# Advanced SALT
# Interactor Usability Study

## Sponser
## Jim Larson

## Team Members


Pauline Amal

Chuck Banaka

Mark Cowlishaw

Tim Smith

Mabel Pecos

Yevgeniya Yufereva

# Advanced SALT
# Interactor Usability

# Introduction

Speech Application Language Tags (SALT) are generic markup tags that can be used to implement speech-based interfaces on top of SALT-enabled browsers. SALT provides low-level speech-to-text, text-to-speech, and telephony services so that application developers can incorporate speech into their applications. SALT provides a great deal of flexibility to its users, it enables developers to produce highly usable and sophisticated speech-enabled applications, but it also enables complex and unusable user interfaces.

One method for solving this problem is to provide higher-level user interface components using speech (high-level speech interactors) based on SALT that developers interactively evaluate. If the high-level interactors are well designed and usable, they should make it easier to develop speech-enabled applications with effective user interfaces.

We used SALT, JavaScript and XHTML to implement high-level speech interactors. We performed experiments to evaluate the usability of the interactors using performance and preference metrics.

In this paper we report the results of the usability tests. This report contains a summary of the test procedure, a description of the metrics used in testing, an analysis of test results for each SALT interactor, test limitations, and valuable lessons that we learned from this experiment. As an addendum, we provide the raw test data used in the report.

To view the state transition diagrams for the interactors:

**http://salt-usability.sourceforge.net/main/8other/InterStateMachines.htm**

# 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 convenient place for test subjects. The test subjects were observed in order 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 to 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 case began. At the end of each test case the test subject was asked to rate the test case (see preference metrics) and were also encouraged to enter comments. Each test case took between 2 to 3 minutes each. A complete test session generally took between 20 to 30 minutes per test subject.

Test sessions were controlled by a cgi script that determined the random order of the test cases, and entered the preference, performance, and comment data collected by each test case into a database.

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:

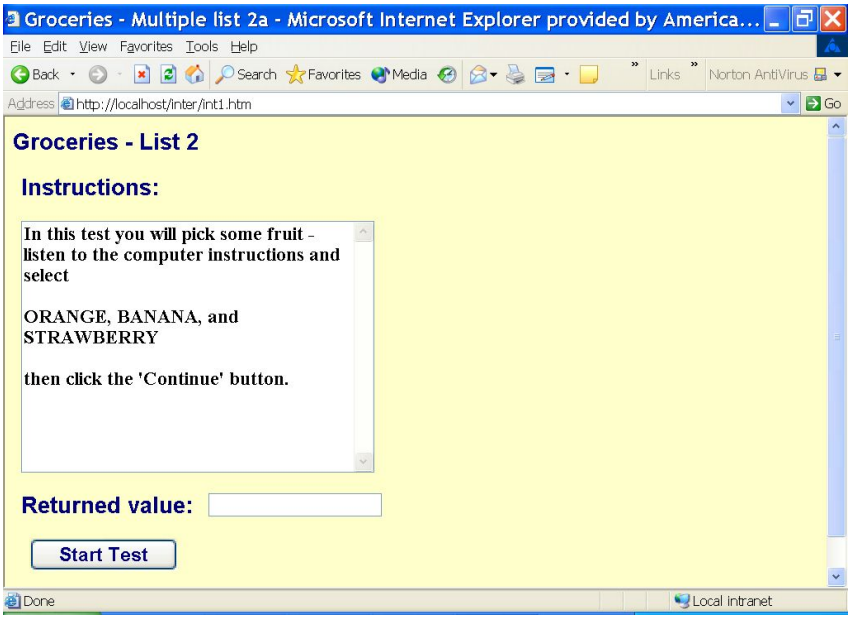


Figure: 1 – Example test case

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. All metrics were gathered electronically by each test case.

Metric targets were our best guess arrived at by experience gained during testing of each interactor and its test case.

## *Preference Metrics*

Preference metrics were measured on a scale from 1 to 5. Test subjects were asked questions which helped to fill in the preference metrics. In phase 1 of the tests, the test subject was simply asked to give a numerical rating for each test case where 1 was generally described as very poor or difficult and 5 described as very good or easy. In phase 2, we assigned specific words to each of the numeric measurements.

Both phases used a web page presented after each test case to collect preference metrics. Phase 1 used a web page at the end of the test session to collect comments for "Best Liked Features", Least Liked Features", "Desired Features", and "General Comments" after the entire test session was complete. In phase 2 we added a text area to the performance metric page for each test case that allowed us to gather comments specific to that test case.

Not all preference metrics were used with all interactors.

**Clarity** – The test subject was instructed, "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 instructed, "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 instructed, "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 instructed, "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 instructed, "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 instructed, "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 instructed, "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.

Definitions:

Low level responses

- Match: when the user says something that is recognized by the grammar.
- No match: when the user says something that is not recognized by the grammar.
- No response: when the user does not respond before the listen times out.

Dialog level responses

- Correct: when the user says something that is a match and is also what the test scenario expected.
- Incorrect: when interactor returns something that is a match but is not what the test scenario expected.
- NULL: interactor returns NULL if user does not select anything.

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

elapsed time /(null + incorrect + correct)

**Error Rate** – The rate at which the interactor returned incorrect results – this is distinct from No Match 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 and is a dialog-level metric.

(null + incorrect) /(nulls + incorrect + correct)

**No Match Rate** – The rate at which the test subject uttered a word or phrase that the interactor did not understand. The No Match rate was measured as a percentage of the total number of times the interactor was listening for a response.

no match /(no match + match + no 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.

(no response) /(match + no match + no responses )

# Interactor Usability Test Results

## *Interactor 01 (Single List Selection)*

The Single List Selection Interactor allows the user to make a single selection from a limited, predefined list of choices, similar to a drop-down menu in a traditional dialog.

The interactor reads each item in the list and prompts the user to speak after an item is read to select the item.

This analysis is based on 27 valid tests in phase 1 and 30 valid tests in phase 2.

## Preference Metrics

| **Phase 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Clarity | 4.19 | 4.5 | 2 | 3 | 1 | 3 | 18 |
| Effectiveness | 4.15 | 4.5 | 3 | 1 | 1 | 6 | 16 |
| Ease of Use | 4.22 | 4.5 | 2 | 2 | 1 | 5 | 17 |
| Prompt Speed | 3.33 | 3 | 0 | 1 | 19 | 4 | 3 |
| **Phase 2** | | | | | | | |
| Clarity | 4.67 | 4.5 | 0 | 0 | 2 | 6 | 22 |
| Effectiveness | 4.5 | 4.5 | 0 | 0 | 3 | 9 | 18 |
| Ease of Use | 4.57 | 4.5 | 0 | 0 | 2 | 9 | 19 |
| Prompt Speed | 3.07 | 3 | 0 | 4 | 22 | 2 | 2 |

For this interactor, clarity, effectiveness, and ease of use should be rated “good” (4) or “excellent” (5), while prompt speed should be rated 3 (just right).

In phase 1, clarity, effectiveness and ease of use metrics were just slightly below their target value, but still in the acceptable range (above 4). Prompt speed was almost exactly on target as well.

In phase 2, users rated this interactor exceptionally high: clarity, effectiveness, and ease of use metrics each met their goals, and no user gave a rating below “average” (3). Prompt speed was within 7/100 of a perfect score.

## Performance Metrics (Phase 1)

| **Phase 1** | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 17512 | 14000 | 14501 | 7104 |
| Error Rate | 0.12 | 0.10 | 0.0 | 0.22 |

| | | | | |
|---|---|---|---|---|
| No Match Rate | 0.03 | 0.05 | 0.0 | 0.07 |
| **Phase 2** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 14744 | 14000 | 14611 | 737.5 |
| Error Rate | 0.077 | 0.10 | 0.0 | 0.13 |
| No Match Rate | 0.015 | 0.05 | 0.0 | 0.06 |

In phase 1, Time per Iteration was off by a few seconds, but the median was very close to the desired time. In phase 2, Time per Iteration improved to very near the desired result.

In phase 1, the Error Rate is slightly higher than we would like – although the median suggests that most users encountered no errors, and that a few users who had problems are skewing the average. A slight change to improve accuracy may be indicated. In phase 2, the Error rate met its target with room to spare – clearly the interactor in phase 2 is much more accurate.

In phase 1, the No Match Rate was very close to the target, and indicates little change. There was some correlation between error rate and No Match rate. In phase 2, the interactor was even better.

# Comments (Phase 1)

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- Choosing options in a single list – saying just 'Yes' or 'Ok' or 'No'.
- Preferred to enter my own choices, rather than waiting for a list and responding ok.

**Least Liked Features**

- The pause was long between choices.

