Comparing Synthesized versus Pre-Recorded Tutor Speech in an Intelligent Tutoring Spoken Dialogue System*

Kate Forbes-Riley and **Diane Litman** and **Scott Silliman** and **Joel Tetreault**
Learning Research and Development Center, University of Pittsburgh,
3939 O'Hara Street, Pittsburgh, PA, 15260
{forbesk, litman, scotts, tetreaul}@cs.pitt.edu

Abstract

We evaluate the impact of tutor voice quality in the context of our intelligent tutoring spoken dialogue system. We first describe two versions of our system which yielded two corpora of human-computer tutoring dialogues: one using a tutor voice pre-recorded by a human, and the other using a low-cost text-to-speech tutor voice. We then discuss the results of two-tailed t-tests comparing student learning gains, system usability, and dialogue efficiency across the two corpora and across corpora subsets. Overall, our results suggest that tutor voice quality may have only a minor impact on these metrics in the context of our tutoring system. We find that tutor voice quality does not impact learning gains, but it may impact usability and efficiency for some corpora subsets.

Introduction

In recent years the development of intelligent tutoring *dialogue* systems has become more prevalent, in an attempt to close the performance gap between human and computer tutors. Although many of these systems are text-based (Evens & Michael 2006; Zinn, Moore, & Core 2002; Aleven, Popescu, & Koedinger 2001; VanLehn *et al.* 2002), with recent advances in speech technology, several systems have begun to incorporate spoken language capabilities (Beck, Jia, & Mostow 2004; Pon-Barry *et al.* 2004; Graesser *et al.* 2005; Rickel & Johnson 2000; Litman & Silliman 2004), hypothesizing that adding speech technology will promote student learning by enhancing communication richness. However, the relationship between the quality of speech technology and student learning is not yet clear; i.e., is high quality speech technology required to maximize the ability of students to learn?

Although results are somewhat mixed, recent work suggests that the quality of the computer tutor voice - i.e., whether the tutor voice is synthesized with a text-to-speech system, or is a human voice that has been pre-recorded - can impact system effectiveness. In the domain of instructional planning, for example, students rate both visual and non-visual intelligent agents as more engaging and human-like when audio recordings of a human voice are used (Baylor, Ryu, & Shen 2003). Student motivation also increases when the human voice is used with the non-visual version of the agent. However, with the visual agent, the synthesized voice increases motivation. In experiments in both laboratory and school settings using a computer learning environment for teaching math, a human voice is preferable even when the agent is animated: students learn more deeply, and give more positive ratings to the agent, than when a machine-generated voice is used (Atkinson, Mayer, & Merrill 2005). As these authors note, however, this finding may change as machine-generated voices improve, and/or if students are first given practice listening to machine-generated voices. Research on other types of spoken dialogue systems has also shown users prefer pre-recorded to synthesized audio (e.g. (Team 1999)). However, a recent study of a "smart-home" spoken dialogue system (Moller, Krebber, & Smeele 2006) found that although users generally prefer the more "natural-sounding" system voice, specific voice characteristics, such as "voice pleasantness" and "listening effort required", seem to have higher importance than whether the voice is synthesized or pre-recorded.

Of course, pre-recorded speech is much more costly than synthesized speech, and it is also less flexible when combined with more dynamic natural language generation capabilities. In this study, we investigate whether pre-recorded speech is more effective than synthesized speech in the context of our intelligent tutoring spoken dialogue system, which unlike (Baylor, Ryu, & Shen 2003; Atkinson, Mayer, & Merrill 2005), has no visual agent, uses speech input and output, and has a full natural language dialogue system as a back-end. We compare corpora collected from two versions of our system: one with a tutor voice pre-recorded by a human, and one with a low-cost text-to-speech synthesized tutor voice. We use two-tailed t-tests to evaluate differences in three evaluation metrics across the two corpora overall and corpora subsets. First, we evaluate differences in student learning gains; student learning is an important evaluation metric for intelligent tutoring systems. We also evaluate differences in system usability (measured by subject surveys) and dialogue efficiency (measured by time on task); these evaluation metrics are important for dialogue systems in general. We hypothesize that in our tutoring system, the more "natural-sounding" pre-recorded tutor voice will per-

