Test Procedure

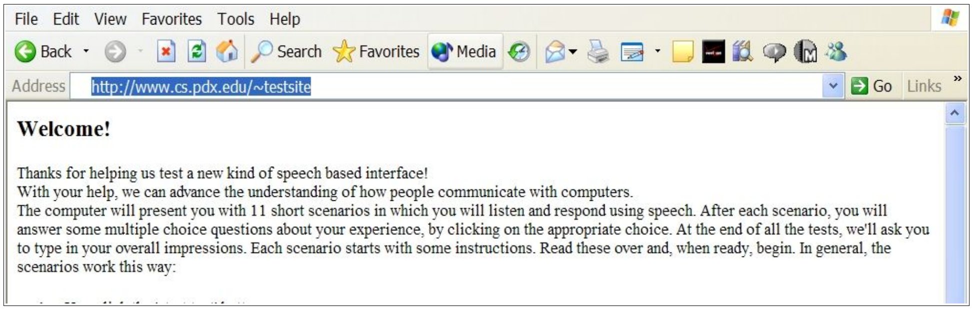
The usability tests consisted of two test phases separated by one week. We tested an initial set of interactors in phase 1, then analyzed the results, made changes to the interactors, and tested the changed interactors in phase 2.

The usability tests were performed on laptop computers configured with Internet Explorer, Microsoft Speech Application SDK version 1.0, and a headset (with both headphones and microphone).

Both usability tests took place in a centrally located, quiet classroom. The test subjects were observed to record their behavior during the test. The test subjects were all volunteers, with about half recruited from the Portland State University campus, and about half recruited from a local church group. Volunteers were over 18 years of age, but were otherwise of diverse ages and backgrounds. We did not collect or use demographic information in recruitment of volunteers – volunteers simply needed to be 18 and without a thick accent. We did not accept computer science students.

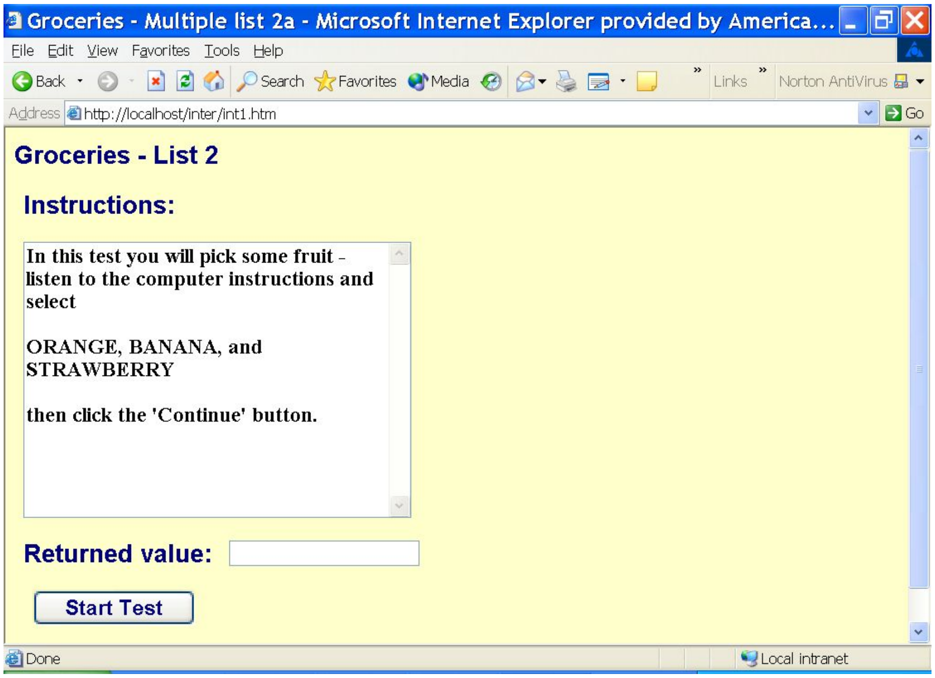
Each test session consisted of 10 to 11 individual test cases, one for each interactor, with the order of test cases randomly selected to avoid any training bias in the results. The test cases were html pages that used a simple test script to run the interactor appropriately and record performance metrics. Each test case consisted of 2-3 test scenarios, in which the test subject was asked to use the interactor in a different way. Each test subject was given a short set of verbal instructions before the test session began. The test procedure for each test case went as follows:

1. The test case is presented to the test subject, with instructions in the 'instructions' box. A sample test case is displayed below:


File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://www.cs.pdx.edu/~testsite
Go
Links

Welcome!