# Comments (Phase 2)

- Good job!
- I'd like to just say what I want. After the third vegetable, I almost forgot to say OK. Maybe others are better listeners
- In the first set of instructions it was confusing as to whether or not I could just say "orange." I realized in the second set of instructions that I could say the name of my choice only after the computer program spoke it.
- Rather than say "OK", I sometimes feel like repeating the name of the vegetable or fruit. I did feel rushed answering.

- In the written instructions, the computer said to say the name of the fruit or vegetable, but did not say that choices would be listed. In the verbal instructions, it gave the option of saying okay or the name of the fruit or vegetable. It felt sort of inconsistent.
- The program did not like me putting an "e" on the end of the vegetable

# Analysis (Phase 1)

The preference metrics were quite close to target – the interactor was well liked and the users seemed to think the speed of the supplied prompts was about correct.

The performance metrics were also close to target – however, the error rate was somewhat higher than we would like. The correlation between error rate and No Match Rate seems to indicate that users are speaking but are not understood and our own observations indicate that users sometimes say the name of the item they want to choose, rather than using an affirmative response, like 'OK'. Allowing the user to say the name of the item chosen appears to be indicated.

Finally, the time per iteration metric was much higher than desired – our observations indicate that users may be waiting before they proceed with tests, unsure that the interactor has completed working, which may be resulting in a false time measurement. A clear indicator of the interactor's finish may help to improve this metric.

# Analysis (Phase 2)

The preference metrics all met their targets, showing a strong, positive change from the phase 1 interactor. No user rated the interactor below average, which is also an excellent sign of improvement.

The performance metrics all improved significantly, each meeting its target.
Clearly, the interactor was improved dramatically from phase 1 to phase 2.

# Changes from Phase 1 to Phase 2

**Change 1**

**What**: Allow the user to say the name of the item when read, as well as 'yes', 'ok', and other affirmative responses.

**Why**: The No Match Rate is somewhat correlated with the error rate, indicating that users are sometimes saying words that were not understood. Our own observations indicate that some users prefer to say the name of the item when it is read.

**How to Measure**: If it is successful, this change should result in a drop in the error and No Match Rates and perhaps a slight increase in the 'ease of use' preference metric.

**Did it work**?: Yes. The error rate did in fact drop significantly, the No Match Rate also fell, and the 'ease of use' metric was also improved,

**Change 2:**

**What**: Repeat the chosen item at the end of the interactor

**Why**: In comments, users indicated a strong preference for verbal confirmation of their choices.

**How to Measure**: If it is successful, this change should result in a small drop in the 'time per iteration' metric, since users will know immediately when the interactor has completed.

**Did it Work**?: Yes. The time per iteration fell by more than three seconds, on average, and times were much more consistent across users. Hearing the verbal confirmation helped to trigger the test subject to press the continue button that would either start the next scenario for the test case or start the next test case.

## Future Changes

The following changes may be indicated in the future.

1. There are still a significant number of users that found the prompting too fast or too slow. Since some users where on either side of this issue, it may be that prompt speed needs to be tailored individually to each user. Measuring response time and adjusting the prompt speed accordingly may be indicated.
2. Many users prefer to speak the name of the item they want, rather than hearing a list, while others prefer a list – it may be desirable to allow the user to simply speak the name of an item and to provide the list only if the user speaks an item that is not on the list, or asks for help explicitly.
3. From observation, users seem to like being able to "barge in" while a prompt is being spoken, however, this creates a problem with excess noise, which can be read as an invalid (or valid!) user response. Special handling of SALT's bargein event may cut down on the effect of excess noise.

## *Interactor 02 (Multiple List Selection without De-selection)*

The Multiple List Selection Interactor allows the user to select one or more items from a short, fixed list of selections, similar to a multiple selection box in a standard graphical interface.

In phase 1, the user was prompted to speak an affirmative after the desired item was read, and was asked if he wanted to select more items after each selection.
In phase 2, we changed the interactor significantly, creating two separate interactors based on the original (the Multiple List Interactor and Help List Interactor).

The Multiple List Interactor prompts the user as the original, but allows the user to speak the name of the item as well as the affirmative. After a selection, the interactor continues reading the list until the user indicates that they have finished their selections.

The Help List Interactor allows the user to simply say the name of each desired item, and indicate when they have completed their selections. If the user speaks a name that is not in the list, or asks for help, the interactor reads the list as in the normal Multiple List Selector.

This analysis is based on 26 valid tests of interactor 2 in phase 1 and 29 valid tests of the Multiple List Interactor and 28 valid tests of the Help List Interactor in phase 2.

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below. The "Speech Preference" metric (for the Help List Selector only) is a measurement of whether the user prefers to speak their choices immediately (1), or speak as the items are listed (5).

| **Phase 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **# 1** | **# 2** | **# 3** | **# 4** | **# 5** |
| Clarity | 4.69 | 4.5 | 0 | 0 | 1 | 6 | 19 |
| Effectiveness | 4.54 | 4.5 | 1 | 0 | 2 | 4 | 19 |
| Ease of Use | 4.54 | 4.5 | 0 | 1 | 1 | 7 | 17 |
| Ease of Multiple Selection | 4.42 | 4.5 | 0 | 2 | 1 | 7 | 16 |
| **Phase 2 (Multiple List Selector)** | | | | | | | |
| Clarity | 4.45 | 4.5 | 0 | 1 | 3 | 7 | 18 |
| Effectiveness | 4.31 | 4.5 | 0 | 1 | 5 | 7 | 16 |
| Ease of Use | 4.28 | 4.5 | 0 | 2 | 4 | 7 | 16 |
| Ease of Multiple Selection | 4.34 | 4.5 | 0 | 3 | 3 | 4 | 19 |
| **Phase 2 (Help List Selector)** | | | | | | | |
| Clarity | 4.46 | 4.5 | 0 | 1 | 4 | 4 | 19 |
| Effectiveness | 4.36 | 4.5 | 0 | 2 | 3 | 6 | 17 |
| Ease of Use | 4.5 | 4.5 | 0 | 1 | 2 | 7 | 18 |
| Ease of Multiple Selection | 4.29 | 4.5 | 0 | 2 | 5 | 4 | 17 |
| Speech Preference | 2.39 | 3 | 9 | 1 | 17 | 0 | 1 |

In phase 1, the only metric that missed its target was ease of multiple selections, which also had the fewest ratings of excellent (5). If we need to make changes in this interactor, a good target would be to make it easier to select multiple items.

In the phase 2 Multiple List Selector, the average metric rating went down very slightly across the board, although all were still close to their desired level, and there were no "1" ratings.

In the phase 2 Help List Selector, the average metric rating was nearly identical to the Multiple List Selector. Speech Preference showed a tendency to prefer speaking the names of the items, but most users preferred to have both choices.

## Performance Metrics

| **Phase 1** | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 11891 | 11000 | 12142 | 2079 |
| Error Rate | 0.198 | 0.10 | 0.0 | 0.191 |
| No Match Rate | 0.06 | 0.05 | 0.0 | 0.07 |
| **Phase 2 (Multiple List Selector)** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 11812 | 11000 | 10833 | 1001 |
| Error Rate | 0.157 | 0.10 | 0.0 | 0.160 |
| No Match Rate | 0.027 | 0.05 | 0.0 | 0.068 |
| **Phase 2 (Help List Selector)** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 10284 | 11000 | 8642 | 2417.5 |
| Error Rate | 0.116 | 0.10 | 0.0 | 0.136 |
| No Match Rate | 0.02 | 0.05 | 0.0 | 0.046 |

In phase 1, Time per Iteration was off by almost a second, and the median suggests that the average underestimates the amount of time taken for most users – a slight change in the interactor may be indicated.

The phase 2 Multiple List Selector showed time per iteration close to the target and the median was below it, indicating a slight improvement over phase 1.

The phase 2 Help List Selector showed a marked improvement in time per iteration, putting it well within the target.

In phase 1, the Error Rate was nearly twice what we would like – although the median suggests that most users encountered no errors, and that a few users who had problems were skewing the average. Clearly, though, we must make the interactor more accurate for those users who had problems.

The phase 2 Multiple List Selector showed a marked improvement in error rate from phase 1, but the interactor still failed to meet the target.

The phase 2 Help List Selector showed an even greater improvement in error rate.

In phase 1, the No Match Rate was very close to the target, but may provide the key to improving the interactor. The median indicates that most users did not encounter any problems, however, there is a fairly strong correlation between those users who encountered bad response problems, and those who had a non-zero error rate (8 of the 9 non-zero error rates also encountered bad response problems). This indicates that expanding the vocabulary of the interactor may have a significant impact on the overall error rate.

In the phase 2 Multiple List Selector, our changes significantly improved the No Match Rate, allowing it to meet its goal.