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

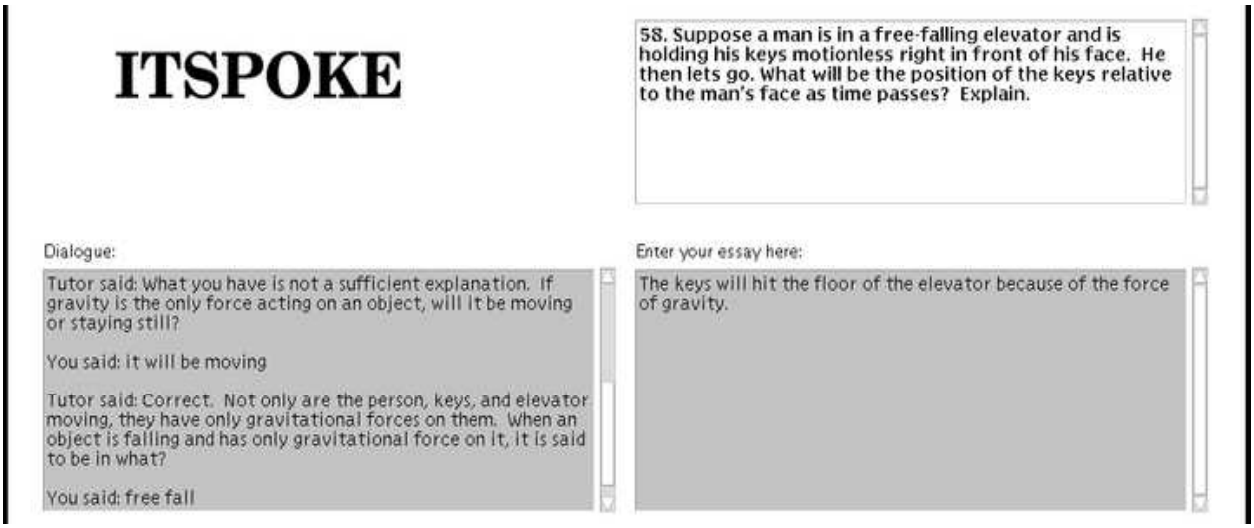

Figure 1: Screenshot during an ITSPOKE Spoken Tutoring Dialogue

form better across all evaluation metrics. Overall, however, our results show that tutor voice quality has only a minor impact on our evaluation metrics in the context of our tutoring system. In particular, tutor voice quality does not significantly impact learning across our corpora overall or any corpora subsets. Tutor voice quality may impact usability and efficiency, but only for certain corpora subsets, and like the studies cited above, we find mixed results with respect to the impact of tutor voice quality: for some corpora subsets the pre-recorded voice may be preferable, while for others the synthesized voice may be preferable.

The Experiments

The ITSPOKE System and Corpora

ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialogue system) (Litman & Silliman 2004) is a *speech-enabled* version of the *text-based* Why2-Atlas conceptual physics tutoring system (VanLehn *et al.* 2002). Dialogues between students and ITSPOKE are mediated by a web interface supplemented with a headphone-microphone unit. An example screenshot of the ITSPOKE interface is shown in Figure 1. As shown in the lower right box, the student first types an essay answering a qualitative physics problem, which is shown in the upper right box. 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 the microphone input and sent to the Sphinx2 recognizer. Sphinx2's most probable "transcription" output is sent to the Why2-Atlas back-end for syntactic, semantic and dialogue analysis. Finally, the text response produced by the back-end is converted to speech as described below, then is played to the student through the headphone and displayed in the lower left box of the interface at the same time as it is spoken. This scrollable box records the entire dialogue history between the student and ITSPOKE.

For this study, we implemented two different versions of ITSPOKE. One version used a synthesized tutor voice, and the other version used a pre-recorded tutor voice.[1] For the synthesized voice, we purchased Cepstral's voice entitled "Frank" for $29.95. Each tutor text response produced via the Why2-Atlas back-end was sent to the Cepstral text-to-speech system to render the response as speech. For the pre-recorded voice, a paid voice talent was first recorded speaking each tutor response in his natural "academic" voice. The appropriate audio file(s) were then played when the corresponding text responses were produced by the back-end. Our voice talent recorded a total of 5.85 hours of audio, which took 25 hours of paid voice talent time (at $120/hr).

