The Effects of Speech Recognition Errors on Learner’s Contributions, Knowledge, Emotions, and Interaction Experience

Sidney K. D’Mello, Brandon King, Michal Stolarski, Patrick Chipman, and Arthur Graesser

Institute for Intelligent Systems
University of Memphis, Memphis, TN 38152, USA
{sdmello|bgking|mstolrsk|pchipman|graesser}@memphis.edu

Abstract
The paper presents a study in which participants learned computer literacy by having a spoken conversation with AutoTutor, an intelligent tutoring system with conversational dialogue. Thirty students completed a multiple-choice pre-test, a 35-minute training session, and a multiple-choice post-test. After completing the post-test, students reviewed their tutorial interaction and judged what emotions they experienced on the basis of the dialogue history and their facial expressions. Our results revealed that many measures of performance were impervious to poor or modest speech recognition accuracy, which is compatible with a soft constraint-based model. Speech recognition errors had a very subtle impact on learning as well as participants’ emotions and attitudes.

Index Terms: spoken dialogue, spoken input, AutoTutor

1. Introduction
The field of human-computer interaction (HCI) may be radically transformed by computer systems equipped with robust automatic speech recognition (ASR) for speech-to-text transcription, coupled with effective natural language processing (NLP) techniques that link the recognized text to specific computer actions. This approach marks a significant departure from current interaction standards, which are all derivatives of the WIMP (window, icon, menu, pointing device) paradigm. The major motivating factor behind redefining the interaction paradigm from WIMP to voice input is to narrow the communicative barrier between the perceptually deficit computer and the highly expressive human. Humans primarily communicate through speech and a host of non-verbal cues, such as facial expressions, posture, and gesture, rather than through typing, pointing, and clicking. Computer systems that are able to recognize and respond to these communication channels will presumably provide a more effective, meaningful, and naturalistic interaction experience.

Computer technologies that would greatly benefit from verbal input are intelligent tutoring systems (ITSs) that promote one-on-one tutoring. ITSs offer a powerful method of adapting to the learner at a fine-grained level and promoting active knowledge construction. It is believed that the implementation of spoken dialogues in ITSs will result in added learning benefits [1, 2], at least when compared to text-based input. However, this needs to be substantiated by further empirical research. The one study that compared text-based to speech-based ITSs reported no significant differences in learning gains between the two groups [2]. However, the tutorial dialogues of students in the speech condition showed a significant increase in some attributes, such as length of the student utterance, that have been shown to be highly correlated with learning in tutoring environments [3]. It is also reasonable to presume that spoken student-tutor turns provide unique interaction benefits that may indirectly translate into learning gains. Spoken dialogues may prime learners to communicate with the computer in a more socially realistic manner than text-based dialogues [4] and thereby result in better learning. From an HCI perspective, the enhanced social communication afforded by speech input would result in a more naturalistic and user-friendly interaction. This would arguably translate into lower attrition rates, a critical problem for situations with high degrees of learner control, where individuals are one-mouse-click-away from ending the session.

Nevertheless, barring a few exceptions [2, 5, 6], almost all ITSs have involved text-based communication [see 2]. This can be attributed to the technological limitations of ASR systems, such as significant recognition errors, long training times, and substantial computational resources. These problems are further complicated in learning environments in which students typically struggle to come up with answers, thereby plaguing their speech with pauses, mispronunciations, and imperfect phrasings, all factors that increase ASR errors.

This trend is changing, however, as recent advances in the speed and accuracy of speech recognition systems have made speech-based ITSs technologically feasible. However, despite these advances, performance of these systems for conversational speech in real world environments is far from optimal and it is unlikely that perfect recognition will ever be achieved, at least not in the near future. Furthermore, it is difficult to establish a suitable upper bound on ASR performance because the recognition accuracies vary according to the particular ASR engine and speech corpus used for the evaluation.

Consequently, within the context of ITSs, an interesting research question is to determine the level of ASR errors that can be tolerated before the learning process is compromised. This can be accomplished by developing ITSs that can compensate for partial failure in the quality of interpreting the learner’s input. According to soft constraint-satisfaction models, the performance of an intelligent system should not rely on the integrity of any one level or module, but rather should reflect the confluence of several levels/modules that are statistically combined. If an evaluation module can adequately compensate for ASR errors, then such errors would have little to no impact on learning, the emotions of the student, and the overall interaction experience.