The phase 2 Help List Selector showed even greater improvements in the No Match Rate.

# Comments (Phase 1)

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- Choosing options in a list – saying yes or no rather than naming the things I wanted
- Picking multiple items from a list
- I liked having verbal confirmation of my choices
- I liked having the computer repeat my choices
- Preferred to say what I wanted, rather than picking from a list

**Least Liked Features**

These comments were given as the features of the interactors that the user liked least. Comments are paraphrased to protect the user's identity:

- The timing of the speech seemed unnatural. Waiting to choose items from a list seemed impractical.
- The pause between choices was so long that I didn't know if there were more items in the list
- Wasn't sure whether to say "yes" or "ok"
- The choices were read somewhat slowly

**Desired Features**

These comments were given as the features the user would like to see in the interactors. The comments are paraphrased to protect the user's identity:

- Don't list choices. I'd prefer to say the things I want rather than waiting for them to be read from a list.

**General Comments**

These are general comments that the users gave. The comments are paraphrased to protect the user's identity:

- Narrow responses to "yes" or "no", instead of "ok"
- Written instructions should tell you to say, "ok" after a response, and "yes" to read more. Additionally, I sometimes didn't have enough time to respond to the list.

# Comments (Phase 2, Multiple List Selector)

- Not sure if I was supposed to say "okay" after the fruit/veggie and "done" at the end??
- The computer didn't hear my response to the celery prompt on the first try.
- I find it difficult to remember the instructions!
- I felt a little rushed between the program's naming of vegetables.
- Program responded equally well to both "done" and "complete"
- Couldn't understand kiwi
- I did not need to say, "done"; the program prompted an end. However, the directions told me to say, "done."

# Comments (Phase 2, Help List Selector)

- On the fruits the computer never gave me a chance to choose one or the other option. It repeated the instructions too fast.
- I am beginning to notice the voiceless consonants of the program, particularly the "f", are a little soft. Also, when I am choosing two items, I want to say "and", but I am worried that it will confuse the program.
- Directions were much clearer than the previous questions regarding options.
- MORE DIFFICULT
- It never did accept my third fruit.
- I like having both options, however, I was not exactly sure if I did or not on the first part.

## Analysis

Other than the error rate, this interactor performed very close to how we expected in phase 1. It was quite highly rated by users and often cited as one of their favorite interactors. It appears that even where problems occurred, they applied to a minority of users. The main improvement needed is in diminishing the rate of errors.

The high error rate and strong correlation between errors and "no match" bad responses seem to indicate improved handling of bad responses. Additionally, the comments indicate and our observations during the tests confirm that some users prefer to name the items they want rather than waiting for the responses in the list. Together, these suggest allowing the user to simply name the items he wants and to use the list as a "fall back," in case the user is not understood, or needs help with his choices.

In phase 2, both changed interactors showed better performance than the phase 1 interactor, but the performance gain did not correlate with a similar gain in user preference. This does not quite meet with observations, as users often cited the Help List Interactor as their favorite. The differences in the interactor preference metrics appear to be insignificant.

The solid performance gains (as well as a comparison to interactor 3 in phase 2) show that users prefer to indicate when they are finished making selections, rather than being asked if they want to continue after each selection.

## Changes

The changes are given in priority order:

**Change 1**

**What**: Allow the user to speak the name of a choice, in addition to saying "yes", "ok," or another affirmative response.

**Why**: No Match Rate correlated with error rate may indicate that the user was saying the name of the item, rather than saying "ok." Additionally, comments indicate and our own observations support that many users find this more natural.

**How to Measure**: Should cause a drop in the No Match Rate and a drop in the error rate if successful.

**Did it Work**: Yes. Both phase 2 interactors showed a marked improvement in both error rate and No Match Rate.

**Change 2**

**What**: Allow the user to speak the names of the items desired BEFORE a list is presented. Only present the list of items if the user does not respond, is not understood, or says "help" or "list." This should probably be a separate interactor.

**Why**: Comments indicate that some users did not like waiting for a list, and felt that it was boring to wait for the choices to be read. Our own observations confirm this. Combining both methods of communication would accommodate both impatient and deliberate users.

**How to measure**: Should cause a drop in the error rate. Additionally, should ask whether the user prefers to say the items themselves, or select them from a list.

**Did it work**: Yes. The Help List Interactor proved to have the lowest error rate of any multiple list interactor. Additionally, users indicated that they liked having the option of speaking items right away or selecting them from a list, with a small preference toward speaking the items at once.

**Change 3**

**What**: Change the interactor to continue reading the list of items after a correct response, until the user says "done," "complete," or "finished," rather than querying after each selected item.

**Why**: The ease of multiple selections was the only preference metric that did not meet its target – not requiring an additional "yes" to continue reading the list may make multiple selections easier. Additionally, this change would decrease the number of times we wait for a user response, which could decrease the amount of time taken, as well as improving the No Matchand error rates.

**How to measure**: Should result in a drop in the time per iteration, as well as slight decreases in No Match and Error rates.

**Did it work**: Again yes. The median time per iteration dropped significantly in both phase 2 interactors, and both showed significant improvements in the and error rates.

**Change 4**

**What:** Change the interactor to confirm the user's choices at the end. Say, "I didn't detect any selections" if no choices are made.

**Why:** Comments indicate that users appreciate having verbal confirmation of their choices – if change 3 is completed, we will need another mechanism for confirming their choices

**How to measure:** Should result in decreases in time per iteration, as well as small increases in Ease of Use and Clarity preference metrics.

**Did it work:** The result is mixed. Both phase 2 interactors showed an improvement in time per iteration and no significant change (in fact a slight drop) in Ease of Use and Clarity. It appears that the use of a confirmation had some impact on time per iteration, but little impact on the users' impressions of the interactor. This may be simply because all interactors confirmed the user's choice at the end in phase 2 testing.

## Future Changes

1. From observation, users seem to like being able to "barge in" while a prompt is being spoken, however, this creates a problem with excess noise, which can be read as an invalid (or valid!) user response. Special handling of SALT's bargein event may cut down on the effect of excess noise.

## *Interactor 03 (Multiple List Selection with De-selection)*

The Multiple List Selection with De-selection interactor allows the user to select one or more items from a short, fixed list of choices, similar to a multiple selection box in a standard graphical interface.

The phase 1 interactor prompted users to speak an affirmative after their desired choice, and asked if they wanted to select additional items after every selection. At the end of all choices, the interactor re-read the list of choices, prompting the user to say 'no' after any undesired selections. The main idea was that the user would be able to correct false positives without re-running the interactor entirely.

In phase 2, the de-selection prompt is changed so that the user can say the item's name or 'yes' (instead of no) for deselection. This interactor served as a control to the other multiple list interactors, as it asked the user if he wanted to continue after each selection. This analysis is based on 26 valid tests of interactor 3 in phase 1 and 27 valid tests in phase 2.

## Preference Metrics

### Phase 1

| Metric | Mean Score | Target | # 1 | # 2 | # 3 | # 4 | # 5 |
|---|---|---|---|---|---|---|---|
| Clarity | 3.96 | 4.5 | 0 | 3 | 6 | 7 | 11 |
| Effectiveness | 3.81 | 4.5 | 1 | 2 | 7 | 8 | 9 |
| Ease of Use | 3.92 | 4.5 | 0 | 3 | 6 | 7 | 10 |
| Ease of Multiple Selection | 3.85 | 4.5 | 1 | 2 | 7 | 7 | 10 |
| Confirmation | 3.89 | 4.5 | 0 | 2 | 8 | 8 | 9 |
| De-selection | 3.59 | 4.5 | 2 | 3 | 7 | 7 | 8 |

### Phase 2

| Metric | Mean Score | Target | # 1 | # 2 | # 3 | # 4 | # 5 |
|---|---|---|---|---|---|---|---|
| Clarity | 3.74 | 4.5 | 1 | 4 | 5 | 8 | 9 |
| Effectiveness | 3.37 | 4.5 | 3 | 6 | 4 | 6 | 8 |
| Ease of Use | 3.56 | 4.5 | 0 | 7 | 6 | 6 | 8 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Ease of Multiple Selection | 3.70 | 4.5 | 2 | 3 | 6 | 6 | 10 |
| Confirmation | 3.44 | 4.5 | 1 | 5 | 5 | 13 | 3 |
| De-selection | 3.26 | 4.5 | 2 | 6 | 7 | 7 | 5 |

In phase 1, each performance metric fell significantly short of the goal
In phase 2, the shortfall was even worse.