We collected two corpora of spoken tutoring dialogues: one using the version of ITSPOKE with the pre-recorded voice, and one using the version of ITSPOKE with the synthesized voice. These corpora were collected in spring, 2005. The pre-recorded corpus contains 28 subjects and the synthesized corpus contains 29 subjects. Subjects were recruited from advertisements on campus at the University of Pittsburgh, were required to have not taken college physics, and were paid for their involvement. Subjects were randomly assigned to one of the conditions after passing a screening test based on their speech recognition performance. The experimental procedure for corpus collection was as follows: Subjects 1) read a small document of background material, 2) took a pretest measuring their initial physics knowledge, 3) used a web and voice interface to work through a set of 5 training problems (dialogues) with the computer tutor, 4) took a posttest similar to the pretest[2], and 5) completed a survey questionnaire (described below).

[1] Readers can hear a tutor turn spoken in each voice at this website: http://www.cs.pitt.edu/itspoke/pub/flairs06/index.html

[2] Our isomorphic pre- and posttests consisted of 40 multiple choice questions originally developed for the Why2-Atlas backend.

Evaluation Metrics

Once the two corpora were collected, we evaluated the significance of the differences between the two corpora with respect to three main evaluation metrics: *student learning gain*, *dialogue efficiency*, and *system usability*.

"Student learning gain" is an important evaluation metric for intelligent tutoring systems.[3] A standard measure of learning gain is: *posttest score - pretest score* (e.g. as in (Chi *et al.* 1994)). Hereafter we refer to this measure as **SLG** (standardized learning gain). However, this measure does not normalize for variation in student pretest scores. A common measure of learning gain that does normalize for pretest is: $\frac{(posttest\ score - pretest\ score)}{(1 - pretest\ score)}$ (e.g. as in (Crouch & Mazur 2001)). Hereafter we refer to this measure as **NLG** (normalized learning gain). These two learning gain measures were calculated for each student in each condition.

"Dialogue efficiency" is another important evaluation metric for most dialogue systems, e.g. how long a given task takes to complete. This metric is important in tutoring dialogue systems too. Here we measure dialogue efficiency using "time on task". For each student in each condition, we calculate the total time of all their dialogues with the system (in minutes). Hereafter we refer to this metric as **TOT**.

Finally, many dialogue systems are evaluated in terms of a "system usability" survey, which encompasses subjective perceptions of likability, ease of use, text-to-speech quality, etc. This metric is important in tutoring systems too, as students won't want to use the system if they don't feel it is usable. For this study, we constructed a survey with the 11 usability statements listed in Figure 2, which each student completed after taking the posttest. Students rated their degree of agreement with each statement on a scale of 1 to 5, as shown at the bottom of the figure. Statements 1-7, taken from (Baylor, Ryu, & Shen 2003), were tailored to the tutoring domain. Statements 8-11, taken from (Walker *et al.* 2002), were more generally applicable to dialogue systems. Hereafter we refer to each statement as **S#** (e.g. **S11**).

S1. It was easy to learn from the tutor.
S2. The tutor interfered with my understanding of the content.
S3. The tutor believed I was knowledgeable.
S4. The tutor was useful.
S5. The tutor was effective on conveying ideas.
S6. The tutor was precise in providing advice.
S7. The tutor helped me to concentrate.
S8. It was easy to understand the tutor.
S9. I knew what I could say or do at each point in the conversations with the tutor.
S10. The tutor worked the way I expected it to.
S11. Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly.

ALMOST ALWAYS (5), OFTEN (4), SOMETIMES (3), RARELY (2), ALMOST NEVER (1)

Figure 2: ITSPOKE Survey

Evaluation Methodology

To evaluate the differences between the two corpora overall, we computed two-tailed t-tests for each evaluation metric over all the *students* in each corpus. These results are discussed in the first part of the Results section below.