We tested this hypothesis by conducting an exploratory study with a new version of AutoTutor that supports speech-based input of student contributions as an alternative to the conventional method of typed input. AutoTutor is a fully automated computer tutor that simulates human tutors and holds conversations with...
students in natural language [7]. AutoTutor helps students learn Newtonian physics and computer literacy by presenting challenging problems (or questions) from a curriculum script and engaging in a mixed-initiative dialog while the learner constructs an answer. Participants’ speech input was recorded while they interacted with AutoTutor. By manually transcribing a sample of the student utterances, we calibrated the accuracy of the speech recognizer and assessed the impact of recognition errors on AutoTutor’s evaluations of students’ answer quality, learning gains, learners’ emotions, and the overall experience of the interaction.

2. Empirical Data Collection

2.1. Methods

The participants were 30 undergraduates at the University of Memphis who participated for course credit. The procedure consisted of a pre-test, an interaction with AutoTutor, a post-test, and judgments of emotions the learner experienced during the session with AutoTutor. The pre-test and post-tests consisted of multiple-choice questions that had been used in previous AutoTutor research.

We used the commercially available Dragon Naturally Speaking™ (v. 6) speech recognition system for speech-to-text translation. Before beginning the tutoring session, participants completed a brief (7-10 minute) training phase in order to produce a model of their speech patterns to improve recognition accuracy. To further boost recognition accuracy, we created a custom dictionary for the recognizer. This dictionary included the common words (e.g., and, because) present in the default dictionary of the recognizer plus 2550 additional content words that were specific to the computer literacy topic (e.g., RAM, ROM, floppy). The content words were added via the Dragon Naturally Speaking Vocabulary Builder™.

After the training session participants completed a pre-test. Participants then interacted with AutoTutor for 35 minutes on one of three randomly assigned topics in computer literacy. A detailed discussion of the architecture, strategies, and effectiveness of AutoTutor is provided in previous publications, as cited above. Each major topic had 6 tutoring questions that required a paragraph of information for a good answer. After interacting with AutoTutor, participants completed a post-test.

Participants then judged the emotions they experienced during the tutoring session from a video of their face that was captured during the interaction. Each video included an audio track with the tutor’s speech and the student’s verbalized contributions.

A list of the affective states and definitions was provided for the learners. The states were boredom, confusion, flow, frustration, delight, surprise, and neutral. These affective states were most frequently experienced in previous studies of AutoTutor [8] that investigated emotions with alternative methods: Emote-aloud procedures during learning and observations of trained judges or peers.

2.2. Transcribing Recorded Contributions

AutoTutor’s log files were parsed from a transcript of the tutorial interactions. A sample of 352 individual student dialogue moves was randomly selected from the total set of 1615 dialogue moves (confidence interval 95%, sampling error 4.62). We ensured that there was an equal distribution of speech excerpts from each participant. In order to evaluate the accuracy of the speech recognition system, the automatically recognized text of each student dialogue turn in the sample was compared to a manual transcription of that dialogue turn, which was prepared by a human from the video recordings of the tutorial interactions.

3. Results and Discussion

3.1. Calibrating ASR Accuracy

We collected two metrics to assess the performance of our ASR system. The first was word recognition rate (WRR), which is a standard metric used to assess the reliability of ASR systems.

WRR is computed as \( \frac{(S + D + I)}{N} \) where \( S \), \( D \), and \( I \) are the number of substitutions, deletions, and insertions in the automatically recognized text when compared to a reference text of \( N \) words, after the two strings are aligned using dynamic string alignment at the word level.

The result of the WRR metric for our ASR system was .45, which lies somewhere between poor and fair when compared to automatic recognition of natural conversational speech. This result implies that our ASR system would be problematic for an ITS that requires deep NLP capabilities in evaluating a learner’s response. Deep NLP-based systems require relatively error-free text for syntactical parsing, etc.

We performed an analysis of the proportions of substitutions, deletions, and insertions, as shown in Figure 1A. A repeated measures ANOVA revealed that there were statistically significant differences among the types of ASR errors, \( F(2, 58) = 117.4, Mse = .006, p < .001 \). Bonferroni post-hoc tests confirmed that substitutions (\( M_{SUB} = .339 \)) constituted the most frequent errors, followed by insertions (\( M_{INS} = .134 \)), and deletions (\( M_{DEL} = .054 \)).

