ITSPOKE: An Intelligent Tutoring Spoken Dialogue System

Diane J. Litman
University of Pittsburgh
Department of Computer Science &
Learning Research and Development Center
Pittsburgh PA, 15260, USA
litman@cs.pitt.edu

Scott Silliman
University of Pittsburgh
Learning Research and Development Center
Pittsburgh PA, 15260, USA
scotts@pitt.edu

Abstract

ITSPOKE is a spoken dialogue system that uses the Why2-Atlas text-based tutoring system as its "back-end". A student first types a natural language answer to a qualitative physics problem. ITSPOKE then engages the student in a spoken dialogue to provide feedback and correct misconceptions, and to elicit more complete explanations. We are using ITSPOKE to generate an empirically-based understanding of the ramifications of adding spoken language capabilities to text-based dialogue tutors.

1 Introduction

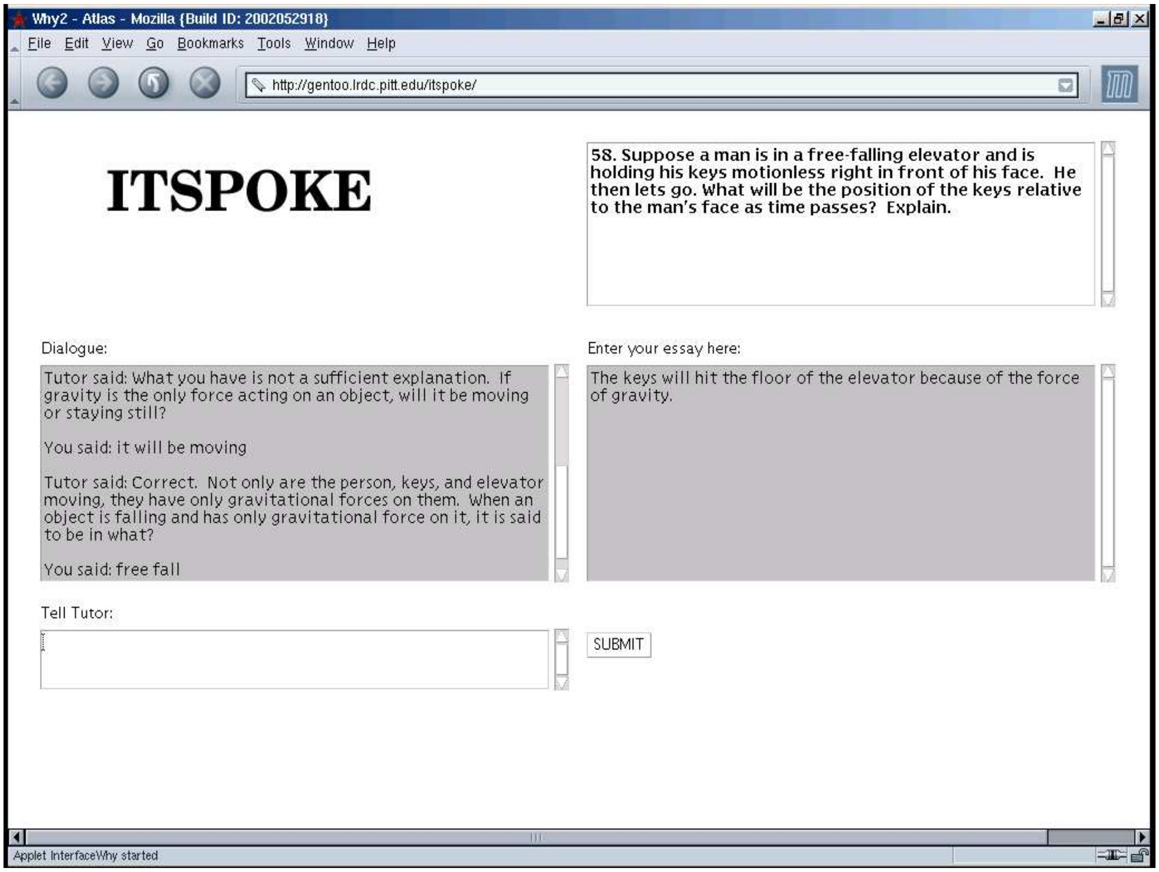
The development of computational tutorial dialogue systems has become more and more prevalent (Aleven and Rose, 2003), as one method of attempting to close the performance gap between human and computer tutors. While many such systems have yielded successful evaluations with students, most are currently text-based (Evens et al., 2001; Aleven et al., 2001; Zinn et al., 2002; VanLehn et al., 2002). There is reason to believe that *speech-based* tutorial dialogue systems could be even more effective. Spontaneous self-explanation by students improves learning gains during human-human tutoring (Chi et al., 1994), and spontaneous self-explanation occurs more frequently in spoken tutoring than in text-based tutoring (Hausmann and Chi, 2002). In human-computer tutoring, the use of an interactive pedagogical agent that communicates using speech rather than text output improves student learning, while the visual presence or absence of the agent does not impact performance (Moreno et al., 2001). In addition, it has been hypothesized that the success of computer tutors could be increased by recognizing and responding to student emotion. (Aist et al., 2002) have shown that adding emotional processing to a dialogue-based reading tutor increases student persistence. Information in the speech signal such as prosody has been shown to be a rich source of information for predicting emotional states in other types of dialogue interactions (Ang et al., 2002; Lee et al., 2002; Batliner et al., 2003; Devillers et al., 2003; Shafran et al., 2003).

