Speech Recognition Performance and Learning in Spoken Dialogue Tutoring

Diane Litman and Kate Forbes-Riley

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

Abstract

Speech recognition errors have been shown to negatively correlate with user satisfaction in evaluations of task-oriented spoken dialogue systems. In the domain of tutorial dialogue systems, however, where the primary evaluation metric is student learning, there has been little investigation of whether speech recognition errors also negatively correlate with learning. In this paper we examine correlations between student learning and automatic speech recognition performance, in a corpus of dialogues collected with an intelligent tutoring spoken dialogue system. We examine numerous quantitative measures of speech recognition error, including rejection versus misrecognition errors, word versus sentence-level errors, and transcription versus semantic errors. Our results show that although many of our students experience problems with speech recognition, none of our measures negatively correlates with student learning.

1. Introduction

While intelligent tutoring has become a domain of increasing interest for dialogue systems research, most current tutorial dialogue systems are still text-based [1, 2]. Although recent studies suggest that *speech-based* tutorial dialogue systems have the potential of being even more effective for student learning[1] [3, 4], and several projects have begun to incorporate spoken dialogue capabilities into their systems [5, 6, 7], there is still a feeling in the intelligent tutoring systems community that automatic speech recognition (ASR) is just not yet up to the task, despite very little empirical investigation of this issue. In particular, while there have been a number of studies showing that ASR errors negatively correlate with user satisfaction in evaluations of spoken dialogue systems [8, 9], there has been little investigation of whether ASR errors negatively correlate with *student learning*, the primary evaluation metric in intelligent tutoring dialogue systems.

To assess the impact of adding spoken language capabilities to dialogue tutoring systems, we have built ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) [7], a spoken dialogue tutor in the domain of conceptual physics. In a first evaluation of ITSPOKE, we found that students learned a significant amount after their spoken dialogue tutoring sessions [4]. Furthermore, a pilot analysis of our data suggested that ASR errors did not negatively correlate with learning.

In this paper we perform a much more extensive analysis of our data, to fully investigate whether ASR performance correlates with learning. We look at numerous ways of measuring ASR performance, including rejection versus misrecognition errors, word versus sentence-level errors, and transcription versus semantic errors. Our results show that although many of our students experience ASR problems, not a single one of our measures negatively correlates with student learning. In contrast, certain types of ASR errors do negatively correlate with dialogue efficiency. Our results suggest that even though current ASR technology is still far from perfect, spoken dialogue systems can indeed effectively support student learning.

[1] [3] found that student knowledge construction (a type of student activity known to correlate with learning) could be increased by using speech, while [4] found that in *human* tutoring, spoken rather than typed "chat" dialogue caused significant improvements in learning.

2. Spoken tutoring system and corpus

ITSPOKE [7] is a *speech-enabled* version of the *text-based* Why2-Atlas conceptual physics tutoring system [10]. When interacting with ITSPOKE, students first type 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. During the dialogue, 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. Next, Sphinx2's most probable "transcription" (one-best 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, then played to the student through a headphone.

The corpus analyzed in this paper consists of 100 spoken tutoring dialogues between 20 students and ITSPOKE. The corpus was collected during a prior study [4], using the following experimental procedure (taking approximately 4-5 hours per student): university students 1) read a small document of background material, 2) took a pretest measuring their physics knowledge, 3) used a web and voice interface to work through a set of 5 training problems (dialogues) with ITSPOKE, and 4) took a posttest similar to the pretest. Because speech recognition is imperfect, after the data was collected, each student utterance in our corpus was manually transcribed by a project staff member. Annotated (Section 3) excerpts from our corpus are shown in Figures 1 and 2 (punctuation added for clarity). The excerpts show both what the student said and what ITSPOKE recognized (the ASR labels). On average, our dialogues contain 22 student turns and 29 tutor turns.

3. Measuring ASR performance

We use the data discussed in Section 2 to compute several types of measures relevant for ASR performance evaluation, as shown in Figure 3. Informally, misrecognitions occur when the best

. . . spoken dialogue excerpt at 8.6 min. into session. . .

ITSPOKE$_1$: Because gravity is the only force, it is trivial to determine the net force (i.e., the vector sum of all forces). So what direction does the net force have?

