Correlating Student Acoustic-Prosodic Profiles with Student Learning in Spoken Tutoring Dialogues

Kate Forbes-Riley and Diane Litman

Learning Research and Development Center
University of Pittsburgh, 3939 O'Hara Street,Pittsburgh, PA, 15260, USA
{forbesk,litman}@cs.pitt.edu

Abstract

We examine correlations between student learning and student acoustic-prosodic profiles, which prior research has shown to be predictive of emotional states. We compare these correlations in two corpora of spoken tutoring dialogues: a human-human corpus and a human-computer corpus. Our results suggest that rather than relying on emotion prediction models developed via the more labor-intensive method of manually labeling emotions, adaptive strategies for our spoken dialogue tutoring system can be developed based on observed acoustic-prosodic profiles that we hypothesize to be reflective of emotion.

1. Introduction

Recent work motivated by spoken dialogue applications has started to use naturally-occurring speech to train emotion predictors [1, 2, 3, 4], for use in developing adaptive strategies to improve system performance by adapting to user emotional states. We have also experimented with predicting manually labeled student emotions in our spoken dialogue tutoring corpora using a combination of acoustic-prosodic, lexical, and dialogue features that are available to our system in real time [5, 6]. However, emotion annotation is extremely labor-intensive. Moreover, predictive results vary from system to system and inter-annotator agreement is usually low. Working in the context of automatic call-center spoken dialogue systems, [2] has shown that instead of manually labeling, predicting and adapting to some pre-defined set of user emotions, a spoken dialogue system can respond to problematic situations in spoken dialogue by detecting speech "peculiarities" that are indicative of "trouble in the communication" between user and system. Such peculiarities include automatically extractable acoustic-prosodic, lexical and dialogue features as well as manually labeled features (e.g. repetitions, hyperarticulation).

In general, the goal of system emotion adaptation is dependent on the type of adaptive system being developed. In call center applications, for example, detecting speech peculiarities indicative of user frustration or anger would enable the system to adapt by turning the caller over to a human operator. In the tutoring domain, however, *student learning* is the key evaluation measure of system performance. Prior studies in the domain of adaptive tutoring systems have investigated the association between student learning and various text-based features representing specific student behaviors by measuring the *correlation*, i.e. the strength of the positive or negative relationship, between the amount students learn and each student behavior being analyzed [7, 8]. Similarly, [9] has used correlations to identify student emotions associated with student learning. Once student states are identified that correlate with student learning, adaptive techniques can be developed to try to perpetuate those states that positively correlate with learning, and try to alter these states that negatively correlate with learning.

In this paper, we investigate whether the approach of detecting speech peculiarities is useful for identifying student emotional states that correlate with student learning. We focus on acoustic-prosodic peculiarities that are automatically extractable from the speech signal. We first extract a set of acoustic-prosodic features from each student turn in our two corpora of spoken tutoring dialogues: a human-computer corpus and a human-human corpus. We then use these extracted features to build an acoustic-prosodic profile for each student in our two corpora. Finally, we examine correlations between student learning and our student acoustic-prosodic profiles in each corpus, and we compare these correlations across our two corpora. Our results show some significant correlations between these profiles and learning, which suggests that adaptive strategies for our spoken dialogue tutoring system can be developed based in part on observed student acoustic-prosodic profiles that we hypothesize to be reflective of emotion, rather than relying on emotion prediction models developed via the labor-intensive method of manually labeling emotions.

2. Spoken dialogue tutoring corpora

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) [10] is a *speech-enabled* version of the *text-based* Why2-Atlas conceptual physics tutoring system [11]. In the ITSPOKE system, student speech is digitized from microphone input and sent to the Sphinx2 recognizer whose stochastic language models have a vocabulary of 1240 words and are trained with 7720 student utterances from evaluations of Why2-Atlas and from pilot studies of ITSPOKE. Sphinx2's most probable "transcription" (recognition output) is sent to the Why2-Atlas back-end for natural language processing (e.g. syntactic, semantic and dialogue analysis). Finally, the text response produced by the back-end is sent to the Cepstral text-to-speech system[1], then played to the student through a headphone.

Our data consists of two corpora of spoken tutoring dialogues, one with the ITSPOKE computer tutor, and the other with a human tutor performing the same task as ITSPOKE. The ITSPOKE Tutoring Corpus consists of 100 dialogues with 20 students, while the Human Tutoring Corpus contains 128 dialogues from 14 students.[2] Both corpora were collected during a prior study [12], using the following experimental procedure: University students 1) read a small document of background

[1] a commercial outgrowth of the Festival system.