With advances in speech technology, several projects have begun to incorporate basic spoken language capabilities into their systems (Mostow and Aist, 2001; Fry et al., 2001; Graesser et al., 2001; Rickel and Johnson, 2000). However, to date there has been little examination of the ramifications of using a spoken modality for dialogue tutoring. To assess the impact and evaluate the utility of adding spoken language capabilities to dialogue tutoring systems, we have built ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system), a spoken dialogue system that uses the Why2-Atlas conceptual physics tutoring system (VanLehn et al., 2002) as its "back-end." We are using ITSPOKE as a platform for examining whether acoustic-prosodic information can be used to improve the recognition of pedagogically useful information such as student emotion (Forbes-Riley and Litman, 2004; Litman and Forbes-Riley, 2004), and whether speech can improve the performance evaluations of dialogue tutoring systems (e.g., as measured by learning gains, efficiency, usability, etc.) (Rosé et al., 2003).

2 Application Description

ITSPOKE is a *speech-enabled* version of the Why2-Atlas (VanLehn et al., 2002) *text-based* dialogue tutoring system. As in Why2-Atlas, a student first types a natural language answer to a qualitative physics problem. In ITSPOKE, however, the system engages the student in a *spoken dialogue* to correct misconceptions and elicit more complete explanations.

Consider the screenshot shown in Figure 1. ITSPOKE first poses conceptual physics problem 58 to the student, as shown in the upper right of the figure. Next, the


Why2 - Atlas - Mozilla {Build ID: 2002052918}
File Edit View Go Bookmarks Tools Window Help
http://gentoo.lrdc.pitt.edu/itspoke/
ITSPOKE
Dialogue:
Enter your essay here:
Tell Tutor:
SUBMIT
Applet InterfaceWhy started

58. Suppose a man is in a free-falling elevator and is holding his keys motionless right in front of his face. He then lets go. What will be the position of the keys relative to the man's face as time passes? Explain.

Tutor said: What you have is not a sufficient explanation. If gravity is the only force acting on an object, will it be moving or staying still?

You said: it will be moving

Tutor said: Correct. Not only are the person, keys, and elevator moving, they have only gravitational forces on them. When an object is falling and has only gravitational force on it, it is said to be in what?

You said: free fall

The keys will hit the floor of the elevator because of the force of gravity.

Figure 1: Screenshot during ITSPOKE Human-Computer Spoken Dialogue

user types in a natural language essay answer (as shown in the essay box in the middle right of Figure 1), and clicks "SUBMIT." ITSPOKE then analyzes the essay, after which the spoken dialogue with the student begins.

During the dialogue, the system and student discuss a solution to the problem relative to the student's essay explanation, using spoken English. At the time the screenshot was generated, the student had just said "free fall." After each system or student utterance, the system prompt, or the system's understanding of the student's response, respectively, are added to the dialogue history (as shown in the dialogue box in the middle left of Figure 1).[1]

At some point later in the dialogue, the system will eventually ask the student to edit the typed essay explanation. The system will then either express satisfaction and end the tutoring for the current problem, or continue with another round of spoken dialogue interaction and typed essay revision. Upon completing the problem, the student is presented with an example of an ideal essay.

Another excerpt with ITSPOKE is shown in Figure 2.

3 System Architecture

The architecture of ITSPOKE is shown in Figure 3. Student speech is digitized from microphone input and sent to the Sphinx2 automatic speech recognizer. Sphinx2's best hypothesis is then sent to Why2-Atlas for syntactic and semantic analysis (Rosé, 2000), discourse and domain processing (Jordan et al., 2003), and finite-state dialogue management (Rosé et al., 2001). The tutor's text output is sent to the Cepstral text-to-speech system, and played through a speaker or headphone.