# Performance Metrics

## Phase 1

| Metric | Mean Score | Target | Median | Standard Deviation |
|---|---|---|---|---|
| Time per Iteration | 19149 | 18000 | 20240.5 | 2898.19 |
| No Match Rate | 0.05 | 0.05 | 0.04 | 0.06 |
| Error Rate | 0.20 | 0.10 | 0.19 | 0.17 |

## Phase 2

| Metric | Mean Score | Target | Median | Standard Deviation |
|---|---|---|---|---|
| Time per Iteration | 17250 | 18000 | 19524 | 3247 |
| No Match Rate | 0.034 | 0.05 | 0 | 0.052 |
| Error Rate | 0.34 | 0.10 | 0.22 | 0.27 |

In phase 1, the No Match Rate met its goal, but the error rate and time per iteration fell significantly short.
In phase 2, time per iteration and No Match Rate both met their goals, but the error rate increased dramatically.

# Comments (Phase 1)

The following comments addressed this interactor:

**Best Liked Features:**

"Allowing to pick from multplies of things and have them on a list."

"Preferred those where I could enter my own choices, rather than waiting while a list was enumerated and responding ok"

"I liked the that the computer confirmed what was ordered on the lists of items."

"I liked that they repeated what I chose rather than just putting the answers in the boxes. It is nice to get verbal confirmation."

### Least Liked Features

"Having to listen to listen and say OK after the correct answer seemed in- practical, but being able to speak the direct destinations was nice."

"Some ambiguity on saying "Yes" or "ok" on some of the question options made it slightly confusing."

### Desired Features

"Where possible, avoid naming lists of choices for users to respond "ok" on their choices ... Id rather state "carrots, broccoli, eggplant" myself than wait thru a list until my choice(s) come up."

### General Comments:

"It would be helpful to tell participants to follow the directions explicitly. The way the first sentence is often worded (You will be asked to tell your favorite city, for example) suggests the participant can give his or her own opinion."

"I would keep response to "yes" or "no", dropping the "ok" response"

## Comments (Phase 2)

- Seems pretty easy to use.
- Rome sounded like Rone.
- Saying "OK" to choose things is not vey intuitive. And saying "yes" to exclude athe than "exclude" is also not vey intuitive.
- I was never given a second country on the first try. The second try was successful. I was not asked to change choices at the end of the program, so the last question was confusing.
- Wasn't sure what to do when confirming cities the first time through.
- Instruct to say yes after the first city is selected if another city is to be selected also.
- I think that I made an error in the last test.
- I didn't like saying "Yes" to a negative confirmation. Perhaps adding another word, such as, "delete" would feel better for me.
- skipped the part of asking for including more counties in the final Question
- It seemed a little counter-intuitive to respond with a YES for something you wished exluded. I would have thought NO would make more sense.
- Asking me to say "yes" to a country I wish to exclude gives a positive response to a negative question.

- i was confused at first -- it took me a while to get the hang of the what the appropriate response should be. after i got familiar with it, it was fine.
- I did this test twice and both times it had difficulity with excluding countries. The first time when I excluded France, it left me with Spain, the second time it left me with England.
- I guess I was confused about what I was to do. I never was able to select a second city only the first and then it would read them back

# Analysis

In phase 1, all preference metrics were below 4. The lowest preference metric was for de-selection from the list. Comments seemed to support changing selection of an item from saying ok during the listing to saying the name of the city. Comments and the poor preference for deselection suggested a change to the way that deselectionw as done.

In phase 1, Time per Iteration and No Match Ratewere right on the target amounts however the Error Rate was high at .20 or 20 percent mean and 19 percent median values where we have a target of 10 percent. In general there were more missing responses (29) than extra response (18) that might indicate that people were having problems selecting items into the list. People who had the highest error rate also tended to have more iterations and more no responses indicating that their responses weren't registering or that they didn't have enough time or weren't sure how to respond.

In phase 2, while allowing the user to use the name of the item selected certainly helped improve the No Match Rate, the error rate skyrocketed and user preference metrics plummeted. From the comments and observation it's clear that there were two main problems with this interactor and the test surrounding it:

1. Users liked saying 'yes' to deselect a choice even less than they liked saying no. This resulted in many users not knowing what to do at the end of the interactor.
2. Many users were trained by the other list selectors that the computer would continue the list after they had made a selection – so many failed to respond when asked if they wanted to select additional items

# Changes (Phase 1 to Phase 2)

### Change 1

**What**: Allow the user to say the name of an item to select it, rather than an affirmative (like OK).

**Why:** The No Match Rate was slightly higher than we wanted, and somewhat correlated with the error rate – indicating that users were consistently saying some theings that the interactor did not understand. Additionally, our observations indicated that some users expected to be understood when speaking the name of a selection, despite the instructions.

**How to Measure:** Should result in a drop in the No Match Rate, and a slight drop in the error rate.

**Did it Work:** Yes. The No Match Rate decreased and users who spoke the names of the items were understood. However, the drop in error rate was not achieved, perhaps due to user expectations from similar interactors and due to the deselection problem.

**Change 2**

**What:** Deselect user choices on a 'yes' response (instead of a 'no' response)

**Why:** Our observations indicated that users where confused by the deslection process – saying 'no' to a choice seemed conunterintuitive.

**How to Measure:** Should result in a decrease in the error rate, and increases in the 'Clarity'; and "Effectiveness' preference metrics.

**Did it Work:** No. Users emphatically disliked saying 'yes' to deselect an option. The error rate increased and the Clarity and Effectiveness metrics decreased.

**Change 3**

**What:** Confirm user choices when the interactor has finished.

**Why:** User comments indicate a strong preference for having their choices confirmed verbally, Additionally, our observations indicate that sometimes users wait when an interactor has completed, unsure that it is finshed – verbal confirmation will provide an additional cue that the user is supposed to continue with the test.

**How to Measure:** Should result in a drop in the time per iteration.

**Did it work:** Yes, time per iteration dropped from phase 1 to phase 2.

# Future Changes

1. The entire idea of deselection needs to be overhauled. Perhaps the user should simply be asked if he made any incorrect selections, and THEN prompted for deselection, Additionally, perhaps the user should be asked if he wants to include each item, with a yes or no response expected. The question, "Do you want to exclude this item," seems to be inherently confusing. Asking, "Do you want to include this item," may be easier to understand.
2. There are still a significant number of users that found the prompting too fast or too slow. Since some users where on either side of this issue, it may be that prompt speed needs to be tailored individually to each user. Measuring response time and adjusting the prompt speed accordingly may be indicated.
3. Many users prefer to speak the name of the item they want, rather than hearing a list, while others prefer a list – it may be desirable to allow the user to simply

speak the name of an item and to provide the list only if the user speaks an item that is not on the list, or asks for help explicitly.

4. From observation, users seem to like being able to "barge in" while a prompt is being spoken, however, this creates a problem with excess noise, which can be read as an invalid (or valid!) user response. Special handling of SALT's bargein event may cut down on the effect of excess noise.

## *Interactor 04 (Double Yes No Using Speech)*

When asked yes/no question, the user says something else (yeah, may be, neh …) Will confirm unclear response by asking "Did you say Yes/No?" Will be used on error (user says something but yes/no) or on low confidence level (user says something close to yes/no).

Computer asks "Would you like fries with your meal?"
User responds "Yes"
If confidence is less than a threshold then the computer asks "Did you say yes?"
User responds "Yes"

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below. The preference metrics used were:
Here is a summary of the measurements:

| **Phase 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Clarity | 4.33 | 4.5 | 0 | 1 | 4 | 7 | 15 |
| Effectiveness | 3.52 | 4.5 | 3 | 5 | 4 | 5 | 10 |
| Ease of Use | 3.81 | 4.5 | 1 | 3 | 6 | 7 | 10 |
| **Phase 2** | | | | | | | |
| Clarity | 4.56 | 4.5 | 0 | 2 | 1 | 4 | 20 |
| Effectiveness | 3.59 | 4.5 | 4 | 5 | 2 | 3 | 13 |
| Ease of Use | 4.22 | 4.5 | 0 | 2 | 5 | 5 | 15 |

## Performance Metrics

For this interactor, we used the following non standard performance metrics:

**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.

($2^{nd}$-level no's) /(match + no match + no response)

Here is a summary of the measurements:

## Phase 1

| Metric | Mean Score | Target | Median | Standard Deviation |
|---|---|---|---|---|
| Time per Iteration | 10153 | 9000 | 9238 | 3713.21 |
| No Match Rate | 0.40 | 0.05 | 0.40 | 0.27 |
| Error Rate | 0.56 | 0.10 | 0.75 | 0.32 |
| Second Level | 0 | 0 | 0 | 0 |

## Phase 2

| Metric | Mean Score | Target | Median | Standard Deviation |
|---|---|---|---|---|
| Time per Iteration | 7173 | 9000 | 7088.5 | 1061 |
| No Match Rate | 0.28 | 0.05 | 0.2 | 0.21 |
| Error Rate | 0.46 | 0.10 | 0.33 | 0.31 |
| Second Level | 0.76 | | 0 | 0.87 |

# Comments (Phase 1)

**Best Liked Features:**

**Least Liked Features**

"The word "neh" in the 2nd fries questions. Was this supposed to be "no"? "

"When I said Neh, it didnt respond with anything and I didnt know what to do. "

**Desired Features**

**General Comments**

# Comments (Phase 2)

- "Petty good."