[2] An ITSPOKE dialogue averages 22.7 student and 29.0 tutor turns; a Human Corpus dialogue averages 46.5 student and 43.0 tutor turns.

material, 2) took a pretest measuring their physics knowledge, 3) used a web and voice interface to work through a set of up to 10 training problems (dialogues) with the (computer or human) tutor, and 4) took a posttest similar to the pretest. In each training problem, students first typed an essay answering a qualitative physics problem; the (computer or human) tutor then engaged the student in spoken dialogue to correct misconceptions and elicit more complete explanations, after which the student revised the essay, thereby ending the tutoring or causing another round of tutoring/essay revision.

An excerpt from our ITSPOKE Corpus in Figure 1 shows both what the student said and what ITSPOKE recognized (the ASR annotations). Figure 2 shows an excerpt from our Human Corpus. Punctuation has been added to both excerpts for clarity.

. . . spoken dialogue excerpt at 5.7 min. into session. . .
ITSPOKE$_8$: According to Newton's Second Law, if an object has a non-zero net force acting on it, what will happen to it?
STUDENT$_9$: It will stay still (ASR: it will stay still)
ITSPOKE$_{10}$: Well, to compute the acceleration of a box, one property that you need to know is its mass. What else do you need to know to find the box's acceleration?
STUDENT$_{11}$: The direction (ASR: add directions)
ITSPOKE$_{12}$: If you see a body accelerate, what caused that acceleration?
STUDENT$_{13}$: Force (ASR: force)
ITSPOKE$_{14}$: Good job. Say there is only one force acting on the box. How is this force, the box's mass, and its acceleration related?
STUDENT$_{15}$: Velocity (ASR: velocity)

Figure 1: ITSPOKE Tutoring Corpus Dialogue Excerpt

. . . spoken dialogue excerpt at 13.6 min. into session. . .
TUTOR$_{59}$: If you take two seconds for example, during these two seconds at every step their velocity will be same. So the distance covered by them will be the same and therefore there is no relative displacement between the keys and the face.
STUDENT$_{60}$: So the keys will hover?
TUTOR$_{61}$: Yes, the keys will appear to hover in front of the face. As they both fall there will be no relative displacement between them and the reason is- what are the reasons?
STUDENT$_{62}$: That they have the same initial velocity and the same acceleration.
TUTOR$_{63}$: Same acceleration- and why they have same acceleration?
STUDENT$_{64}$: 'Cause they're in freefall.

Figure 2: Human Tutoring Corpus Dialogue Excerpt

3. Extracting acoustic-prosodic features from student turns

As discussed in Section 1, the acoustic-prosodic features of speech, alone or in combination with lexical and dialogue features, have been shown to be useful predictors of manually labeled user emotional states in spoken dialogue systems. However, manual emotion labeling is labor-intensive and subjective. Thus [2] has suggested that a spoken dialogue system can detect and respond directly to the speech features (called "peculiarities"), which are interpreted as indicating states of "trouble in communication" between user and system.

In this paper, we investigate whether this approach of detecting speech peculiarities is useful for identifying student emotional states that correlate with student learning in our spoken tutoring dialogues. We focus on the 24 *turn-based* acoustic-prosodic features itemized in Figure 3 below, which we used previously to predict manually labeled student emotions in our two corpora [5, 6][3] and are also motivated by the emotion research cited in Section 1. A real benefit of these features is that they can be computed automatically and in real-time, and thus could be used to trigger online adaptation in ITSPOKE.

Max$_T$ **F0**: maximum F0 value in turn
Min$_T$ **F0**: minimum F0 value in turn
Ave$_T$ **F0**: average F0 value in turn
STD$_T$ **F0**: standard deviation of F0 values in turn
Max$_T$ **RMS**: maximum RMS value in turn
Min$_T$ **RMS**: minimum RMS value in turn
Ave$_T$ **RMS**: average RMS value in turn
STD$_T$ **RMS**: standard deviation of RMS values in turn
IntSilence$_T$: amount of internal silence in turn
Duration$_T$: duration of turn (seconds)
PrePause$_T$: duration of pause preceding turn (seconds)
Tempo$_T$: tempo (speaking rate) of turn
Norm. <feature>$_T$ each of the above 12 features normalized by the first student turn

Figure 3: 24 Acoustic-Prosodic Features per Student Turn

From each student *turn* in our two corpora, we extracted the 12 raw acoustic-prosodic features shown first in Figure 3. A normalized value for each of these 12 features (shown last in Figure 3) was then computed by dividing the value of each feature for the current student turn by the value of that feature for the first student turn. This normalization has the benefit of measuring change across student turns rather than absolute value, and it also removes gender dependency of the pitch features.