STUDENT$_2$: downward

ASR$_2$: downward (*WER = 0%, ASR MIS* = FALSE*)*

ITSPOKE$_3$: Yeah. Does the packet have an acceleration? If yes, please specify its direction.

STUDENT$_4$: yes downward

ASR$_4$: is downward *(WER = 50%, ASR MIS* = TRUE*)*

ITSPOKE$_5$: Uh huh. Let's look now at the velocity of the packet. In general, for two dimensional motion it is useful to analyze the velocity of an object in terms of its horizontal component and its vertical component separately. So let's begin by analyzing the vertical component of the packet's velocity after the release. How would you describe the vertical component of the packet's velocity? (e.g., decreasing, zero, etc.)

STUDENT$_6$: increasing

ASR$_6$: decreasing *(WER = 100%, ASR MIS* = TRUE*)*

Figure 1: Excerpt illustrating Misrecognitions

hypothesis that is output by ITSPOKE's automatic speech recognizer (ASR) differs from what the student actually says (as in Figure 1), and rejections occur when ITSPOKE does not accept the ASR output (as in Figure 2).

More precisely, with respect to misrecognitions, we first compute the word error rate (**WER**) for each turn, by comparing the manual transcription and the ASR output using the SCLITE scoring algorithm from the NIST Scoring Toolkit Version 0.1. In addition, we compute a boolean version of WER called **ASR MIS**, which is TRUE when any portion of the turn is misrecognized (WER $>$ 0), and FALSE when the ASR output exactly corresponds to the transcription (WER=0). Figure 1 shows the transcription, the one-best ASR output, and the resulting values for our continuous and boolean word error measures, for three of the student utterances in our corpus.

Although such word-based measures are a traditional way to evaluate speech recognizers, semantic measures are often used instead in dialogue evaluations, as they do not penalize for word errors that are unimportant to overall utterance interpretation. We compute a semantic notion of misrecognition for each student turn in our corpus using the following procedure. First, we apply ITSPOKE's natural language understanding (NLU) component in batch mode to both the manual transcription and the ASR output. We then score the semantic results as in the word-level analysis. ITSPOKE's NLU component is provided by the Why2-Atlas back-end, and associates a semantic grammar with each tutor question (grammars across questions may share rules). The categories in the grammar correspond to the expected responses for the question, and represent both correct answers and typical student misconceptions [10]. Given a student's utterance, the output of the NLU component is thus a subset of the semantic concepts that were expected as answers to the tutor's prior question, and that were found when parsing the student's utterance. For instance, the semantic concept *downward* is used in many of the semantic grammars, and would be the NLU output for a variety of utterances such as "down-

. . . spoken dialogue excerpt at 23.1 min. into session . . .

ITSPOKE$_1$: The direction of the gravitational force acting on the apple is vertically down, or toward the center of the earth. Therefore, what is the magnitude of this gravitational force in the horizontal direction?

STUDENT$_2$: significant

ASR$_2$: significant (*REJ* = TRUE, *TRUE REJ* = FALSE, *FALSE REJ* = TRUE)

ITSPOKE$_3$: Could you please repeat that?

STUDENT$_4$: great

ASR$_4$: crate (*REJ* = TRUE, *TRUE REJ* = TRUE, *FALSE REJ* = FALSE)

ITSPOKE$_5$: I'm sorry, I'm having trouble understanding you. Please try again.

Figure 2: Excerpt illustrating Rejections

- Misrecognitions
  - **WER**: word error rate
  - **ASR MIS**: boolean word error
  - **SER**: semantic error rate
  - **SEM MIS**: boolean semantic error
- Rejections
  - **TRUE REJ**: boolean true rejection
  - **FALSE REJ**: boolean false rejection
  - **REJ**: **TRUE REJ** $\vee$ **FALSE REJ**
- Problems (Misrecognitions or Rejections)
  - **ASR PROB**: **ASR MIS** $\vee$ **REJ**
  - **SEM PROB**: **SEM MIS** $\vee$ **REJ**

Figure 3: Turn-level ASR Performance Measures

wards", "towards earth", "is it downwards", "down", etc.