We also evaluated differences between specific subsets of students in each corpus who might be more susceptible to tutor voice quality with respect to learning, system usability, or dialogue efficiency. For each of three criteria discussed below, we partitioned the students in each corpus into "high" and "low" subsets, based on the median value for each criterion in each corpus; if a median value occurred, it was excluded. Because students with the very highest or lowest values for a criterion may display more differences, we also partitioned students into "highest" and "lowest" subsets, as in (Chi *et al.* 1994), using a cutoff value above/below the median for each criterion. Cutoff values were chosen to obtain roughly equal numbers of students per subset and exclude at least half the students in the high/low subsets.[4] We then computed two-tailed t-tests for each subset.

Our first partition criterion was **pretest score**, to evaluate differences between only those students in each condition who began the study with more or less physics knowledge or intuition than other students. We hypothesized that students with low pretest scores might be more influenced by what the tutor was saying, so that we might see greater differences in student learning, usability or efficiency across conditions for these students. This hypothesis was also motivated by (Vanlehn *et al.* submitted), who found that different tutoring methods only impacted learning for students with low pretest scores. T-test results for pretest partitions are discussed in the second part of the Results section below.

Our second partition criterion was **word error rate**, to evaluate differences between only those students in each condition whose speech was more or less understood than other students. Word error rates were computed by comparing the manual transcription of the student's dialogue and the recognized output, using the SCLITE scoring algorithm from the NIST Scoring Toolkit Version 0.1. We hypothesized that how well (or badly) the students were being understood might impact how much they listen to the tutor, e.g. we might see more significant differences in learning, usability or efficiency across conditions for students with high word error rates. T-test results for word error rate partitions are discussed in the third part of the Results section below.

Our third partition criterion was **time on task**. We used this as a partition criterion (as well as an evaluation metric), to evaluate differences between only those students in each condition who spent more or less time with the tutor

[3] A 2-way ANOVA with condition by repeated test measures design showed a robust main effect for test phase, $F(1,55) = 178.28$, $p = 0.000$, $MSe = 0.005$, with no reliable interaction effect for condition, indicating that students in both conditions learned a significant amount independently of condition.

[4] Cutoff values were chosen prior to any analyses, based on the range of observed values. In future work we will also try other procedures, but since multiple students can have the same value, all subsets usually won't contain equal numbers of students.

than other students. We hypothesized that students who had more exposure to the tutor's voice might show more significant differences in learning, usability or efficiency across conditions. T-test results for time on task partitions are discussed in the last part of the Results section below.

Results and Discussion

T-Test Results for All Subjects

Table 1 presents the results of the t-tests for each evaluation metric across all the students in each condition. For all tables hereafter, the first column shows the evaluation metric being tested. The second and third columns show the student mean for that metric in each condition, with standard deviation shown in parentheses. The pre-recorded condition is labeled **PR**, and the synthesized condition is labeled **SYN**. The fourth column shows the mean difference across conditions, and the last column shows the significance (p-value) of this difference, where $p \leq .05$ indicates that the mean difference was significant and $p \leq .10$ indicates a trend for a significant difference.[5] The caption indicates the number of subjects in each condition.

Metric	PR Mean	SYN Mean	Diff	p
SLG	.17 (.09)	.17 (.10)	.00	.92
NLG	.41 (.20)	.35 (.24)	.06	.28
TOT	121.04 (32.9)	121.34 (33.4)	-.30	.97
S1	3.46 (.74)	3.45 (.78)	.02	.94
S2	2.29 (.76)	2.21 (.56)	.08	.66
S3	**3.50 (.88)**	**3.00 (.96)**	**.50**	**.05**
S4	3.71 (.85)	3.86 (.83)	-.15	.51
S5	4.04 (.84)	4.00 (.76)	.04	.87
S6	3.79 (1.03)	3.72 (.84)	.06	.81
S7	2.93 (1.05)	2.93 (.92)	.00	.99
S8	3.93 (1.05)	3.83 (.85)	.10	.69
S9	3.64 (.91)	3.86 (.83)	-.22	.35
S10	3.71 (.71)	3.83 (.81)	-.11	.58
S11	2.21 (1.20)	2.59 (1.02)	-.37	.21

Table 1: T-Tests: All, **PR** (28) versus **SYN** (29)