F0 and RMS values, representing pitch and loudness, respectively, are computed via Entropic Research Laboratory's pitch tracker, *get_f0*, with no post-correction.[4] Internal silence is approximated as the proportion of zero f0 frames for the turn. Duration and PrePause are computed via the start and end turn boundaries of the student turns. Tempo is computed based on the number of syllables in the turn transcription divided by the turn duration. The "*T*" subscript emphasizes that these features are all computed for each *turn* in the two corpora.

4. Correlating student acoustic-prosodic profiles and learning

4.1. Correlation methodology

We next created an acoustic-prosodic profile for each student, to investigate whether any of the features in the profile correlated with learning on a *per-student* basis, i.e., across all the students in each corpora. To create these acoustic-prosodic profiles, we computed, for each student, a single maximum, minimum, average, and standard deviation across all of his/her turns, for each of the 24 per-turn features listed in Figure 3. In addition, for each student, we computed a sum (by summing the values across all turns) for 6 of the per-turn temporal features: IntSilence, Duration, and PrePause (both raw and normalized). This yielded an acoustic-prosodic profile for each student that

[3] We use "emotion" to cover affects and attitudes that can impact student learning; acoustic-prosodic features reflect *expressions* of emotion.

[4] RMS (root mean squared amplitude) values are based on a 0.03 second window within frame steps of 0.01 seconds.

contained a total of 102 acoustic-prosodic feature values (24 x 4 + 6). Note that we label these *per-student* features with the subscript "*S*", to distinguish them from the *per-turn* features, which are subscripted with "*T*".

To exemplify how an acoustic-prosodic profile is produced, suppose the dialogue in Figure 1 constituted our entire ITSPOKE corpus. Each student turn in Figure 1 has its own value for each of the 24 features in Figure 3. For example, **STUDENT**$_9$ has a value for Max$_T$ F0, and **STUDENT**$_{11}$ has a different value for Max$_T$ F0. To get our acoustic-prosodic profile for this student, we compute the average, maximum, minimum, and standard deviation for each of the 24 features across all the turns of this student. For example, if the single Max$_T$ F0 values for **STUDENT**$_9$ - **STUDENT**$_{15}$ were respectively: 200Hz, 210Hz, 220Hz, 230Hz, then the acoustic-profile for this student would contain the following: Ave$_S$ Max$_T$ F0 = 215Hz, Max$_S$ Max$_T$ F0 = 230Hz, Min$_S$ Max$_T$ F0 = 200Hz, STD$_S$ Max$_T$ F0 = 12.9Hz.

Next, for each feature in the acoustic-prosodic profile, we computed a Pearson's Correlation between the feature value and posttest score.[5] However, because the pretest and posttest scores were already significantly correlated in both the Human Tutoring (R=.72, p =.01) and ITSPOKE Tutoring corpora (R=.46, p=.04), we controlled for the effect of pretest score by first regressing it out of the correlation, thus following the same correlation methodology as used in our prior work [12].

The Human Tutoring student means for the multiple-choice pre- and posttests were 0.42 and 0.72, respectively, and the ITSPOKE Tutoring student means were 0.48 and 0.69. These means reflect the percent of correctly answered questions on these tests. Clearly, students learned significantly in both conditions; the goal of this study is to investigate whether student acoustic-prosodic profiles correlate with these learning gains.

4.2. ITSPOKE Tutoring Corpus results

Table 1 presents our best results on the correlations of our 102 raw acoustic-prosodic features with learning in our ITSPOKE Tutoring Corpus, i.e. results where the correlation was significant ($p \leq .05$) or a trend ($p \leq .1$), after regressing out pretest. The first two columns list the acoustic-prosodic feature and its mean and standard deviation (across all students), while the last two columns show the Pearson's correlation between posttest and the feature after regressing out pretest.

