

IBM Research Report

Usability and Learning in a Speech-Enabled Reading Tutor: A Field Study

Paula M. Bach

The Pennsylvania State University
University Park, PA 16802

Jennifer Lai

IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598



Research Division

Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Usability and Learning in a Speech-Enabled Reading Tutor: A Field Study

Paula M. Bach

The Pennsylvania State University
University Park, PA 16802
pbach@ist.psu.edu

Jennifer Lai

IBM T.J. Watson Research Center
Hawthorne, NY 10598
jlai@us.ibm.com

Abstract

In this paper we discuss outcomes of a field study conducted to evaluate usability and learning associated with a speech-enabled reading tutor application for adults. The evaluation compared learning outcomes and usability measures between two versions of the software, as well as with traditional classroom instruction. Findings indicate small usability improvements and no significant difference in learning between versions, and equivalent learning levels between the groups who used the software exclusively and those that had a teacher cover the material instead. We discuss these results and how challenges associated with the field evaluation may have impacted the findings.

Keywords

Usability evaluation methods, reading tutor, learning technology, speech recognition, field usability testing

ACM Classification Keywords

K.3.1. Computer Uses in Education: Computer-assisted instruction

Introduction

In this paper we report on a study that evaluated the usability and learning outcomes of a speech-enabled

reading tutor called Reading Companion. Two versions of the software were evaluated: an older desktop version that has been in use for years and a new web-based version that had been in beta testing for only a few months before the evaluation. Students participating in the evaluation were part of a welfare-to-work program called Job Readiness, and are primarily 'English as a Second Language' (ESL) learners. In some cases they learned how to read in their native language, while in others the education level did not go beyond the second grade. Due to the nature of this program, students are only at the site for a short number of weeks.

Reading Companion

Reading Companion consists of a speech recognition component, a student modeling component, and a Flash client that renders the e-Books on the web and controls a tutor character that interacts with the readers. Readers use the software to practice reading and receive feedback from the character who 'speaks' to them guiding them through the book. Currently the spoken prompts consist of pre-recorded human speech, although a move to using dynamically generated synthesized speech is being investigated. In addition to helping and coaching the readers, the character 'listens' to them read. Figure 1 shows the character in a listening mode. The software uses speech recognition technology to determine if the words are read correctly. If a match occurs, then the character provides positive feedback such as "you sound great!" If no match occurs, then the character prompts the user to repeat the incorrect word. If no match has occurred after several attempts (the number of attempts is determined by the reading level the reader has achieved) the tutor moves on to the next word or

phrase. The goal of the Reading Companion is to provide reading practice with feedback, much like a human reading tutor would confirm or correct words read by an emerging reader. Figure 2 shows an e-book displayed with the character waiting for the user to indicate that she is ready to read.

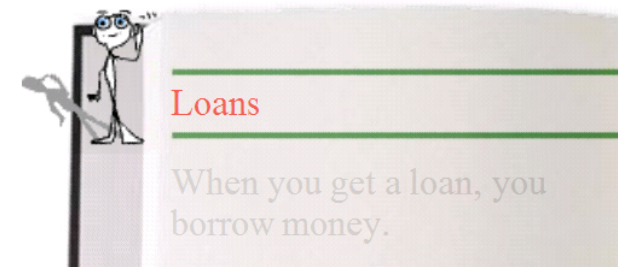


Figure 1. Tutor character 'listening' with text to be read highlighted

The move to a web-based version of the software was motivated by both the availability of better speech recognition technology (the old version did not use a continuous speech recognizer) and the desire to include student modeling functionality¹. These advantages are in addition to all the inherent advantages of a web-based solution. The goal of the evaluation was to test if learning or usability had been impacted, either positively or negatively, by the move from the older

¹ Information from the student modeling is directed at teachers to help them better understand the student's strengths and weaknesses. Teachers were not included as participants in this study and as such the paper does not report on the impact of this component.

version to the newer version and to compare learning outcomes with those of a traditional classroom setting.

Methodology and Results

We were intrigued with the ability to compare the two versions given the changes in speech recognition technology and function. We began by assessing the

functionality of both systems using an analytical heuristic method [2]. Next, we spent time in the field observing students and noting where errors occurred, or where the class instructor had to repeatedly show a student how to use a function.