As shown, when considering all subjects in the two conditions, there was no significant difference in either calculation of learning gains (SLG or NLG). This result does not support the hypothesis that student learning gains will be higher with pre-recorded tutor speech than with less "human-sounding" synthesized tutor speech. Similarly, there was no significant difference in dialogue efficiency, i.e. the mean time on task (TOT). Moreover, there was no significant difference for most of the individual survey questions, which measure system usability. However, there was a significant difference across the two conditions for student responses to **S3**: *The tutor believed I was knowledgeable.* The mean difference is positive (.50), indicating students in the pre-recorded condition scored this question significantly higher than the students in the synthesized condition. This result suggests both that students attributed more "human-like" qualities to the more "human-sounding" voice (i.e. the students believed the pre-recorded tutor possessed beliefs about the students' knowledge states), and that students overall showed a preference for the pre-recorded tutor voice.

Hereafter, we only tabulate results where the mean difference across conditions was (unadjusted) significant ($p \leq .05$) or showed a trend for significance ($p \leq .10$).

T-Test Results for Partitions by Pretest Score

Students with the highest pretest scores showed a trend for a significant difference in dialogue efficiency across conditions, as shown in Table 2. The mean difference in time on task (TOT) is negative, indicating that students with high pretest scores showed a trend to take significantly longer to complete the dialogues in the synthesized condition, as compared to the pre-recorded condition. One possible interpretation of this result is that these more knowledgeable students in the synthesized condition took more time to read the dialogue transcript shown on the ITSPOKE interface as noted above. Although student responses to **S8**: *It was easy to understand the tutor.* yielded no significant differences or trends overall (Table 1) or for highest pretest students, in both cases the mean score for this question was slightly lower for students in the synthesized condition. As in other recent work evaluating the impact of speech synthesis quality on spoken dialogue system usability (Moller, Krebber, & Smeele 2006), future versions of our survey will contain more questions aimed at teasing out any relationship between tutor voice understandability and efficiency or usability, e.g. how much effort was required to understand the voice, and how often the students read the transcript.

Metric	PR Mean	SYN Mean	Diff	p
TOT	100.9 (18.4)	121.9 (25.6)	-21.0	.09

Table 2: T-Tests: Highest Pretest, **PR** (6) versus **SYN** (9)

There were no significant differences or trends in student learning gains or system usability, for any subsets of students partitioned by pretest score. Our initial hypotheses that students with low pretest scores would show significant differences in learning gains or system usability across the two conditions, are thus not supported.

T-Test Results for Partitions by Word Error Rate

Students with high word error rates showed trends for significant differences across conditions in their scores on **S3** and **S11**, as shown in Table 3. The mean difference on **S3** was positive; this is the same result that was found using all students (Table 1), although it is less significant here.

For **S11**: *Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly*, the mean difference was negative. This trend suggests that students

[5] These p-values are not adjusted for the fact that 14 t-tests are performed on each dataset, increasing the chance of a false positive result. If we adjust with the Bonferroni correction, then $p \leq .004$ (.05/14) indicates a significant difference and $p \leq .007$ (.1/14) indicates a trend, but the chance of a false negative result increases.

with high word error rates would use the system with the synthesized voice more regularly than the system with the pre-recorded voice. One possible interpretation of this result is that these students who aren't being understood very well prefer the "machine-sounding" voice because it is more consistent with their experience; i.e. in the synthesized condition it is more clear to these students that a machine (not a more intelligent human) is misunderstanding them. As noted above, (Baylor, Ryu, & Shen 2003) also found conditional benefits of a synthesized tutor voice, in that students were more motivated by the animated agent with a synthesized voice. They suggest that when the system is too human-like, learners' expectations about being understood may be too high and negatively impact their experience.

Metric	PR Mean	SYN Mean	Diff	p
S3	3.43 (1.02)	2.64 (1.15)	.79	.07
S11	1.86 (1.17)	2.64 (1.08)	-.79	.08

Table 3: T-Tests: High WER, **PR** (14) versus **SYN** (14)

As shown in Table 4, student with low word error rates showed a trend for a significant difference across conditions in their scores on **S2**: *The tutor interfered with my understanding of the content.*, where the mean difference was positive, suggesting that students with low word error rates felt that the tutor with the pre-recorded voice interfered more with their understanding, as compared to the tutor with the synthesized voice. One possible interpretation of this result is that when all other things go well (i.e. good speech recognition and "human-sounding" tutor voice), deeper problems can come into the student's focus, i.e. the inflexibility of the underlying natural language understanding and generation of the system. Moreover, like the result for **S3**, students in the pre-recorded condition seem to be attributing more human-like qualities to the more "human-sounding" tutor voice (i.e., the ability to "interfere").

