Evaluation of Response Time for Speech, Visual & Bi-Modal Cues

Hooman Baradaran
Department of Computer Science
York University
Toronto, ON
cs213240@cs.yorku.ca

ABSTRACT

Speech is expressive and can be used when doing other things, however, listening to speech is slower than reading text. Response time for listening to computer-generated speech as an alternative to visual cues in graphical user interfaces as well as their combination are evaluated. Computer-generated speech was used to instruct 11 participants to hit one of four numbered keys in equal distances from a start button after randomly listening to speech, reading a text or both. The mean response time for visual cue was significantly shorter than speech (1200 vs 1470 ms) and addition of speech to visual cues had no significant improvement (also 1200 ms). The participants did find the speech easy to understand, however, did not find it helpful in visual interfaces in general and further improvements were suggested.

Keywords

Interactive listening, text-to-speech, graphical user interfaces, speech user interfaces, multi-modal feedback

INTRODUCTION

Speech is a rich, expressive and powerful communication medium and can be used while doing other things[1]. There is an extensive amount of research on use of speech interfaces for the visually impaired users[2,3] and yet many researchers suggest that it is not enough. Addition of speech to graphical user interfaces to improve accessibility for older adults[4] is also an active area of research, however, at least up to my knowledge, there have been no published papers that study speech as an additional feedback and cue system in graphical user interfaces for average users. Unlike non-speech audio feedback and cues in visual interfaces[5], which in addition to research have been commercially used for many years, speech has never been added to visual interfaces. A possible reason is that while for input it is faster to speak than it is to write or type, it is an opposite case for output: it is slower to listen than it is to read. Speech is harder to skim and it is usually a burden on the listener[1]. While based on this claim, improvements have been suggested for systems that use speech-only interfaces, it is hard to find published experiments to support this claim, at least in available libraries of human computer interaction publications. Also, this claim is about reading text alone and not the time required to scan visual interfaces to find the message to read.

I present a user study to experimentally compare listening to speech with reading text and scanning visual interfaces based on participant's response time to a given cue. In addition, I examine whether the addition of speech to visual cues in a graphical user interface can have any improvements to this response time.

In addition to this quantitative comparison, a qualitative evaluation is performed to find out if users are able to understand currently available text-to-speech systems when compared to a real person's voice. The users were also asked whether they would desire speech in graphical user interfaces.

METHOD

A user study was carried out for a statistical comparison of participants' response time for speech and visual cues as well as bi-modal cues (their combination). The qualitative study of the speech prototype was performed via questionnaires and short informal interviews.

Participants

Eleven undergraduate university students (6 males and 5 females) all computer literate, working as part-time help desk staff, were selected for the user study. All participants were between ages 20-25 and in their final (third or forth) years on studies. This is a reasonable choice as age can be a factor in the response time to speech versus visual cues. Most participants studied either computer science or information technology with the exception of two who studied psychology and education. Only four participants indicated previous experience with text-to-speech systems and none of the participants was familiar with the prototype used. While English was the second language for some participants, they were all fluent in English.

Apparatus

The prototype was written in Java, allowing the experiment to be perform in various locations. The graphical user interface includes five buttons and a large text display for visual cues. One button placed in the middle of the screen was used as a start button for each trial and the other four buttons were numbered and placed in four corners in equal distances from the start button. The distance between the center of two targets was exactly twice the diameter of each target. The display would disappear when speech was the only method used, so a simple message was displayed after the program was started to indicated the position of the display. While visual tracking is needed, the position of the display is fixed since the position of items in actual graphical user interfaces does not change either. About five seconds after the start of the application, recorded instructions would start playing with the same computer-generated voice as the one used in the other parts of the experiment.

