause detection program can be found at http://www.helsinki.fi /˜lennes/praat-scripts/

Feature Set(s)	Accuracy
baseline	66.00 %
t_cxt	68.30 %
t_cur	74.73 %
t_cur + t_cxt	75.96 %
bg_cur	75.96 %
t_cur + t_cxt + bg_cur	76.42 %

Table 2: Classification accuracy of *certain*, *uncertain*, and *neutral* student turns using different feature sets.

7. Discussion and Future Directions

We have explored the role of student certainness in a corpus of tutorial spoken dialogues. Our annotation schema distinguishes between the perception of *certain*, *uncertain*, *mixed*, and *neutral* student turns. We have found empirical evidence that tutors respond differently to students based on their perception of the certainness of a student turn. This finding motivates the exploration of certainness detection in spoken dialogue ITS systems and suggests that such systems could be improved (or, at least, be made more human-like) by responding to this perception as humans tutors do.

From our annotated student turns we extracted acoustic-prosodic features at different levels of segmentation. Features were extracted over the current turn and also over varying degrees of context. Our results show that the addition of contextual features aids in the automatic classification of certainness. In addition, we demonstrated that breath groups — approximations of intonational phrases — can reliably be predicted using a semi-automatic algorithm. Including features defined on these units, we can increase classification accuracy above both turn-level feature sets. The highest classification accuracy we observed was 75.96%, a 15.8% relative improvement over baseline performance.

We noticed that the breath group feature set **bg_cur** performed better than the current turn feature set **turn_cur**. While this would seem to imply that breath groups are more appropriate units of analysis for emotion classification in general, the fact that all feature sets combined performed the best overall would seem to suggest that acoustic-prosodic information at both levels is useful. Turn level features may describe global prosodic activity whereas breath groups might describe what happens at certain points in the turn. Or, the relation between the two may be of interest. Relating turn features to features of component breath groups will be a subject of our future research. Similarly, including contextual features at the level of breath groups could increase the predictive power of our learning techniques still further.

Additional avenues of future research will be to annotate our corpus for certainness not only for student turns, as was done for this study, but also for each breath group. This would lend itself to a more direct comparison between acoustic-prosodic features of turns and breath groups as meaningful units of analysis. Also, the inclusion of non-acoustic-prosodic features, most notably lexical features, may also increase certainness prediction accuracy, as has been shown for other emotional states in spoken dialogue [1, 2, 3, 4, 6].

8. Acknowledgments

This research was supported in part by NSF grant IIS-0328295. We thank Diane Litman, Kate Forbes-Riley, and Scott Silliman from the Research and Development Center at the University of Pittsburgh for data collection, annotation, and discussion. We also thank Jeansun Lee for her work in annotating the data.

9. References

[1] L. Devillers and L. Vidrascu, “Reliability of lexical and prosodic cues in two real-life spoken dialog corpora,” in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, May 2004.

[2] C. M. Lee and S. Narayanan, “Towards detecting emotions in spoken dialogs,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, March 2005.

[3] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” in *Proceedings of ICSLP*, Denver, Colorado, USA, 2002, pp. 2037–2039.

[4] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, “How to find trouble in communication,” *Speech Communication*, vol. 40, no. 1-2, pp. 117–143, April 2003.

[5] K. Forbes-Riley and D. Litman, “Predicting emotion in spoken dialogue from multiple knowledge sources,” in *Proceedings of the 4th Meeting of HLT/NAACL*, Boston, MA, May 2004.

[6] D. Litman and S. Silliman, “Itspoke: An intelligent tutoring spoken dialogue system,” in *Proceedings of the 4th Meeting of HLT/NAACL* (Companion Proceedings), Boston, MA, May 2004.

[7] J. Cohen, “Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit,” *Psychological Bulletin*, vol. 70, pp. 213–220, 1968.

[8] K. Forbes-Riley, D. Litman, A. Huettner, and A. Ward, “Dialogue-learning correlations in spoken dialogue tutoring,” in *Proceedings 12th International Conference on Artificial Intelligence in Education (AIED 2005)*, Amsterdam, July 2005.

[9] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001. [Online]. Available: http://www.praat.org

[10] M. Ostendorf and K. Ross, “A multi-level model for recognition of intonation labels,” in *Computing Prosody*, Y. Sagisaka, W. N. Campbell, and N. Higuchi, Eds. Springer-Verlag, 1997.

[11] C. W. Wightman and R. C. Rose, “Evaluation of an efficient prosody labeling system for spontaneous speech utterances,” in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Keystone, Colorado, USA, December 1999.

[12] D. Hirst and A. D. Cristo, *Intonation Systems. A Survey of Twenty Languages*. Cambridge University Press, 1998, ch. A survey of intonation systems, pp. 1–44.

[13] D. Beeferman, A. Berger, and J. Lafferty, “Statistical models for text segmentation,” *Machine learning, special issue on Natural Language Processing*, vol. 34, no. 1-3, pp. 177–210, 1999, c. Cardie and R. Mooney (editors).