The second metric used to quantify the performance of the ASR system formulates the recognition problem as an information retrieval problem. These measures included precision (PRC, specificity) and recall (RCL, sensitivity), which were computed as \( PRC = \frac{|C_P \cap C_U|}{|C_P|} \) and \( RCL = \frac{|C_U \cap C_H|}{|C_U|} \), where \( C_P \) and \( C_H \) are the set of target words in the reference and hypothesized texts respectively. Also included in this metric was the F-measure, which combines precision and recall and is computed as \( FMS = 2 \times (PRC \times RCL) / (PRC + RCL) \).

As mentioned above, our speech recognizer was provided with a custom dictionary that consisted of common words in the English language and content words that are highly specific to
the computer literacy topic. In order to test the efficacy of the dictionary we computed the precision and recall in two separate modes: (1) common + content words and (2) content words only. These results are presented in Figure 1B. Note that F-measure scores associated with content words ($M_{\text{CONTENT}} = .784$) are significantly higher than the corresponding scores of all the words taken together ($M_{\text{ALL}} = .693$), $F(1, 29) = 28.7$, $Mse = .004$, $p < .001$. Therefore, content words were being recognized at an acceptable rate.

3.2. Effects of ASR Errors on AutoTutor’s Evaluation of Learners’ Contributions

AutoTutor measures student response quality by comparing each of the student’s verbal contributions with potential good answers (called expectations) to the topic being discussed. The metric used to evaluate the conceptual similarity between the student’s response and the current expectation is based on Latent Semantic Analysis (LSA) and Inverse Word Overlap (WOL).

LSA is a statistical technique that measures the conceptual similarity of two text sources [9]. LSA computes a geometric cosine (ranging from 0 to 1) that represents the similarity between the two text sources (i.e., learner’s response and current expectation). The WOL algorithm is a word-matching algorithm in which each word is weighted on a scale from 0 to 1.0, relative to its inverse frequency in a corpus of text. As a consequence, higher frequency words, such as closed class words, have low weights and therefore little impact on the WOL match score. Lower frequency words (content words) have higher weights.

In order to assess the effects of ASR errors on AutoTutor’s evaluation of learners’ contributions, we first computed the LSA and WOL values associated with the automatically recognized students responses (hypothesized text) and expectations (correct answers). We then compared these values with the LSA and WOL scores associated with the manually transcribed contributions (reference text) and expectations. These results are presented in Figure 2A.

![Evaluation of Learners’ Contributions](image)

**Figure 2** Evaluation of Learners’ Contributions

The lack of statistically significant differences ($F < 1$) between LSA scores associated with comparisons of the hypothesized versus reference text to the expectations reveals that recognition errors have no impact on LSA-based measures of student contributions. Perhaps LSA will be a robust technique for learning environments with noisy speech.

In contrast to LSA, statistically significant differences between automatically recognized and manually transcribed contributions were discovered for the inverse word overlap measure, $F(1, 29) = 6.6$, $Mse = .006$, $p < .05$. This suggests that the use of WOL as a metric to evaluate spoken utterances might be problematic in environments with noisy speech.

The version of AutoTutor in this study used an amalgamation of LSA and WOL to evaluate the conceptual quality of a learner’s response. Answer quality was measured as a combined score: $CMB = 0.33 \cdot LSA + 0.67 \cdot WOL$. Statistical tests on the combined score indicated that there were no statistically significant differences for the combined score when the two versions of learner contributions (reference vs. hypothesized) were compared to expectations.

We also investigated the possibility of using content word overlap between student contributions and expectations as a potential measure of student answer quality. Figure 2B shows precision, recall, and F-Measure scores for this metric of comparisons for hypothesized-expectation and reference-expectation. Statistical tests on the F-measure revealed that there were significant differences between these two comparisons, $r(1, 27) = 6.301$, $Mse = .001$, $p < .05$, indicating that this metric is not robust enough to compensate for ASR errors.

3.3. Effects of ASR Errors on Learning

The results presented above indicate that AutoTutor’s evaluation module used to evaluate the quality of student responses was insensitive to speech errors. This suggests that shallow NLP-based methods, such as LSA (but not WOL or content word overlap), can be effective in environments with noisy speech recognition. However, a fundamental issue is whether learning is occurring.

We computed pre-test scores and post-test scores of participants before and after they interacted with AutoTutor. As expected, the post-test scores were significantly higher than the pre-test scores, $t(29) = 2.75$, $p < .05$, $S = .044$. So clearly, learning occurred during the 35-minute interaction with AutoTutor, showing an effect size of 0.50 sigma (approximately half a letter grade).