Metric	PR Mean	SYN Mean	Diff	p
S2	2.36 (.75)	1.93 (.48)	.43	.08

Table 4: T-Tests: Low WER, **PR** (14) versus **SYN** (14)

As shown in Table 5, students with the highest word error rates showed the same trend as those with high word error rates (Table 3) in their scores on **S11**; this trend came closer to significance (p=.06).

Metric	PR Mean	SYN Mean	Diff	p
S11	1.50 (.55)	2.71 (1.38)	-1.21	.06

Table 5: T-Tests: Highest WER, **PR** (6) versus **SYN** (7)

There were no significant differences or trends in student learning or efficiency, for any subsets of students partitioned by word error rate. Note that whether word misrecognition impacts the tutoring or dialogue length depends on which word(s) is misrecognized. In a prior study of another ITSPOKE corpus we found no correlations between word error rate and learning, but we did find correlations between time on task and a boolean version of word error rate that measures whether or not any words in a turn are misrecognized (Litman & Forbes-Riley 2005).

T-Test Results for Partitions by Time on Task

Students with the highest time on task showed a trend for a significant difference across conditions for **S3**, as shown in Table 6. The mean difference is positive; this is the same result found for students with high word error rates (Table 3) and all students (Table 1).

Metric	PR Mean	SYN Mean	Diff	p
S3	3.40 (.70)	2.64 (1.03)	.76	.06

Table 6: T-Tests: Highest TOT, **PR** (10) versus **SYN** (11)

There were no significant differences or trends in student learning gains for any subsets of students partitioned by time on task. These results do not support our initial hypothesis that those students who had a higher total amount of exposure to the tutor's voice might show more differences in learning across conditions.

Conclusions and Current Directions

We evaluated the impact of tutor voice quality in the context of ITSPOKE, by comparing differences in student learning (measured by learning gains), system usability (measured by a survey) and dialogue efficiency (measured by time on task), in two ITSPOKE corpora: one with a pre-recorded tutor voice, and the other with a synthesized tutor voice.

We hypothesized that the pre-recorded tutor voice would yield higher student learning than the low-cost synthesized tutor voice, thus indicating the need for implementing high quality speech technology in ITSPOKE and/or in similar intelligent tutoring spoken dialogue systems. Contrary to our hypothesis and (Atkinson, Mayer, & Merrill 2005), however, our t-test results showed no trends or significant differences in student learning gains across our two conditions, suggesting that tutor voice quality does not impact learning in ITSPOKE. However, this result can only be interpreted in the context of ITSPOKE (or similar systems), where the dialogue transcription is available to the student. This likely diluted the impact of tutor voice quality, because since students could read the transcription simultaneously, their learning was not wholly dependent on understanding the tutor's speech. Future versions of ITSPOKE will investigate the best combination of modalities, e.g. by showing the transcript only after the students hear the tutor speech and/or not showing it at all. However, if it benefits learning for students to both hear and read the tutor speech, then the transcription should be shown, even if this dilutes the impact of the tutor voice. Another benefit of ITSPOKE is that students have a lot of time to get used to the voice; average time on task is relatively long (121 min. on average), and students also hear the voice during screening. As (Atkinson,

Mayer, & Merrill 2005) suggest, this may have decreased the impact of tutor voice quality on learning.