- "Very use friendly."
- "EASY"
- "Replied yes or no per instructions; pogram did not seem to detect my responses very well, even tho I changed the orientation of the microphone several times"
- "The frys question was not picking up my voice and regestering it on the screen."
- "DId not detect my YES an d NO also when asked  was that a NO I answered YES kind of odd"

# Analysis

Phase 1:

So, all three metrics met their targets, these ratings don't indicate any change by themselves.

Time per Iteration was off by more than two seconds, however, the median is very close to the target, indicating that the average is overestimating the time taken, due to a few users who had difficulty.

The Error Rate is right on target, and the median suggests that it overestimates the error rate – in fact, only 3 users experienced any errors whatsoever.
The No Match Rate was a little high, and, since this interactor expects the user to speak more varied responses, this may be due to similar-sounding words in the grammar, or difficulty understanding particular users.  This is especially true, since the low median indicated that most users experienced no problems. Since this did not correlate particularly with the error rate, it appears that the interactor may have overcome any shortcomings in the recognition engine.  Small changes in the grammar may be indicated by this result.

The No Response Rate was quite low, which suggests that users were not confused about what to say and generally had enough time to respond.

In all important respects, this interactor performed very close to how we expected.  It appears that even where problems occurred, they applied to a very small minority of users. The main improvement needed is in decreasing the No Match Rate.

The somewhat high No Match Rate, along with some of the textual comments suggest that users may not have understood the written instructions on what to say and that the grammar used for the test contain some choices that are too close together phonetically.

Phase 2:

In general the performance test results were close to or bettered our goals. The most notable improvement was in Time per Iteration. The most notable worsening was in the No Response Rate which was the only measure that exceeded our goal.

The preference results were all a little worse than the first test. Changes to the test application and the webpage that we used to gather the preference data may have contributed to the changes.

Clarity:

There was a slight improvment (4.33 first test, 4.56 second test) between the first phase and the second phase. The principle change effecting clarity between the tests was to provide verbal confirmation of a choice and changes to the test application.

Effectiveness:

The difference in effectiveness essentially remained the same (3.52 first test, 3.59 second test) between the first phase and the second phase. The low numbers for both phases could be related to the poor error rate. As noted below both phases generated a lot of nulls which means that the interactor had trouble understanding people.

Ease of Use:

Ease of Use increased (3.81 first phase, 4.22 second phase) from the first phase. This could be attributed to the changes made to the test application that presents instructions when needed and the addition of verbal confirmation.

Error Rate: (nulls + incorrect)/iterations

The error rate improved (0.56 first phase to 0.46 second test) from the first phase however both phases were well above our goal of 0.10. The largest contributing factor for both phases is the number of nulls indicating that there may be some problem with the grammar (unable to find a match perhaps because the grammar for this interactor has several phonetically similar elements), or timing (running out of time).

No Match Rate: (Number of No-Matches)/(Number of listens)

The No Match Rateimproved from the first phase 0.40 to 0.28 though it was still below our goal of 0.10.

Second Level: In the first phase this measured when the second level answer was "no". There was a problem with the interactor where no one reached the second level. In the second phase we were able to implement the confidence level of the first yes/no answer as an ambiguous answer causing people to reach the second level. We also changed the interactor to use this measure to count how often people reached the second level. With 2 different scenarios per test subject, people reached an average of .76 second levels. That is out of the 2 opportunities for reaching the second level people reached it 38 percent of the time.

Of the people who reached the second level:

8 people reached the second level twice:

2 of the 8 reached the second level and the correct answer for each scenario on the first iteration.

6 reached the second level and the correct answers in three iterations with 1 no match

6 people reached the second level once:

4 reached the second level once in three iterations with one correct answer.
1 correct, 2 nulls
2 reached the second level once in four iterations with one correct answer.
1 correct, 3 nulls

The number of nulls indicates that these people had problems with the second level listen element understanding their response which could be related to the problems with the error rate.

# Changes (Phase 1 to Phase 2)

The changes are given in priority order:

**Change 1**

**What:**
Provide verbal verification or no response detected message.

**Why:**
In the first test there were some comments that verbal response would be preferred to seeing the response in a text box.

**How to measure:**
Should be a decrease in the number of incorrect responses if test subjects understand what their response was.

**Did it work?**
There was a decrease in the number of incorrect responses also ease of use increased.

**Change 2**

**What:**
Change the scenario to use yes and no for the correct first responses.

**Why:**
The first phase indicated that the interactor was having problems dealing with the more ambiguous “neah”, and “yeah”.

**How to measure:**

There should be a decrease in the number of nulls.

**Did it work?**

The number of nulls decreased from 72 to 38 indicating an improvement.

**Change 3**

**What:**

Implement using the confidence level to generate ambiguous answers.

**Why:**

Explore the area between a match and a confidence level threshold. Using that grey area to generate a second level disambiguating prompt.

**How to measure:**

Should see second level entries.

**Did it work?**

Yes. 14 of the test subjects reached the second level at least once.

## Future Changes

**Change 1**

Simplify the grammar used for the first yes/no prompt to reduce the number of nulls by using fewer phonetically similar elements.

**Change 2**

Use more fields to record the different combinations of answers:

No/no
No/yes
Yes/yes
Yes/no

## *Interactor 05 (Double Yes/No using visual dialog)*

When asked yes/no question, the user says something else (yeah, may be, neh …) Will confirm an unclear response by displaying a dialog box asking the original question. Will be used on error (user says something but yes/no) or on low confidence level (user says something close to yes/no).

Computer asks “Would you like fries with your meal?”
User responds “Yes”
If confidence is less than a threshold then the computer asks “Would you like fries with your meal?”
User responds “Yes”

## Preference Metrics

For this interactor, most preferences should be rated “good” (4) or “excellent” (5). We would like to minimize the number of metric scores of “average” (3) or below. The preference metrics used were:

Clarity – How well did you understand what the computer wanted you to do?
Effectiveness – How well were you able to accomplish your task?
Ease Of Use – How easy was this program to use?
Here is a summary of the measurements:

| **Phase 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Clarity | 4.08 | 4.5 | 2 | 1 | 2 | 9 | 12 |
| Effectiveness | 3.54 | 4.5 | 1 | 4 | 8 | 6 | 7 |
| Ease of Use | 3.65 | 4.5 | 1 | 5 | 4 | 8 | 8 |
| **Phase 2** | | | | | | | |
| Clarity | 4.43 | 4.5 | 1 | 0 | 2 | 8 | 17 |
| Effectiveness | 4.04 | 4.5 | 2 | 1 | 4 | 3 | 13 |
| Ease of Use | 4.21 | 4.5 | 0 | 2 | 4 | 8 | 14 |

## Performance Metrics

For this interactor, we used the following non standard performance metrics:
2nd level no rate – the percentage of the time that the user said the opposite of what computer understood - taken as percentage of the total number of choices. We would like to keep this rate below 5 percent.
Here is a summary of the measurements:

| **Phase 1** | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 10073 | 10000 | 9535 | 2458 |
| No Match Rate | 0.33 | 0.10 | 0.29 | 0.25 |
| Error Rate | 0.508 | 0.10 | .5 | 0.307 |
| Second Level | 0 | | 0 | 0 |

| Phase 2 | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 7913 | 9000 | 8391 | 2035.85 |
| No Match Rate | 0.11 | 0.05 | 0.0 | 0.19 |
| Error Rate | 0.29 | 0.10 | 0.0 | 0.32 |
| Second Level | | | 0.0 | |

# Comments (Phase 1)

**Best Liked Features:**

**Least Liked Features**

- Having to say "neh".
- Contradicting instructions.
- Was confused by the instructions.
- Computer stopped after saying "heh" and was unsure how to proceed.

**Desired Features**

**General Comments**

- Need to work out bug.
- Has conflicting instructions.

# Comments (Phase 2)

- I'm not sure if the problem was due to my not speaking clearly or if this is to test what happens when there are problems - even so, corrections are rather straight forward and easy to make. I hesitated when the program asked me to confirm a "No" answer, because I needed to confirm the answer with a "Yes." I was worried, incorrectly albeit, that the program wouldn't differentiate between the two.
- Much better!
- This is a very cool application!!!