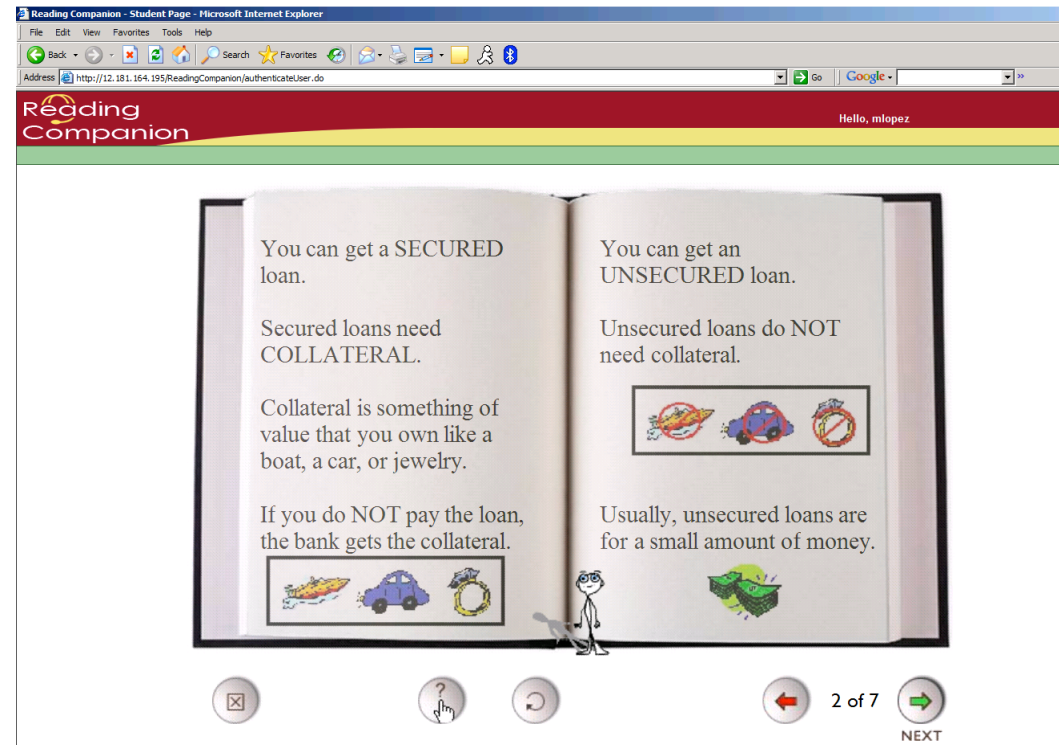


Figure 2. An e-Book with the tutor character waiting

We compared these field notes with the heuristic analysis and found common problems present in both.

We then developed a usability specification for performance with regard to number of errors and time

on task. We designed tasks to test performance for the common problems which included interaction with the character, clarity of the function icons, and navigation. In addition to performance measures, we collected subjective user reactions with a questionnaire using a five-point Likert scale. The questionnaire also included three open-ended questions that asked what the user liked, disliked, or would change in the system.

The measures were collected during class time over a period of five weeks with a total of 40 participants who were new to the Job Readiness program. They were randomly assigned into one of four groups, two of which used the Reading Companion software exclusively, one group received only teacher instruction and did not use the software, and the last group was a control group that took the test without the benefit of instruction.

Reading Companion is used during scheduled class times to read e-books. While some books are designed to increase vocabulary, others are for students to learn about specific topics such as applying for work, or conducting oneself appropriately in an interview. We chose to evaluate learning using the first e-Book in a series about applying for work because it is basic enough for new students and they are motivated to learn the content. In this book readers practice reading and learn how to fill in their personal information on a job application form.

In order to measure learning outcomes, we administered a test developed by the welfare-to-work program educators. The test asked students basic questions such as the types of information required for a job application. An example question: "Answer these

questions with full sentences - What is your first name?" Students also completed a partial job application as part of the test. The content of the job application e-Book covered all of the questions in the test. We conducted this learning evaluation in conjunction with the usability evaluation and thus used the same two groups of users. For the groups that used Reading Companion, 18 of the 20 participants were female, and all were Latina/o, which is a predominant characteristic of the user population at this adult literacy location. The age range was 22 to 51 with an average age of 33.

Results - Performance

Since speech recognition in the newer version of the software is taking place on a remote, network-connected server rather than on the client, the lag between speech and recognition was longer with the web version. However, in spite of this latency, time on task performance figures were significantly faster with the new version of the software, $t(16) = 2.28, p < .05$, than with the old. No significant difference was found for number of errors $t(15) = -0.06, p > .05$ between the versions. For 88% of the tasks, subjective measures indicated that users found the web-based system better in terms of confidence, fun, ease of use, and lack of confusion.

Results - Learning

To test if learning was affected by the change from desktop to web-based, test scores were compared between the group of users who only used the older desktop version ($M = 24.10, SD = 2.01$) and the group that used only the newer web-based version ($M = 24.35, SD = 2.27$), with no significant difference $t(18) = -0.26, p > .05$.