Festival Speech Synthesis System[6], an open source text-to-speech application developed at the university of Edinburgh, was used as the speech engine. While it is possible to use this system with Java, the libraries are written in C, requiring the system to be recompiled on different systems. To keep the portability of the prototype which was needed to find a good environment to perform the experiment, the Wizard of Oz technique was used and the default American male voice in Festival was prerecorded and played back during the experiment. This would not cause a problem in this user study since the instructions and the number of the key to press would not need to change during the experiment.

Procedure

The experiment was performed in a quiet sound-proof room to avoid any interruption and noise when listening to speech. A desktop computer running Windows with a usual mouse were available in the room and used for the experiment. The computer's speakers were used for the speech and volume level was set to highest. Glass walls on one side of the room allowed the experimenter to observe the participant without causing any distraction, however, when necessary the participant could talk to the experimenter and ask questions.

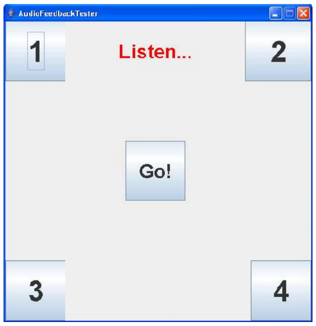


Figure 0. The system at start

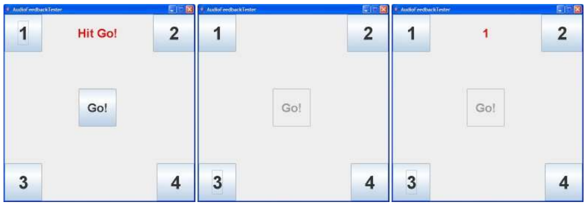


Figure 1. The system during experiment. Left to right: before each trial, speech only cue, visual or bi-modal cue

After filling out a pre-questionnaire, each participant was asked to listen to the recorded instructions and wait for the experimenter to go back into the room in order to answer questions and clarify instructions if necessary. Another reason the participants were asked not to start the experiment immediately was that, due to incorrect intonations of the speech system, there was a delay after the phrase “after you hit go...” in the instructions and a few of the initial participants thought the system was actually asking them to hit the start button right after the phrase was heard. It was required that the participants listen to all of the instructions and do not start. Also if the participants started too soon and the the first trial needed to play back another speech phrase, the second voice would not be heard and the experiment would have to be repeated with a new set of data for that participant.

All participants were ready to perform the experiment after listening to the recorded instructions and did not require any further clarification.

Each trial involved pressing the start button followed by a number button. The number of the button to press was given randomly via one of the three modes. The participants performed all trials simultaneously.

In the end the participants were asked to fill out the post questionnaire and provide their feedback on the system in a very short interview. The complete experiment took less than 10 minutes on average.

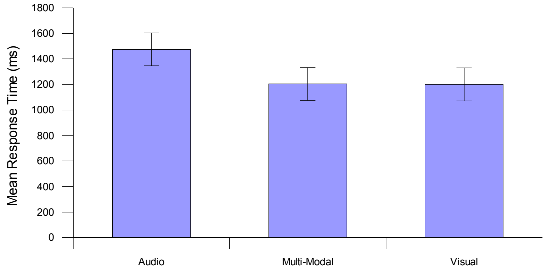
Design

The task and factors (audio, visual or bi-modal) were within subjects and used in random orders. The input tasks for each participant included all 12 possible combinations (four trials for each factor) and were examined to make sure different orders are practiced in order to counterbalance any possible learning effect. While no practice session was used, one purpose of the recorded instructions was to make the participants familiar with the computer-generated speech.

Trials were recorded in the order performed by each participant and for each trial the key pressed, mode of the given cue, number of errors made and the response time in milliseconds was recorded.

The response time recorded for this experiment was the time between a press of the start button and the press of the correct target button. If an incorrect button was pressed, an error would be recorded and the timer would keep

Figure 2. Cue mode vs. response time



incrementing until the correct target key was pressed. None of the participants made any error during the experiment so no error factor was considered in the analysis of the results.