Table 1 shows first that numerous Tempo features are predictive of student learning in our ITSPOKE Corpus. The trend for a positive correlation between Max$_S$ Tempo$_T$ and learning relates increased maximum speaking rate in student turns with increased learning; however, this is the only positive result found. The trend for a correlation between Min$_S$ Norm. Tempo$_T$ and learning, and the significant correlation between STD$_S$ Norm. Tempo$_T$ and learning, are both negative. These results relate increased minimum speaking rates and increased variation in speaking rates with decreased learning.[6]

[5] A Pearson's Correlation measures the strength of the linear relationship between two variables. Pearson's Correlation Coefficient is signified by R, and can take on the values from -1.0 to 1.0, where -1.0 is a perfect negative (inverse) correlation, 0.0 is no correlation, and 1.0 is a perfect positive correlation. The statistical significance of R is tested using a t-test. A low p-value for this test means that there is a statistically significant relationship between the two variables.

[6] "Increased/decreased learning" means that the correlation relates increases/decreases in posttest score to increases in the feature value. The higher the R value, the stronger this correlation.

Table 1 also shows that numerous raw Pitch features of student turns are predictive of decreased student learning in our ITSPOKE corpus. In particular, there are trends for negative correlations with learning for Ave$_S$ Max$_T$ F0 and well as for Max$_S$, Ave$_S$, and STD$_S$ of Min$_T$ F0. These results all relate increased maximum and minimum raw pitch values in student turns with decreased student learning. There is also a significant negative correlation with learning for Ave$_S$ Ave$_T$ F0. This result relates increased average raw pitch values in student turns with decreased student learning. However, because these are all correlations between non-normalized pitch values and learning, and females generally produce higher pitch values than males, further analysis is required to discount the influence of any correlation between gender and learning.

Table 1: ITSPOKE Corpus: Trends and Significant Correlations between Student Acoustic-Prosodic Profiles and Learning

AP Feature	Mean (SD)	R	p
Max$_S$ Tempo$_T$	.69 (.18)	.40	.09
Min$_S$ Norm. Tempo$_T$	.09 (.13)	-.44	.06
STD$_S$ Norm. Tempo$_T$	.97 (.35)	-.47	.04
Ave$_S$ Max$_T$ F0	223.1 (54.8)	-.45	.05
Max$_S$ Min$_T$ F0	206.9 (57.2)	-.44	.06
Ave$_S$ Min$_T$ F0	133.4 (33.8)	-.45	.05
STD$_S$ Min$_T$ F0	39.6 (14.2)	-.42	.07
Ave$_S$ Ave$_T$ F0	176.1 (43.0)	-.48	.04

4.3. Human Tutoring Corpus results

Table 2 presents our best results on the correlations of our 102 acoustic-prosodic features with learning in our Human Tutoring Corpus, i.e. results where the correlation was significant ($p \leq .05$) or a trend ($p \leq .1$), after regressing out pretest.

Table 2: Human Corpus: Trends and Significant Correlations between Student Acoustic-Prosodic Profiles and Learning

AP Feature	Mean (SD)	R	p
Max$_S$ Duration$_T$	32.3 (9.9)	.47	.10
Sum$_S$ Norm. Duration$_T$	1817.3 (1267.1)	-.48	.09
Ave$_S$ Norm. Duration$_T$	4.1 (1.9)	-.60	.03
Ave$_S$ Tempo$_T$	4.2 (.37)	-.50	.08

Table 2 shows first that numerous Duration features of student turns are predictive of student learning in our Human Tutoring Corpus. The trend for a positive correlation between Max$_S$ Duration$_T$ and learning relates increased turn length with increased learning; however, this is the only positive result found. The trend for a correlation between Sum$_S$ Norm. Duration$_T$ and learning, and the significant correlation between Ave$_S$ Norm. Duration$_T$ and learning, are both negative. These results relate increases in the total time a student speaks and in average student turn length with decreased learning.

Table 2 also shows a trend for a negative correlation between Ave$_S$ Tempo$_T$ of student turns and learning in our Human Corpus. This result relates increased average student speaking rate in student turns with decreased student learning.

4.4. Cross-Corpora comparison

Our Human Tutoring Corpus represents an upper bound for the speech and natural language processing capabilities of our IT-

SPOKE Tutoring Corpus. As such, cross-corpora differences in how the acoustic-prosodic profiles relate to student learning can shed light on how system improvements might positively impact learning. To a large extent, we find that the acoustic-prosodic features that are predictive of student learning differ across the two corpora. First, in the Human corpus, Duration features are predictive of both increased and decreased learning, while in the ITSPOKE corpus, they are not predictive of learning at all. Interestingly, the ITSPOKE corpus also shows less variation in turn durations than the Human corpus. For example, in the ITSPOKE corpus, where the average turn duration is 3.45 seconds, the average maximum turn duration is 8.94 seconds, while in the Human corpus, where the average turn duration is 2.28 seconds, the average maximum turn duration is 32.3 seconds, and this feature is predictive of increased learning. Visual inspection of the student turns in the Human corpus with above average maximum turn Durations shows that often in these turns the student is reasoning through the problem, as exemplified below. Previous research has shown that learning positively correlates with students' construction of knowledge [13], such as occurs through reasoning. Future system versions will encourage increased student reasoning via more open-ended questioning.