Thanks for helping us test a new kind of speech based interface!
With your help, we can advance the understanding of how people communicate with computers.
The computer will present you with 11 short scenarios in which you will listen and respond using speech. After each scenario, you will answer some multiple choice questions about your experience, by clicking on the appropriate choice. At the end of all the tests, we'll ask you to type in your overall impressions. Each scenario starts with some instructions. Read these over and, when ready, begin. In general, the scenarios work this way:


Groceries - Multiple list 2a - Microsoft Internet Explorer provided by America...
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Links
Norton AntiVirus
Address
http://localhost/inter/int1.htm
Go
Start Test
Done
Local intranet

Groceries - List 2

Instructions:

In this test you will pick some fruit - listen to the computer instructions and select

ORANGE, BANANA, and STRAWBERRY

then click the 'Continue' button.

Returned value:

Figure: 1 – Example test case

2. Test subjects read the instructions and click on the start button.
3. The test script starts up the interactor.
4. The interactor executes, recording performance metric data on the fly, and returns a result.
5. The test script examines the result returned from the interactor. If the result is correct, the test script goes on to the next test scenario; if not, it will repeat the current test scenario at most twice before going on.
6. After each test case is complete, the performance metric data are recorded in the database and the test subject is presented with a set of questions that make up the preference metrics for the interactor.
7. The test subject fills in the answers to the preference metric questions and submits the form.

8. The preference metrics are recorded in the database and a new test case is presented. This process is repeated until all the test cases have been executed.
9. At the end of all the test cases, the test subject is presented with a screen that allows them to type in textual comments

Metrics

In order to measure the usability of the software, we employed two types of metrics: preference metrics and performance metrics. Preference metrics are a subjective measurement of how the test subject evaluated the software – whether it was easy to use, if he/she enjoyed the experience, etc. Performance metrics are objective measurements of the results of the test subject's interaction with the software – how often errors occurred, how often the test subject spoke a word that was not understood, etc.

Preference Metrics

Preference metrics were measured on a scale from 1 to 5. In phase 1 of the tests, the test subject was simply asked to give a numerical rating for each. In phase 2, we assigned specific words to each of the numeric measurements. Not all preference metrics were used with all interactors.

Clarity – The test subject was told, "Rate how well you understood what the computer wanted you to do." In phase 2, the value assignments were: 1=very poorly, 2=somewhat poorly, 3=ok, 4=well, 5=very well

Effectiveness – The test subject was told, "Rate how well you were able to accomplish your task." In phase 2, the value assignments were: 1=very poorly, 2=somewhat poorly, 3=ok, 4=well, 5=very well

Ease of use – The test subject was told, "Rate how easy this program was to use." In phase 2, the value assignments were: 1=very difficult, 2=somewhat difficult, 3=ok, 4=easy, 5=very easy

Ease of Multiple Selection – The test subject was told, "Rate how easy it was to select multiple items." In phase 2, the value assignments were: 1=very difficult, 2=somewhat difficult, 3=ok, 4=easy, 5=very easy

Confirmation – The test subject was told, "Rate how you liked the confirmation of your choices at the end of the dialog." In phase 2, the value assignments were: 1=did not like at all, 2=liked somewhat, 3=neutral, 4=liked, 5=liked a lot

Deselection – The test subject was told, "Rate how well changing your selections at the end of the program worked." In phase 2, the value assignments were: 1=very poorly, 2=somewhat poorly, 3=ok, 4=well, 5=very well

Prompt Speed – The test subject was asked, "Were the options provided too slow, too fast, or just right?" In phase 2, the value assignments were: 1=very slowly, 2=somewhat slowly, 3=just right, 4=somewhat quickly, 5=very quickly

Speech Preference – The test subject was asked, "Do you prefer to say the items you want immediately, or select them from a list of options." In phase 2, the value assignments were: 1=I strongly prefer to speak the items, 2=I slightly prefer to speak the items, 3=I like having both options, 4=I slightly prefer to select from a list, 5=I strongly prefer to select from a list

Second Prompt – The test subject was told, "Rate how helpful the second prompt was." This metric was not used in the second round of testing.

Performance Metrics

The following performance metrics were used for these tests, not all metrics were used with all interactors.

Time per iteration – The time (in milliseconds) it took to make a single choice in the interactor.

Error Rate – The rate at which the interactor returned incorrect results – this is distinct from bad response rate, because it judges the end result of the interactor execution, and not any intermediate data. The error rate was measured as a percentage of the total number of results returned from the interactor.

Bad Response Rate – The rate at which the test subject uttered a word or phrase that the interactor did not understand. The bad response rate was measured as a percentage of the total number of times the interactor was listening for a response.

No Response Rate – The rate at which the test subject said nothing when the interactor expected a response. The no response rate was measured as a percentage of the total number of times the interactor was listening for a response.

Second-level No Rate – The rate at which test subjects answered in the negative when asked to confirm their response. The second level no rate was measured as a percentage of the total number of results returned from the interactor.