ITSPOKE is implemented in Python and acts as a proxy server between the Why2-Atlas server and client. ITSPOKE monitors the XML being sent between the two and decides what text should be spoken and when to listen. After speech recognition, ITSPOKE submits what it thinks the student said to Why2-Atlas (as if the student had typed it). Sphinx2's C API is tied into Python by using SWIG to generate a Python wrapper API. Cepstral's speech synthesis engine is currently run out of process using a fork/exec, but we plan to use SWIG to tie its C API into Python in the future. For speed, we have in-

[1] The "Tell Tutor" box is used for typed student login only.

ITSPOKE: I didn't understand. What force acts on all objects within close proximity to earth?
STUDENT: uh the force of gravity
ITSPOKE: Fine. Besides the Earth's gravitational force, are there any other forces acting on the person, keys, and elevator after the release?
STUDENT: no
ITSPOKE: Yep. Thus, keep in mind that after the release, the only force on the person, keys, and elevator is the force of gravity. If gravity is the only force acting on an object, will it be moving or staying still?
STUDENT: it will be moving
ITSPOKE: Fine. Not only are the person, keys, and elevator moving, they have only gravitational forces on them. When an object is falling and has only gravitational force on it, it is said to be in what?
STUDENT: freefall
ITSPOKE: Good. Thus, after the release the only force on the keys and man is the downward force of earth's gravity, so they are in freefall. Before the release, what is the relationship between the man's velocity and the keys' velocity?
STUDENT: they're equal

Figure 2: ITSPOKE Excerpt (3.5 minutes into session)

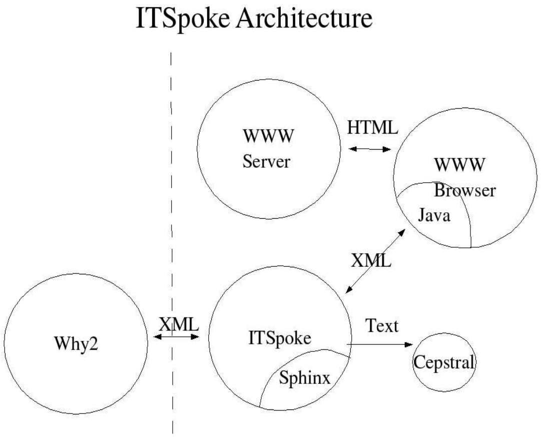


Figure 3: The Architecture of ITSPOKE

stalled Sphinx2 and Cepstral on the ITSPOKE machine. The dashed line in the figure reflects that Why2-Atlas can be installed on a different machine, although we are currently running it on the same machine.

4 Performance Analysis

A formal evaluation comparing ITSPOKE and other tutoring methods began in November 2003, and is still ongoing. Subjects are University of Pittsburgh students who have taken no college physics and are native speakers of American English. Our experimental procedure, taking roughly 4 hours/student, is as follows: students 1) read a small document of background material, 2) take a pretest measuring their physics knowledge, 3) use ITSPOKE to work through 5 physics problems, and 4) take a post-test similar to the pretest.

As of March 2004, we have collected 80 dialogues from 16 students (21 total hours of speech, mean dialogue time of 17 minutes). An average dialogue contains 21.3 student turns and 26.3 tutor turns. The mean student turn length is 2.8 words (max=28, min=1).[2]

ITSPOKE uses 56 dialogue-state dependent language models for speech recognition; 43 of these 56 models have been used to process the data collected to date.[3] These stochastic language models were initially trained using 4551 typed student utterances from a 2002 evaluation of Why2-Atlas, then later enhanced with spoken utterances obtained during ITSPOKE's pilot testing. For the 1600 student turns that we have collected, ITSPOKE's current Word Error Rate is 31.2%. While this is the traditional method of evaluating speech recognition, semantic rather than transcription accuracy is more useful for dialogue evaluation as it does not penalize for word errors that are unimportant to overall utterance interpretation. Semantic analysis based on speech recognition is the same as based on perfect transcription 92% of the time. An average dialogue contains 1.4 rejection prompts (when ITSPOKE is not confident of the speech recognition output, it asks the user to repeat the utterance), and .8 timeout prompts (when the student doesn't say anything within a specified time frame, ITSPOKE repeats its previous question).