Consider how this impacts our scoring of the two misrecognized student utterances shown in Figure 1:

- STUDENT$_4$ versus ASR$_4$
  NLU("yes downward"): *downward*
  NLU("is downward"): *downward*
  SER = 0%, SEM MIS = FALSE
- STUDENT$_6$ versus ASR$_6$
  NLU("increasing"): *increase*
  NLU("decreasing"): *decrease*
  SER = 100%, SEM MIS = TRUE

Although there is a word error when comparing STUDENT$_4$ and ASR$_4$, the semantic analyses are identical (namely, both the transcription and ASR output are mapped to a single semantic concept (*downward*))[2]. From this semantic perspective, ASR$_4$ is no longer considered an error because ITSPOKE will behave the same as if ASR$_4$ had exactly matched STUDENT$_4$. Thus,

[2]Despite the system's yes/no question, *yes* was not one of the semantic concepts that the system was looking for.

while ASR MIS = TRUE, SEM MIS = FALSE. On the other hand, differences between the semantic analysis of the transcription and the ASR output for $STUDENT_6$ cause this turn to remain a misrecognition, even at the semantic level (that is, ASR MIS = TRUE and SEM MIS = TRUE). In particular, the different semantic interpretations will impact ITSPOKE's response: the error will cause ITSPOKE to begin a subdialogue to correct a falsely perceived wrong answer.

We now turn to cases where ITSPOKE has low confidence in the ASR output and rejects a student's utterance. As shown by the two examples in Figure 2, after such rejections (**REJ**) ITSPOKE asks the student to repeat the last utterance. When measuring this type of system error, it is sometimes useful to distinguish between two types of Rejections. In a true rejection (**TRUE REJ**), the ASR output for the student turn *is* different from what the student said (Figure 2, $STUDENT_4$). In contrast, for a false rejection (**FALSE REJ**), the ASR output *is not* different from what the student said (Figure 2, $STUDENT_2$).

Finally, our last error category classifies an ASR output as problematic if it yielded either a misrecognition or a rejection. This allows us to explore whether the cumulative effect of different types of ASR problems negatively correlates with learning (even if there are no correlations with the individual error types). **ASR PROB** combines transcription misrecognitions and total rejections, while **SEM PROB** combines semantic misrecognitions and total rejections.

4. Correlation methodology

As discussed in Section 1, in our prior work we demonstrated that students learned a significant amount after being tutored by ITSPOKE. Furthermore, in a pilot study based on two of the measures in Figure 3 (namely WER and SER), we found no negative correlation between either measure and how much students learned [4].[3] Here we revisit the question of whether speech performance errors correlate with learning in our spoken tutoring dialogues, but extend our previous analysis to include a much more comprehensive set of measures derived from the boolean measures enumerated in Figure 3. We call these measures our *Speech Performance Measures*, and compute them using the following procedure.

First, we computed the value of each boolean error measure, for *each* of the 2354 *student turns* in our corpus, as explained above. Next, for *each student* and *each boolean measure*, we computed a **total (#)** and a **percentage (%)** representing how frequently that type of error occurred, across *all* of the dialogues with that student. Each Total was computed by counting the number of times the measure's value was TRUE, across all of the student's turns. Each Percentage was computed by dividing the Total for the measure by the total number of turns for the student, and can be viewed as a normalized version of the raw totals. We further computed two **ratios (/)** of totals: the ratio of Semantic Misrecognitions to ASR Misrecognitions (S/A MIS), and the ratio of Semantic Problems to ASR Problems (S/A PROB). If the excerpt in Figure 1 constituted all of the dialogue data for this particular student, the following would be the values for the (student-level) Speech Performance Measures corresponding to the (turn-level) measures ASR PROB and SEM PROB: # ASR PROB = 2, % ASR PROB = .66, # SEM PROB = 1, % SEM PROB = .33, and S/A PROB = .5.