RESULTS AND DISCUSSION

Response Time

Mean response time for speech, visual and bi-modal cues, as shown in figure 2, were 1470, 1200, 1200 milliseconds respectively. The analysis of variance (ANOVA) for response time showed a significant difference for the cue mode factor ($F_{2,20}$ = 10.8, $p < .005$). There was no significant learning effect for the order of the number key pressed for each cue mode ($F_{3,30}$ = 2.50, ns).

Furthermore, the response time for visual cue was significantly faster than speech alone ($F_{1,10}$ = 12.5, $p < .05$) and the addition of speech to visual cues had no improvement over visual cues alone ($F_{1,10}$ = 0.004, ns).

Based on these results, it is now proven that listening to speech is slower than reading text. We can also conclude that speech is slower that reading and visual scanning together. At least when dealing with speed and response time for average users, graphical user interfaces are fast enough without speech.

Qualitative Results and More Observations

In the questionnaire the participants were asked to rate their level of understanding of the speaker. To make sure they answer "how well they can understand the system" rather than "how natural the computer-generated voice feels [in comparison to a real person]", the participants were first asked to compare the speaker with a real person's voice in a separate question before rating their level of understanding of the system. The purpose of the later was to know if currently available text-to-speech systems, regardless of the artificiality of the speaker, are clear and understandable enough to be used instead of natural speech. On a 7-point Likert scale (1 = not understandable at all, 4 = somehow understandable, 7 = very well understandable) all participants gave a rating of 4 or higher to this question. The answer on average was 5.5 and the median was 6.

Those participants who also gave a comment on improvements, wanted the speaker to be more natural and clear. They also expected better intonations and stops in future versions. Some suggested a higher volume, specially for the speech-only mode. Some participants did not like the selected voice and one participant suggested a female voice would be easier to understand for her. Considering Festival is an educational open-source system that is under development and that some of these issues are resolved in commercially available systems, text-to-speech systems are understandable enough to be used instead of natural voice where necessary. Based on this and the quantitative results, we can expect to see slower response time for speech even if the text-to-speech system is replaced by speech recorded by a human.

The participants were also asked whether they would like their operating system to "speak back to them" and on a similar scale of 1 to 7 (1 = never, 4 = sometimes, 7 = always) no preference was given on average. Both the mean and the median for this question were 4, however, based on the interviews and written comments, those who did want speech in the operating system found it "interesting" rather than useful.

CONCLUSION

This experiment has proven that in practice, listening to speech is slower than reading text and visually scanning graphical user interfaces. Based on this and the qualitative results, there is neither any improvement nor any demand in adding speech to graphical user interfaces, however, this experiment does not consider possible benefits adding speech can have in productivity.

These results show that when an speech interface is needed current text-to-speech technology is usable enough to replace natural speech and more effort must be put in speech recognition rather than it's synthesis.

REFERENCES

1. Arons, B. SpeechSkimmer: a system for interactively skimming recorded speech. In *ACM Transactions on Computer-Human Interaction (TOCHI)*, ACM Press (1997), 3-38.

2. Mynath, E. D., Weber, G. Nonvisual presentation of graphical user interfaces: contrasting two approaches. In *Proceedings of CHI '94*, ACM Press (1994), 166-172.

3. Pitt, I. J., Edwards, A. D. N. Improving the usability of speech-based interfaces for blind users. In *Proceedings Of second ACM conference on Assistive technologies*, ACM Press (1996), 124-130.

4. Emery, V. K., Edwards, P. J., Jacko, J. A., Moloney, K. P., Barnard, L., Kongnakorn T., Sainfort F., Scott I. U. Toward achieving universal usability for older adults through multi modal feedback. In *Proceedings of 2003 conference on Universal Usability CAPH 2003*, ACM Press (2003), 46-53.

5. Albers, M. C., Bergman, E. The audible web: auditory enhancements for Mosaic. In *Proceedings of CHI '95*, ACM Press (1995), 318-319.

6. Festival Speech Synthesis System http://festvox.org/festival