- What was the flash of confirming the response? Since the computer did not detect my answers to either question.
- Having the pop up box to verify a "yes" or "no" choice seemed redundant. It would be quicker to just click a box to begin with.
- The delay between opening the small sound meter attached to the pointer icon caused me to have to repeat answers several times.
- It may be simple to stat the program by clicking "stat" and then giving the directions of what to expect.
- Unsure about the confirmation box that popped up. I didn't know if it popped up because the computer didn't hear my verbal answer.
- could not detect my answer
- Was easy to understand
- I did not like to be asked to answer and then have to click on the answer again.
- I had problems as I was speaking the answers before the computer was ready for my response.

# Analysis

Phase 1:

N/A

Phase 2:

Preference metrics improved across the board although all three preference measures still did not meet our goals.

Performance metrics also improved across the board.

Clarity:

There was good improvement (4.08 first phase, 4.43 second phase) between the first phase and the second phase. The principle change effecting clarity between the tests was to provide verbal confirmation of a choice and changes to the test application and providing instructions in context instead of all at the beginning of the test.

Effectiveness:

The difference in effectiveness essentially remained the same (3.54 first phase , 4.04 second phase) between the first phase and the second phase. The low numbers for both phases could be related to the poor error rate. As noted below both phases generated a lot of nulls which means that the interactor had trouble

understanding people. Although using the dialog box seems to have improved the error rate from Double Yes/No using Speech test.

Ease of Use:

Ease of Use increased (3.6 first phase, 4.21 second phase) from the first phase. This could be attributed to the changes made to the test application that presents instructions when needed and the addition of verbal confirmation. Note that the second phase Ease of Use with the Double Yes/No using Speech test was almost identical at 4.22

Error Rate: (nulls + incorrect)/iterations

The error rate improved considerably (0.508 first phase to 0.29 second phase) from the first phase although still worse than our goal. The largest contributing factor for both phases is the number of nulls indicating that there may be some problem with the grammar (unable to find a match perhaps because the grammar for this interactor has several phonetically similar elements), or timing (running out of time). This measure was better than that for the Double Yes/No using Speech test possibly because using the dialog box eliminates the errors that would have been generated using the grammar for the second Yes/No.

No Match Rate: (Number of No-Matches)/(Number of listens)

The No Match Rateimproved from the first phase 0.33 to the second phase 0.11 which brought it very close to our goal of 0.10.

Second Level:

# Changes (Phase 1 to Phase 2)

The changes are given in priority order:

**Change 1**

**What:**

Provide verbal verification or no response detected message.

**Why:**

In the first test there were some comments that verbal response would be preferred to seeing the response in a text box.

**How to measure:**

Should be a decrease in the number of incorrect responses if test subjects understand what their response was.

**Did it work?**

There was a decrease in the number of incorrect responses also ease of use increased.

**Change 2**

**What:**
Change the scenario to use yes and no for the correct first responses.

**Why:**
The first phase indicated that the interactor was having problems dealing with the more ambiguous “neah”, and “yeah”.

**How to measure:**
There should be a decrease in the number of nulls.

**Did it work?**
The number of nulls decreased from 55 to 20 indicating an improvement.

**Change 3**

**What:**
Implement using the confidence level to generate ambiguous answers.

**Why:**
Explore the area between a match and a confidence level threshold. Using that grey area to generate a second level disambiguating prompt.

**How to measure:**
Should see second level entries.

**Did it work?**

# Future Changes

**Change 1**

Simplify the grammar used for the first yes/no prompt to reduce the number of nulls by using fewer phonetically similar elements.

**Change 2**

Use more fields to record the different combinations of answers:

No/no
No/yes
Yes/yes
Yes/no

## *Interactor 06 (N-Best List )*

Confirm an answer to a "textual" question using a proximity threshold in the n-best list. After user's response, get the list of words that match that response the most, and prompt the user again with that list of words: "Did you say item1, item2 or item3?"

e.g. "What city would you like to visit?" User says "Athens". Computer response: "More than one city matches your choice. Please indicate which of the following cities you meant by saying, Oak A, after the correct city is read."

User says yes after hearing Athens Greece. (The city user wanted to visit was Long View)

First Phase:
This analysis is based on 25 valid tests of interactor 6, for which all 25 returned preference metrics.

Second Phase:
This analysis is based on 28 valid tests of interactor 6, for which all 28 returned preference metrics.

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below. The preference metrics used were:

Prompt Speed – Was the prompt speed, listing options too slow (value of 1), just right (value of 3), or too fast (value of 5)

Here is a summary of the measurements:

| **Phase 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Clarity | 4.76 | 4.5 | 0 | 0 | 0 | 6 | 19 |
| Effectiveness | 4.56 | 4.5 | 0 | 0 | 3 | 5 | 17 |
| Ease of Use | 4.48 | 4.5 | 0 | 1 | 2 | 6 | 16 |
| Prompt Speed | 3.2 | 3 | 0 | 4 | 15 | 3 | 3 |
| **Phase 2** | | | | | | | |
| Clarity | 4.61 | 4.5 | 0 | 0 | 4 | 3 | 21 |
| Effectiveness | 4.5 | 4.5 | 0 | 2 | 3 | 2 | 21 |
| Ease of Use | 4.54 | 4.5 | 0 | 0 | 4 | 5 | 19 |
| Prompt Speed | 3.39 | 3 | 0 | 1 | 19 | 4 | 4 |

## Performance Metrics

Here is a summary of the measurements:

| **Phase 1** | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 17646 | 17000 | 16861 | 1994 |
| Error Rate | .06 | 0.10 | 0 | .109 |
| No Match Rate | .009 | 0.05 | 0 | .031 |
| **Phase 2** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 11664 | 17000 | 11697 | 1084 |
| Error Rate | .1 | 0.10 | 0 | .17 |
| No Match Rate | .08 | 0.05 | 0 | .12 |

## Comments (Phase 1)

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- Responding after prompt to select item.
- Prompt speed was not too fast or slow.

**Least Liked Features**

These comments were given as the features of the interactors that the user liked least. Comments are paraphrased to protect the user's identity:

- Computer was supposed to give "Paris Texas" as a choice and it didn't.

**Desired Features**

These comments were given as the features the user would like to see in the interactors. The comments are paraphrased to protect the user's identity:

- None were noted.

**General Comments**

These are general comments that the users gave. The comments are paraphrased to protect the user's identity:

- None were noted.

# Comments (Phase 2)

"Vey use fiendly."
"What options?"
"i dont know if the options wee actually povided too quickly, but i felt the need to huy with my answe."
"The wod choice fom t he voice was petty loud and freakish"

# Analysis (Phase 1)

Time per Iteration was off by 646 milliseconds.

The Error Rate exceeded our target goal.
The No Match Rate exceeded our target goal.

# Analysis (Phase 2)

In general the test results were close to our goals. The most notable improvement was in Time per Iteration.

Clarity:

There was a difference of 3.2 percent (4.76 first phase, 4.61 second phase) between the first phase and the second phase. The principle change effecting clarity between the phases was to provide verbal confirmation of a choice. Both numbers are better than our goal. This possibly correlates with a slight worsening of Error Rate from 0.06 to 0.1 and the No Match Rate from 0.009 to 0.08.

Effectiveness:

The difference in effectiveness was essentially unchanged (4.56 first phase, 4.50 second phase).

Ease of Use:

Ease of Use increased (4.48 first phase, 4.5.4 second phase) slightly from the first phase correlating to a 34 percent improvement of Time per Iteration from 17646 to 11664. The changes in Time per Iteration were primarily due to changes made to the test application template making the flow of the interactor scenario more intuitive.

Some of the increase in Ease of use may also be attributed to the addition of verbal confirmation and moving the instructions to be in context with the scenario instead of at the beginning of the test.

Prompt Speed:

Interestingly while we didn't change the interactor prompt speed this worsened slightly from 3.2 to 3.39 with 3.0 representing the goal. Out of the eight people who gave scores other than 3, seven thought that the prompt speed was too fast. Possibly some people were using this measure to register their preferences for the speed of the new test application template.

Error Rate:

The error rate increased slightly from 0.06 to 0.1 from the first phase but still meets our goal 0.1. However the makeup of these results is quite different. The error rate is calculated as (Incorrect Responses + Nulls)/ (Number of Iterations). The Number of Iterations changed from a mean of 3.24 iterations per test subject to 2.18 due to changes in the number scenarios from 3 to 2.

No Match Rate:

The No Match Rate is the (Number of No-Matches)/(Number of listens). As it was for the error rate the number of listens decreased from a mean of 6.56 per test subject to 3.29 probably due to the changes to the test scenarios.