In addition, motivated by prior studies of learning correlations [11], we fit regression lines to the occurrences of five of the turn-level measures across the chronologically-ordered dialogues of each student: ASR MIS, SEM MIS, REJ, ASR PROB, and SEM PROB. This yielded a **slope (S)** and **intercept (I)** per student for each of these types of errors, which allowed us to compare where these errors were occurring for each student's interaction with the system. E.g., a high intercept and decreasing slope for ASR MIS for a student would indicate that most of the ASR Misrecognitions occurred early on during that student's interaction with the system. This could occur if the student adapted his/her speaking style and thereby improved the system understanding. A low intercept and rising slope for ASR MIS would indicate that ASR Misrecognitions increased throughout that student's interaction with the system. This could occur if the student reacted to system misunderstanding with frustration and decreased effort in speaking clearly.

Finally, for each of these Speech Performance Measures (totals, percentages, ratios, slopes and intercepts), we computed a Pearson's correlation between the measure and posttest score, across all of the students in our corpus. However, because the pretest and posttest scores were significantly correlated in our corpus (R=.46, p=.04), we controlled for pretest score by regressing it out of the correlation, as in other studies of correlations with student learning [11, 4].[4] We also computed a Pearson's correlation between the measure and the total time that the student interacted with the system (computed by summing the total dialogue times across all dialogues of a student). While learning is the primary evaluation measure used in the tutoring community, other analyses such as dialogue efficiency are also of interest. Again, because the pretest and total interaction time were significantly correlated in our corpus (R= -.55, p=.01), we controlled for pretest by regressing it out of the correlation.

5. Results and discussion

Table 1 presents our results on the correlations of our Speech Performance (SP) Measures with both learning and total interaction time in our corpus. The first column lists the measure (total (#), percentage (%), ratio (/), intercept (I), or slope (S) of the speech recognition performance error per student). The second column shows the mean and standard deviation (across all students). The third and fourth column show the Pearson's correlation between posttest and the measure after the correlation with pretest is regressed out. The last two columns present the Pearson's correlation between total interaction time and the measure after the correlation with pretest is regressed out. For example, the third row shows that there are 32.3 total turns containing an ASR misrecognition per student on average, and that there is no correlation between # ASR MIS with learning (p=.26), but there is a significant (p=.05) positive correlation (R = .46) between # ASR MIS and total interaction time.

Bolded correlations indicate that the correlation was statistically significant ($p \leq .05$) or a trend ($p \leq .1$). As the table shows, there were *no* significant correlations or trends for a correlation between any of our Speech Performance Measures and learning. Although this result is contrary to our predictions, it is a welcome result, and generalizes our previous findings.

However, we do see two positive correlations between Speech Performance Measures and total interaction time. The correlation for ASR misrecognitions is somewhat mysterious, as these errors alone do not impact how the system responds to the student. However, perhaps a student's awareness of such er-

[3] The details were not published, so are included in Table 1.

[4] The student means for the pre- and posttests (each consisting of 40 multiple choice physics questions) were 0.48 and 0.69, respectively.

rors impacted how the student responded to the computer (the dialogue history was displayed on the screen, so the student was in fact potentially aware of ASR misrecognitions). The trend for ASR problems to correlate with total interaction time may in part reflect the fact that Rejections require the student to repeat his/herself; however we see no correlation between Rejections alone and interaction time.

Table 1: Speech Performance Measure Correlations with Student Learning and Interaction Time (20 students)