5 Summary

The goal of ITSPOKE is to generate an empirically-based understanding of the implications of using speech instead of text-based dialogue tutoring, and to use these results to build an improved version of ITSPOKE. We are currently analyzing our corpus of dialogues with ITSPOKE to determine whether spoken dialogues yield increased performance compared to text with respect to a variety of evaluation metrics, and whether acoustic-prosodic features only found in speech can be used to better predict pedagogically useful information such as student emotions. Our next step will be to modify the dialogue manager inherited from Why2-Atlas to use new tutorial strategies optimized for speech, and to enhance ITSPOKE to predict and adapt to student emotion. In previous work on adaptive (non-tutoring) dialogue systems (Litman and Pan, 2002), adaptation to problematic dialogue situations measurably improved system performance.

Acknowledgments

This research has been supported by NSF Grant Nos. 9720359, 0328431 and 0325054, and by the Office of

[2] Word count is estimated from speech recognition output.

[3] The remaining language models correspond to physics problems that are not being tested in the current evaluation.

Naval Research. Thanks to Kurt VanLehn and the Why2-Atlas team for the use and modification of their system.

References

G. Aist, B. Kort, R. Reilly, J. Mostow, and R. Picard. 2002. Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens. In *Proc. of Intelligent Tutoring Systems*.

V. Aleven and C. P. Rose. 2003. Proc. of the AIED 2003 Workshop on Tutorial Dialogue Systems: With a View toward the Classroom.

V. Aleven, O. Popescu, and K. Koedinger. 2001. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In J. D. Moore, C. L. Redfield, and W. L. Johnson, editors, *Proc. of Artificial Intelligence in Education*, pages 246–255.

J. Ang, R. Dhillon, A. Krupski, E.Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. of ICSLP*.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. How to find trouble in communication. *Speech Communication*, 40:117–143.

M. Chi, N. De Leeuw, M. Chiu, and C. Lavancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18:439–477.

L. Devillers, L. Lamel, and I. Vasilescu. 2003. Emotion detection in task-oriented spoken dialogs. In *Proc. of ICME*.

M. Evens, S. Brandle, R. Chang, R. Freedman, M. Glass, Y. Lee, L. Shim, C. Woo, Y. Zhang, Y. Zhou, J. Michaeland, and Allen A. Rovick. 2001. Circsim-tutor: An Intelligent Tutoring System Using Natural Language Dialogue. In *Proc. Midwest AI and Cognitive Science Conference*.

K. Forbes-Riley and D. Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proc. Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics*.

J. Fry, M. Ginzton, S. Peters, B. Clark, and H. Pon-Barry. 2001. Automated tutoring dialogues for training in shipboard damage control. In *Proc. SIGdial Workshop on Discourse and Dialogue*.

A. Graesser, N. Person, and D. Harter et al. 2001. Teaching tactics and dialog in Autotutor. *International Journal of Artificial Intelligence in Education*.

R. Hausmann and M. Chi. 2002. Can a computer interface support self-explaining? *The International Journal of Cognitive Technology*, 7(1).

P. Jordan, M. Makatchev, and K. VanLehn. 2003. Abductive theorem proving for analyzing student explanations. In *Proc. Artificial Intelligence in Education*.

C.M. Lee, S. Narayanan, and R. Pieraccini. 2002. Combining acoustic and language information for emotion recognition. In *Proc. of ICSLP*.

D. J. Litman and K. Forbes-Riley. 2004. Annotating student emotional states in spoken tutoring dialogues. In *Proc. SIGdial Workshop on Discourse and Dialogue*.

D. J. Litman and S. Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12.

R. Moreno, R.E. Mayer, H. A. Spires, and J. C. Lester. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents. *Cognition and Instruction*, 19(2):177–213.

J. Mostow and G. Aist. 2001. Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus and P. Feltovich, editors, *Smart Machines in Education*.

J. Rickel and W. L. Johnson. 2000. Task-oriented collaboration with embodied agents in virtual worlds. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*.

C. P. Rosé, P. Jordan, M. Ringenberg, S. Siler, K. VanLehn, and A. Weinstein. 2001. Interactive conceptual tutoring in Atlas-Andes. In *Proc. Artificial Intelligence in Education*.

C. P. Rosé, D. Litman, D. Bhembe, K. Forbes, S. Silliman, R. Srivastava, and K. VanLehn. 2003. A comparison of tutor and student behavior in speech versus text based tutoring. In *Proc. HLT/NAACL Workshop: Building Educational Applications Using NLP*.

C. P. Rosé. 2000. A framework for robust sentence level interpretation. In *Proc. North American Chapter of the Association for Computational Lingusitics*.

I. Shafran, M. Riley, and M. Mohri. 2003. Voice signatures. In *Proc. of Automatic Speech Recognition and Understanding*.

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. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems*.

C. Zinn, J. D. Moore, and M. G. Core. 2002. A 3-tier planning architecture for managing tutorial dialogue. In *Proc. Intelligent Tutoring Systems*.