These results are significant because they indicate that learning gains can be achieved despite the fact that student’s responses were degraded by speech recognition errors. The 0.50 sigma effect size obtained in this new speech recognition version of AutoTutor is consistent with six previous experiments where learning gains were evaluated on approximately 500 college students [10].

Correlation analyses did not reveal any explicit relationship between recognition errors and pre-test and post-test scores. However, a negative correlation was found between the proportion of insertions (INS) and the learning measures. Specifically, INS correlates negatively with prior knowledge ($r = -.388$, $p < .05$) and post-test scores ($r = -.466$, $p < .01$). This implies that this type of recognition error is more prominent in low domain knowledge students who presumably speak with more uncertainty and less confidence, particularly due to their lack of familiarity with the topic.

3.4. Effects of ASR Errors on Learner’s Emotions

It is widely accepted that emotions are tightly bound to the learning process, in that they modulate and scaffold learning. However, it is important to distinguish the emotions that arise due to the tutoring content from those that occur as an artifact of the learning medium, i.e., the interface. Since current state-of-the-art speech recognition systems are far from perfect, it is likely that repeated recognition failures are able to be detected by learners, and subsequently invoke negative emotions that will potentially
disrupt the learning process. This is a legitimate concern to consider for AutoTutor versions with speech recognition.

Our study included an affect judgment procedure in which learners rated their affective states (boredom, flow, confusion, frustration, delight, surprise, and neutral) based on video recordings of their face and the audio captured during their interaction with AutoTutor. Correlational analyses indicated that there were no statistically significant relationships between ASR errors and the proportions of the various emotions experienced. However, the learners did experience appreciable amounts of boredom (M = .190), confusion (M = .188), flow (M = .232), and frustration (M = .129), which are on par with the state of neutral (M = .189). Experiences of delight (M = .051) and surprise (M = .021) were more rare, as we found in previous studies [8]. As expected, learning computer literacy with AutoTutor invoked emotions in learners. However, these emotions were not related to ASR errors.

3.5. Effects of ASR Errors on Participants’ Impressions
At the conclusion of the study, participants completed a short survey consisting of questions assessing their interaction with AutoTutor. No significant relationship was found between ASR errors and participant ratings on the following questions: (Q1) “I enjoyed interacting with AutoTutor”, (Q2) “I felt the conversational smoothness of my interaction with AutoTutor was comparable to an interaction with a human tutor”, (Q3) “I felt that AutoTutor understood what I said”, and (Q4) “I felt that AutoTutor was easy to use and work with.” However, a fine-grained analysis on the types of errors (Figure 1A) revealed that there was a significant negative correlation between deletion errors and participant responses to Q1 (r = -.477, p < .01) and Q4 (r = -.407, p < .05). With the exception of deletions, participants were not aware of most speech recognition errors. As shown in Figure 2A, AutoTutor’s shallow NLP based evaluations were resilient to these errors.

4. Conclusions
We have demonstrated that the speech-enabled AutoTutor can be effective in promoting learning despite significant recognition errors. This was achieved by using a shallow NLP technique (i.e., LSA) to evaluate the conceptual quality of learners’ contributions without adversely affecting learning gains and with minimal impact of ASR errors on learners’ emotions and overall interaction experience.

One of the major limitations of this research is that the current study did not adhere to an experimental paradigm. Therefore, at this time we are unable to compare our dependent measures obtained from the speech recognition version of AutoTutor with appropriate controls such as the typed contribution based AutoTutor or a wizard of oz type speech recognition condition (where an experimenter transcribes the participants speech and sends the error free transcribed text to AutoTutor). In order to address these limitations we are currently conducting a formal experiment comparing the speech-enabled and text-based interfaces of AutoTutor on measures of learning and user preference. In addition, we will conduct experiments that include other ASR systems in order to improve recognition accuracy.

We are also exploring the possibility of automatically detecting learner emotions from acoustic-prosodic features of speech as part of a larger project aimed at transforming AutoTutor into an affect-sensitive ITS [8]. Our hope is that an affect-sensitive AutoTutor with voice input will broaden the communicative bandwidth between the learner and the computer, thereby providing a more naturalistic interaction experience. This will presumably optimize learner satisfaction and positively impact learning.

5. Acknowledgements
We thank Bethany McDaniel and Kristy Tapp for their help with the data collection. This research was supported by the National Science Foundation (REC 0106965, ITR 0325428, and REC 0633918). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

6. References