		Learning		Time	
SP Measure	Mean	R	p	R	p
WER	31.3 (8.9)	-.20	.41	-.14	.56
SER	7.0 (4.0)	-.11	.65	-.03	.90
# ASR MIS	32.3 (11.0)	.27	.26	**.46**	**.05**
# SEM MIS	6.7 (4.0)	-.05	.83	.25	.29
# TRUE REJ	6.5 (5.7)	-.22	.36	.18	.46
# FALSE REJ	1.7 (2.1)	-.26	.28	.27	.26
# REJ	8.2 (7.5)	-.24	.31	.21	.38
# ASR PROB	40.5 (16.3)	.06	.82	**.41**	**.08**
# SEM PROB	14.9 (10.6)	-.19	.43	.25	.31
% ASR MIS	28.2 (7.7)	.08	.73	-.20	.42
% SEM MIS	5.7 (2.6)	-.10	.68	-.04	.88
% TRUE REJ	5.1 (3.6)	-.21	.39	-.03	.90
% FALSE REJ	1.1 (1.2)	-.33	.17	.14	.58
% REJ	6.2 (4.1)	-.27	.27	.01	.97
% ASR PROB	34.5 (7.8)	-.06	.81	-.19	.44
% SEM PROB	11.9 (5.6)	-.24	.32	-.01	.97
S/A MIS	.20 (.09)	-.09	.71	.10	.68
S/A PROB	.34 (.14)	-.20	.42	.11	.66
I ASR MIS	.28 (.10)	.17	.48	-.12	.63
I SEM MIS	.03 (.04)	-.18	.46	.26	.28
I REJ	.06 (.10)	-.33	.17	-.14	.58
I ASR PROB	.34 (.14)	-.08	.74	-.17	.48
I SEM PROB	.09 (.13)	-.31	.20	-.00	.99
S ASR MIS	-.00 (.00)	-.04	.88	-.01	.96
S SEM MIS	.00 (.00)	.07	.79	-.35	.14
S REJ	.00 (.00)	.25	.30	.15	.53
S ASR PROB	.00 (.00)	.13	.60	.08	.73
S SEM PROB	.00 (.00)	.23	.34	-.06	.82

6. Related work

A preliminary evaluation of student interactions with the SCoT spoken dialogue tutoring system (in the domain of shipboard damage control) has achieved results that complement our findings [6]. While the SCoT analysis was based on only a few ASR statistics (none of which were semantic), two different measures of learning were examined: performance on a written test (as in our ITSPOKE study), and performance in a simulator. Also, their analysis showed that while ASR error did not correlate with learning, it did negatively correlate with a student's desire to use the system again (measured via a questionnaire item).

7. Conclusions and current directions

We computed correlations between student learning and numerous ways of measuring automatic speech recognition performance, in a corpus of spoken dialogues between students and the ITSPOKE dialogue tutoring system. Overall, our results indicate that although there are plenty of speech recognition errors in our system, these problems do not negatively correlate with how much students learn. We conclude that contrary to the fear of many in the tutoring community, ASR is feasible to use in spoken dialogue applications where learning is the primary goal. However, our results do suggest that ASR problems may be predictive of secondary factors relevant to tutoring dialogue system evaluations, such as total interaction time. We are in the process of conducting a new evaluation of ITSPOKE, which will quadruple the size of our current corpus. We have also added a user satisfaction survey to further study interactions between ASR, learning and other evaluation factors.

8. Acknowledgements

We thank the ITSPOKE group. This research is supported by ONR (N00014-04-1-0108) and NSF (#0325054).

9. References

[1] V. Aleven and C. P. Rose, Eds., *Proceedings of the AIED 2003 Workshop on Tutorial Dialogue Systems: with a view toward the classroom*, Sydney, Australia, July 2003.

[2] N. Heffernan and P. Wiemer-Hastings, Eds., *Proceedings of the ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions*, Brazil, 2004.

[3] R. Hausmann and M. Chi, "Can a computer interface support self-explaining?" *The International Journal of Cognitive Technology*, vol. 7, no. 1, pp. 4–14, 2002.

[4] 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.

[5] J. Mostow and G. Aist, "Evaluating tutors that listen: An overview of Project LISTEN," in *Smart Machines in Education*, K. Forbus and P. Feltovich, Eds. MIT/AAAI Press, 2001, pp. 169–234.

[6] H. Pon-Barry, B. Clark, E. Bratt, K. Schultz, and S. Peters, "Evaluating the effectiveness of SCoT: A spoken conversational tutor," in *Proceedings of ITS 2004 Workshop on Dialogue-based Intelligent Tutoring Systems: State of the Art and New Research Directions*, 2004.

[7] 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.

[8] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with paradise," *Natural Language Engineering*, vol. 6, no. 3, 2000.

[9] M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard, "Darpa communicatior: Cross-system results for the 2001 evaluation," in *Proceedings of ICSLP*, 2002.

[10] 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.

[11] C. P. Rosé, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn, "The role of why questions in effective human tutoring," in *Proc. Artificial Intelligence in Education*, 2003.