However the Number of No-Matches increased from 2 in the first phase to 7 in the final phase. Meaning more people responded with an answer that was not in the grammar.

# Changes from Phase 1 to Phase 2

### Change 1

**What:**

Provide verbal verification or no response detected message.

**Why:**

In the first phase there were some comments that verbal response would be preferred to seeing the response in a text box.

**How to measure:**

Should be a decrease in the number of incorrect responses if test subjects understand what their response was.

**Did it work?**

The number of incorrect responses did decrease from 2 in the first phase to 0 in the second phase.

## Future Changes

### Change 1

Increase listen time between n-best list elements to give people more time to answer.

## *Interactor 07 (n-ary prompt)*

Interactor 7 seeks to correct bad user responses with a set of clarifying prompt. Clarifying prompts are played (and the interactor listens again) if the user does not respond, gives an invalid response, or asks for help.

This analysis is based on 27 valid tests of interactor 7 (for which 26 returned preference metrics) in phase 1.

Because this interactor scored well and is similar to interactor 8, it was not tested in phase 2.

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below.

| Metric | Mean Score | Target | # 1 | # 2 | # 3 | # 4 | # 5 |
|---|---|---|---|---|---|---|---|
| Clarity | 4.69 | 4.5 | 0 | 0 | 3 | 2 | 21 |
| Effectiveness | 4.73 | 4.5 | 0 | 0 | 2 | 3 | 21 |
| Second Prompt | 4.18 | 4 | 2 | 0 | 4 | 5 | 15 |

So, Clarity, Effectiveness and second prompt rated just above their targets. This interactor was not rated on ease of use due to a glitch in the preference metric form.

## Performance Metrics

Here is a summary of the measurements:

| Metric | Mean Score | Target | Median | Standard Deviation |
|---|---|---|---|---|
| Time per Iteration | 15980 | 12000 | 15162 | 3801 |
| Error Rate | 0.018 | 0.10 | 0.0 | 0.064 |

Time per Iteration was off by more than two seconds, however, significant numbers of bad responses and no responses were not in evidence – by itself, this metric does not suggest a change..

The Error Rate scored better than the target and the median suggests that it overestimates the error rate – in fact, only 2 users experienced any errors whatsoever.

## Comments

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- Quick feedback; indication of doing something with answer.
- Easy to understand and spoke clearly.
- The speed of the interactor was not too fast or not that slow.
- Extra supports and cues helped to answer the question.

**Least Liked Features**

- None

## Analysis

This interactor met or exceeded all of its targets. The most important indicators – the clarity and second prompt preference metrics seem to indicate that the additional prompting provided by the interactor was quite effective in clarifying the initial question. On the whole, this interactor proved to be quite usable.

## Future Changes

**Change 1**
What: Adding voice confirmation dialog after selection is made.
Why: From users commands we come to know that they like voice confirmation dialog after the selection. This change will make the interactor more user friendly.

Since, this interactor 07 is almost same as interactor 08 (context- sensitive help) and also interactor 07's ratings were right at the target value, We, group member decided to drop it for our second level testing. Instead we are planning to add a new interactor that has new features and the name of the interactor to be determined later.

# *Interactor 08 (Context-Sensitive Help)*

Confirm an answer to a "textual" question using a proximity threshold in the n-best list. After user's response, get the list of words that match that response the most, and prompt the user again with that list of words: "Did you say item1, item2 or item3?"

e.g. “What city would you like to visit?” User says “Athens”. Computer response: “More than one city matches your choice. Please indicate which of the following cities you meant by saying, Oak A, after the correct city is read.”

User says yes after hearing Athens Greece. (The city user wanted to visit was Long View)

First Phase:
This analysis is based on 25 valid tests of interactor 6, for which all 25 returned preference metrics.

Second Phase:
This analysis is based on 28 valid tests of interactor 6, for which all 28 returned preference metrics.

## Preference Metrics

For this interactor, most preferences should be rated “good” (4) or “excellent” (5). We would like to minimize the number of metric scores of “average” (3) or below. The preference metrics used were:
Here is a summary of the measurements:

| **Phase 1** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **#1** | **#2** | **#3** | **#4** | **#5** |
| Clarity | 4.63 | 4.5 | 0 | 1 | 1 | 5 | 20 |
| Effectiveness | 4.52 | 4.5 | 1 | 0 | 2 | 5 | 19 |
| Ease of Use | 4.52 | 4.5 | 0 | 0 | 3 | 7 | 17 |
| **Phase 2** | | | | | | | |
| Clarity | 4.29 | 4.5 | 0 | 1 | 6 | 5 | 16 |
| Effectiveness | 4.29 | 4.5 | 1 | 1 | 4 | 5 | 17 |
| Ease of Use | 4.18 | 4.5 | 0 | 2 | 4 | 9 | 13 |

## Performance Metrics

Here is a summary of the measurements:

| **Phase 1** | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |

| | | | | |
|---|---|---|---|---|
| Time per Iteration | 13924 | 12000 | 12675 | 2588 |
| Error Rate | 0.047 | 0.10 | 0.0 | 0.106 |
| No Match Rate | 0.058 | 0.05 | 0.0 | 0.110 |
| No Response Rate | 0.039 | 0.05 | 0.0 | 0.065 |
| **Phase 2** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 11663 | 12000 | 11145.5 | 2004.84 |
| Error Rate | 0.05 | 0.10 | 0.0 | 0.14 |
| No Match Rate | 0.03 | 0.05 | 0.0 | 0.07 |
| No Response Rate | 0.07 | 0.05 | 0.0 | 0.13 |

# Comments (Phase 1)

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- I appreciated the additional cues to help the user
- I preferred the interactors where I could speak my choice

**Least Liked Features**

These comments were given as the features of the interactors that the user liked least. Comments are paraphrased to protect the user's identity:

- I liked speaking the destination names
- Some people prefer speaking as soon as they know what they need to respond, even if the computer is still speaking – they should be able to do this.

**Desired Features**

These comments were given as the features the user would like to see in the interactors. The comments are paraphrased to protect the user's identity:

- I would like audio confirmation of my choices.

**General Comments**

These are general comments that the users gave. The comments are paraphrased to protect the user's identity:

- The instructions were somewhat ambiguous – at first I thought I should choose my favorite city, instead of the city I was supposed to choose.

## Comments (Phase 2)

- Why not stat with the city/county?
- A little vague when instucted to "select". Pefe the pompt "speak".
- I didnt do well at eading the second diection.
- I was told I would be asked to say Help when pompted but I was not pompted.
- in the fist test, i thought the compute would give me a list of choices, so i didnt immediately say what i was supposed to. then the compute said speak the name, and i knew that i was just supposed to say the city without having the compute give me a list.

## Analysis

First Phase:

So, all three metrics met their targets, these ratings don't indicate any change by themselves.

Time per Iteration was off by more than two seconds, however, the median is very close to the target, indicating that the average is overestimating the time taken, due to a few users who had difficulty.

The Error Rate is right on target, and the median suggests that it overestimates the error rate – in fact, only 3 users experienced any errors whatsoever.

The No Match Rate was a little high, and, since this interactor expects the user to speak more varied responses, this may be due to similar-sounding words in the grammar, or difficulty understanding particular users. This is especially true, since the low median indicated that most users experienced no problems. Since this did not correlate particularly with the error rate, it appears that the interactor may have overcome any shortcomings in the recognition engine. Small changes in the grammar may be indicated by this result.

The No Response Rate was quite low, which suggests that users were not confused about what to say and generally had enough time to respond.

In all important respects, this interactor performed very close to how we expected. It appears that even where problems occurred, they applied to a very small minority of users. The main improvement needed is in decreasing the No Match Rate.

The somewhat high No Match Rate, along with some of the textual comments suggest that users may not have understood the written instructions on what to say and that the grammar used for the test contain some choices that are too close together phonetically.

Second Phase:

In general the performance test results were either very close to or bettered our goals. The most notable improvement was in Time per Iteration. The most notable worsening was in the No Response Rate which was the only measure that exceeded our goal.

The preference results were all a little worse than the first phase. Changes to the test application and the webpage that we used to gather the preference data may have contributed to the changes.

Clarity:

There was a difference of 7.3 percent (4.63 first phase, 4.29 second phase) between the first phase and the second phase. The principle change effecting clarity between the phases was to provide verbal confirmation of a choice and the addition of contextual instructions . This possibly correlates with a slight worsening of No Response Rate from 0.039 to 0.07.

Effectiveness:

The difference in effectiveness worsened (4.52 first phase, 4.29 second phase) between the first phase and the second phase.