We also established a base level of test scores to see how learning with Reading Companion compared to learning in a traditional teacher-student classroom setting. Thus we compared test scores from the two groups that used the reading tutor, to both the control group and the teacher-only group. The control group consisted of 10 students in the program who took the test without the benefit of instruction. The classroom group consisted of 10 different students from the program, who took the test as part of a job readiness class. This class has an instructor preparing students for employment. Although the class did not read a specific book on how to fill in personal information on a job application, their classroom lessons covered all of the questions in the test. Both the reading tutor groups ($M = 24.10, SD = 2.01$), ($M = 24.35, SD = 2.27$) and the classroom group ($M = 21.85, SD = 2.38$), scored significantly higher than the control group ($M = 16.15, SD = 5.59$), ANOVA using Dunnett's test as post-hoc test, $F(3, 39) = 12.60, p < .05$. The results of the post-hoc test indicate that the higher scores were likely to be the result of learning delivered through the tutor or the job readiness class.

To evaluate how learning with the tutor compared with learning in the classroom we randomly selected 10 test scores from both of the tutor groups ($M = 23.60, SD = 2.13$) and compared them with the 10 test scores from the classroom group ($M = 21.85, SD = 2.38$). No significant difference was found between the tutor and the classroom when learning about completing job application forms $t(18) = 1.73, p > .05$.

While the findings indicate that the new system performed better than the old in terms of usability and learning, and compared favorably (as good) with

classroom learning, we faced several challenges evaluating learning and usability in the field.

Discussion

The challenge in evaluating learning and usability in the field is to design a methodology that produces useful results despite a chaotic environment that is ripe with potentially confounding variables. Below we discuss several challenges encountered in our field study that are not uncharacteristic of field studies.

For each task we compared errors, time to complete task, and subjective responses against targets in the usability specification. For 88% of the tasks, the subjective rating was higher than the observed performance, indicating that the users did not perceive the errors and longer times as problematic. One explanation may be that we underestimated the target goals in the usability specification. These targets were established based on results from informal testing with novice, intermediate, and expert users coupled with initial observations from the field.

Another explanation could be that users did not mind making mistakes since the experience of interacting with the character is fun enough to counterbalance the longer time spent on the task. Also the questionnaire measured slightly different factors than the performance testing. For example, questionnaire results indicate that users found the 'listening to a word' function very helpful. Users can click on an icon which puts the software into a mode that thereafter, any word they click on, will be read to them. We observed problems with both entering and exiting this mode. Users may find the function so helpful that they simply

disregard problems associated with entering and exiting the function mode.

Additionally, all of the users who participated in our evaluation were ESL speakers, which introduced a language barrier. Some of them had difficulty understanding the instructions for the task, questionnaire and questions. For these users the test was also difficult. We addressed this issue in several ways. The task instructions were spoken rather than having the users read them, and we made sure they understood the task before starting the timer. For the questionnaire, we moved to having the questions in both English and Spanish. For the test, participants were allowed to ask for help understanding the questions, and we had the instructor translate if necessary, but we were careful not to disclose the answers.

An additional challenge for the questionnaire is that this set of users was not familiar with a Likert scale. Near the end of the evaluation a different instructor was in the classroom and explained how the scale worked. While introducing a different instructor interferes with control this is a good example of uncontrollable challenges encountered in field studies.

Lastly, because participants are part of a welfare-to-work program, they may have thought that if they were too critical of the reading tutor, they might lose their training benefits. Even though we explained that their comments would only be used to make the system better, and that they could be as critical as they wanted, they did not produce any really negative comments about the system. Oliver [1] also noted that

getting users to feel empowered about learning with any learning technology evaluation is difficult. Given the language issue, they may not have known how to comment critically about the system's functionality.

Conclusion

Articulating the challenges of evaluating learning and usability in the field is a first step to refining methodologies so that field results can be useful. While our findings indicate that learning with Reading Companion compares favorably to classroom instruction and was not impacted by the move to a web-based solution, the challenges may point to some caution in interpreting the results. This study and the development of the underlying technology is part of an on-going effort to positively impact literacy in both adults (ESL population) and children. Further studies could look specifically at how the usability of a system affects learning.

Acknowledgements

We thank all participants and the administration at the adult literacy site for being so helpful in conducting this evaluation. We also thank Jack Carroll and Thomas George for comments on a very early presentation of this work that helped shape the ideas presented here.

References

- [1] Oliver, M. An introduction to the evaluation of learning technology. *Educational Technology & Society* 3, 2 (2000), 20-30.
- [2] Rosson, M.B. and Carroll, J.M. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. Morgan-Kaufmann, San Francisco, CA, USA, 2002