TUTOR: Your answer is right but can you give some reasons?
STUDENT: Ok, uh the balls they'll hit the ground at the same time because uh the weight uh or the uh- oh man- the- if the ball's twice as big then the force acting on it will be twice as bibig twice as much uh towards the ground than and if the ball's half as big as the one ball the ball will have half as much uh gravitational force on it. Does that sound right?

Second, although Tempo features are predictive of learning in both corpora, the specific Tempo features that predict learning differ, and in the ITSPOKE corpus, we find that Tempo features relate to both increased and decreased learning, while in the Human corpus, they relate only to decreased student learning. Moreover, in the ITSPOKE corpus, Pitch features are predictive of decreased student learning, while in the Human corpus they are not predictive of learning at all. Overall, a better understanding of how student acoustic-prosodic profiles relate to learning across our two corpora requires further analysis. However, that most of the correlations found in this study are negative suggests that in both of our corpora, acoustic-prosodic signatures are most indicative of "trouble in communication", i.e. places in the dialogue where things start to go wrong.

5. Conclusions and current directions

We examine correlations between student learning and observed student acoustic-prosodic profiles that we hypothesize are predictive of emotional states. We compare our results across a computer tutoring corpus and a human tutoring corpus. Although we found significant correlations and trends in both corpora, the results for specific features differed. This suggests the importance of training emotion predictors from appropriate data. Overall, our results provide preliminary evidence that instead of relying on predictive models based on manually labeled emotions, adaptive techniques for our spoken dialogue tutoring system can be based directly on student profiles containing features that we hypothesize to be reflective of emotions and that correlate with student learning. We are currently working on adding additional acoustic-prosodic and lexical and dialogue features to our student profiles, focusing on automatically extractable features such as verbatim repetitions, swear words, and approximations of rising intonation. Our larger goal is to incorporate a predictive model based on these profiles into our system so that student states that correlate with learning can be recognized and adapted to. We can then compare the performance of this adaptive system with the non-adaptive version.

6. Acknowledgements

This research is supported by NSF (#0328431).

7. References

[1] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003, pp. 31–36.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth, "How to find trouble in communication," *Speech Communication*, vol. 40, pp. 117–143, 2003.

[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 the International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 203–207.

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

[5] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proc. Assoc. for Computational Linguistics*, 2004, pp. 352–359.

[6] K. Forbes-Riley and D. Litman, "Predicting emotion in spoken dialogue from multiple knowledge sources," in *Proc. of the Human Language Technology Conf.: 4th Meeting of the N. American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL)*, 2004, pp. 201–208.

[7] M. G. Core, J. D. Moore, and C. Zinn, "The role of initiative in tutorial dialogue," in *Proc. European Chap. Assoc. Computational Linguistics*, 2003.

[8] G. Jackson, N. Person, and A. Graesser, "Adaptive tutorial dialogue in AutoTutor," in *Proc. Workshop on Dialog-based Intelligent Tutoring Systems at ITS'04*, 2004.

[9] S. D. Craig, A. C. Graesser, J. Sullins, and B. Gholson, "Affect and learning: an exploratory look into the role of affect in learning with AutoTutor," *Journal of Educational Media*, vol. 29, no. 3, pp. 241–250, 2004.

[10] D. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system," in *Proc. of the Human Language Technology Conf.: 4th Meeting of the North American Chap. of the Assoc. for Computational Linguistics (HLT/NAACL) (Companion Vol.)*, 2004, pp. 233–236.

[11] K. VanLehn, P. W. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson, "The architecture of Why2-Atlas: A coach for qualitative physics essay writing," in *Proc. Internat. Intelligent Tutoring Systems Conf. (ITS)*, 2002, pp. 158–167.

[12] D. Litman, C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman, "Spoken versus typed human and computer dialogue tutoring," in *Proc. Internat. Intelligent Tutoring Systems Conf.*, 2004, pp. 368–379.

[13] M. T. H. Chi, S. A. Siler, H. Jeong, T. Yamauchi, and R. G. Hausmann, "Learning from human tutoring," *Cognitive Science*, vol. 25, pp. 471–533, 2001.