We also hypothesized that the pre-recorded tutor voice would yield higher system usability or dialogue efficiency than the synthesized voice, as in (Baylor, Ryu, & Shen 2003) and (Atkinson, Mayer, & Merrill 2005). As in (Baylor, Ryu, & Shen 2003), however, our t-test results were mixed. Most results showed no significant differences or trends in usability or efficiency across conditions. Certain results supported our hypothesis, but others showed a preference for the synthesized tutor voice. In particular, students overall, as well as those with high word error rates and highest time on task (as trends), felt that the tutor in the pre-recorded condition believed them more knowledgeable (S3). However, students with high(est) word error rates showed a trend to prefer to use the system more regularly in the synthesized condition (S11). Like (Baylor, Ryu, & Shen 2003), we suggested that these students may prefer the synthesized voice because it is clearly a machine that is so frequently misunderstanding them. As ITSPOKE's ASR improves, this result may change. In addition, students with low word error rates showed a trend to feel that the tutor interfered more with their understanding in the pre-recorded condition (S2). We suggested that when all else is "human-like" (tutor voice and word error rate), deeper system misunderstandings come into focus. As ITSPOKE's natural language understanding improves, this result too may change. Finally, students with high pretest scores showed a trend to take longer (TOT) in the synthesized condition; we suggested that these more knowledgeable students may have taken more time to read the transcription. Note that our result interpretations are speculative and require further research.

Overall, our major result was the *lack* of many differences in student learning, system usability, or dialogue efficiency across the two conditions (even our few results go away if the stricter Bonferroni correction is applied). In essence, we found that tutor voice quality has only a minor impact in the context of our ITSPOKE tutoring system, which has no visual agent, is a full natural language dialogue system that uses both speech input and output, and makes the dialogue transcript available to the student.

References

Aleven, V.; Popescu, O.; and Koedinger, K. 2001. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proc. Internat. Conf Artificial Intelligence in Education*, 246–255.

Atkinson, R. K.; Mayer, R. E.; and Merrill, M. M. 2005. Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology* 30(1):117–139.

Baylor, A. L.; Ryu, J.; and Shen, E. 2003. The effect of pedagogical agent voice and animation on learning, motivation, and perceived persona. In *Proc. ED-MEDIA*.

Beck, J. E.; Jia, P.; and Mostow, J. 2004. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning* 2:61–81.

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

Crouch, C., and Mazur, E. 2001. Peer instruction: Ten years of experience and results. *American Association of Physics Teachers* 69(9):970–977.

Evens, M., and Michael, J. 2006. *One-on-One Tutoring by Humans and Computers*. Lawrence Erlbaum Associates.

Graesser, A. C.; Chipman, P.; Haynes, B. C.; and Olney, A. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education* 48(4):612–618.

Litman, D., and Forbes-Riley, K. 2005. Speech recognition performance and learning in spoken dialogue tutoring. In *Proceedings of Interspeech/Eurospeech*, 161–164.

Litman, D., and Silliman, S. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Proc. of Human Language Technology/Association for Computational Linguistics (HLT/NAACL) (Companion Vol.)*, 233–236.

Moller, S.; Krebber, J.; and Smeele, P. 2006. Evaluating the speech output component of a smart-home system. *Speech Communication* 48:1–27.

Pon-Barry, H.; Clark, B.; Bratt, E.; Schultz, K.; and Peters, S. 2004. 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*.

Rickel, J., and Johnson, W. L. 2000. Task-oriented collaboration with embodied agents in virtual worlds. In Cassell, J.; Sullivan, J.; Prevost, S.; and Churchill, E., eds., *Embodied Conversational Agents*. MIT Press. 95–122.

Team, D. S. R. 1999. User attitudes towards real and synthetic speech. Technical report, Centre for Communication Interface Research, University of Edinburgh.

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

Vanlehn, K.; Graesser, A.; Jackson, G.; Jordan, P.; Olney, A.; and Rose, C. submitted. Natural language tutoring: A comparison of human tutors, computer tutors and text.

Walker, M.; Rudnicky, A.; Prasad, R.; Aberdeen, J.; Bratt, E.; Garofolo, J.; Hastie, H.; Le, A.; Pellom, B.; Potamianos, A.; Passonneau, R.; Roukos, S.; Sanders, G.; Seneff, S.; and Stallard, D. 2002. Darpa communicator: Cross-system results for the 2001 evaluation. In *Proc. Internat. Conf. on Spoken Language Processing (ICSLP)*, 273–276.

Zinn, C.; Moore, J. D.; and Core, M. G. 2002. A 3-tier planning architecture for managing tutorial dialogue. In *Proc. Intelligent Tutoring Systems Conf. (ITS)*, 574–584.