Ease of Use:

Ease of Use decreased (4.52 first phase, 4.18 second phase) from the first phase even though the time per iteration improved slightly. The comments from the second phase indicated some confusion with understanding how to respond with the first prompt.

Error Rate:

The error rate remained essentially the same (0.047 first phase to 0.05 second phase) from the first phase. Well under our goal of 0.10.

No Match Rate:

The No Match Rate is the (Number of No-Matches)/(Number of listens). It improved from the first phase 0.058 to 0.03.

No Response Rate:

This measure (No Response)/(Number of listens) worsened from 0.039 in the first phase to 0.07 in the second phase indicating the interactor had more difficulty picking up an utterance to match or not match.

# Changes (Phase 1 to Phase 2)

The changes are given in priority order:

**Change 1**

**What:**

Provide verbal verification or no response detected message.

**Why:**

In the first phase there were some comments that verbal response would be preferred to seeing the response in a text box.

**How to measure:**

Should be a decrease in the number of incorrect responses if test subjects understand what their response was.

**Did it work?**

The total number of incorrect remained the same however the number of incorrect responses per test subject went down very slightly. Also in the first phase the incorrect responses were spread out between 2 people where in the second phase 1 person had both incorrect responses. This indicates that fewer people had problems.

## Future Changes

**Change 1**

Change the prompts or instructions to make it more clear how to respond.

**Change 2**

Add more time to the listen to improve the No Match Rate.

## *Interactor 09 (Whisper Prompt)*

The Whisper Prompt Interactor helps the user in case of incorrect response or no response, by playing an example of what is expected. The original idea was to play this prompt at a lower volume, which turned out to be impossible at this time due to IE browser limitations. Therefore, the better name for this interactor is Example Prompt, although we kept the original naming.

In phase 1, the interactor paused after the initial question prompt was played, giving the user some time to respond. If the user responded incorrectly or did not respond at all, the interactor played the example prompt.

In phase 2, in response to the sponsor's suggestion, we created a new interactor 09.A and also kept the original Whisper Prompt 09 unchanged, which was now called 09.B.

Interactor 09.A provides the user with the example prompt right after the initial question prompt is played, without the pause.

This analysis is based on 27 valid tests of interactor 09 in phase 1, and 30 valid tests of the interactor 09 in phase 2, and 30 valid tests of the interactor 09.A in phase 2.

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below.

| **Phase 1 (Whisper Prompt 09)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Mean Score** | **Target** | **# 1** | **# 2** | **# 3** | **# 4** | **# 5** |
| Clarity | 4.67 | 4.5 | 0 | 0 | 2 | 5 | 20 |
| Effectiveness | 4.52 | 4.5 | 0 | 0 | 4 | 5 | 18 |
| Ease of Use | 4.59 | 4.5 | 0 | 0 | 3 | 5 | 19 |
| **Phase 2 (Whisper Prompt 09.B)** | | | | | | | |
| Clarity | 4.13 | 4.5 | 1 | 3 | 2 | 9 | 15 |
| Effectiveness | 4.23 | 4.5 | 1 | 1 | 3 | 10 | 15 |
| Ease of Use | 4.13 | 4.5 | 0 | 2 | 6 | 8 | 14 |
| **Phase 2 (Whisper Prompt 09.A)** | | | | | | | |
| Clarity | 4.34 | 4.5 | 0 | 2 | 5 | 3 | 19 |
| Effectiveness | 4.07 | 4.5 | 0 | 2 | 7 | 7 | 13 |
| Ease of Use | 4.34 | 4.5 | 0 | 1 | 5 | 4 | 18 |

In phase 1, all the metrics for Whisper Prompt 09 met or exceeded their target. No changes need to be made.

In phase 2, all of the metrics for Whisper Prompt 09.A (with no pause) missed their target by very little. The effectiveness metric was the furthest from its target and got the fewest number of excellent scores. Considering how close the metrics were to their targets, we can say the interactor did pretty well. If there are any improvements to be made, we should put this interactor into a real life scenario.

In phase 2, all of the metrics for Whisper Prompt 09.B (original 09) were under their target by very little. There were very few people who were dissatisfied with the interactor. The majority of users gave it excellent scores.

## Performance Metrics

| **Phase 1 (Whisper Prompt 09)** | | | | |
|---|---|---|---|---|
| **Metric** | **Mean Score** | Target | **Median** | **Standard Deviation** |
| Time per Iteration | 11234 | 11000 | 11101 | 3063 |
| Error Rate | 0.062 | 0.10 | 0.0 | 0.132 |
| No Match Rate | 0.04 | 0.05 | 0.0 | 0.11 |

| | | | | |
|---|---|---|---|---|
| No Response Rate | 0 | 0.05 | 0.0 | 0.0 |
| **Phase 2 (Whisper Prompt 09.A)** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 9165 | 10000 | 9673 | 1593.21 |
| Error Rate | 0.23 | 0.10 | 0.0 | 0.25 |
| No Match Rate | 0.21 | 0.05 | 0.0 | 0.24 |
| No Response Rate | 0.0 | 0.05 | 0.0 | 0.0 |
| **Phase 2 (Whisper Prompt 09.B)** | | | | |
| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
| Time per Iteration | 8040 | 10000 | 7279 | 2959.28 |
| Error Rate | 0.06 | 0.10 | 0.0 | 0.14 |
| No Match Rate | 0.05 | 0.05 | 0.0 | 0.13 |
| No Response Rate | 0.0 | 0.05 | 0.0 | 0.0 |

In phase 1, Time per Iteration was off by almost two seconds, and the median suggests that the average is slightly higher than the amount of time taken for most users – some change may be indicated.

In phase 2, Time per Iteration was almost 2 seconds under its target, which shows a big improvement from phase 1.

In phase 2 for Interactor 09.A, time per iteration was less than a second under its target, which shows that having no pause between the question and example prompts did not have much influence on test duration.

In phase 1, the Error Rate is as expected. We were expecting the data to indicate a low error rate, and the average score indicates such.

In phase 2, the Error Rate is also low.

In phase 2 for Interactor 09.A, the Error rate was twice the desired amount. The median of 0 suggests that the average overestimates this rate.

In phase 1, the No Match Rate was very close to the target, although it still met our expectations.

In phase 2, the No Match Rate was right on target, which shows consistency between the data from two usability tests.

In phase 2 for Interactor 09.A, The No Match Rate was twice the desired amount. The median of 0 suggests that the average overestimates this rate.

In phase 1, the No Response Rate is better than expected. We expected a no response rate of less than 5%, and the data indicates a 0%.

In phase 2, Response rate was also zero, which is below its target.

In phase 2 for Interactor 09.A, Response rate was zero, which is below its target.

Overall, the new interactor 09.A (example prompt with no pause) did not show any improvement over the original interactor 09 with the pause, according to the metrics.

# Comments (Phase 1)

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- Comment

**Least Liked Features**

These comments were given as the features of the interactors that the user liked least. Comments are paraphrased to protect the user's identity:

- Comment

**Desired Features**

These comments were given as the features the user would like to see in the interactors. The comments are paraphrased to protect the user's identity:

- Comment

**General Comments**

These are general comments that the users gave. The comments are paraphrased to protect the user's identity:

- Comment

# Comments (Phase 2)

The following comments addressed this interactor:

- Why did the computer say to specify lamb and soda when I was supposed to want chicken and beer?
- Why do I order lamb and soda when I feel like having chicken and beer? The computer didn't understand the fist time I said lamb

- What is "testll"? I didn't know whether to wait for the drink options or just to say "soda" after the prompt. The directions didn't indicate this.
- It should accept input as soon as the question ends. This would be more intuitive
- Not enough time to say o.k. other word before they moved on
- more difficult
- Again the computer was having difficulty picking up my voice.
- Did not register that I said SODA at first until after it read the list

## Comments (Phase 2 Interactor 09.A)

The following comments addressed this interactor:

- Why did the computer say to specify lamb and soda when I was supposed to want chicken and beer?
- It should accept input as soon as the question ends. This would be more intuitive
- Confusing to tell me I was thinking of having chicken and beer but to choose lamb and soda.
- The opening explanation said I wanted chicken and beer. The directions said to order lamb and soda. I followed the directions, but was I supposed to change my mind? I wasn't given that option.
- Why do I order lamb and soda when I feel like having chicken and beer? The computer didn't understand the fist time I said lamb

## Analysis (Phase 1)

Whisper Prompt interactor 09 performed exceptionally according to all the metrics besides the Time per Iteration. Most of the complaints from the users were about confusing directions. Some of the comments suggest that some users thought that this interactor offered selection from the list.

## Analysis (Phase 2)

In phase 2, the Time per Iteration rate changed for the better and now was under its target, which showed a great improvement. The reason for it might be the addition of verbal confirmation, since now the users knew when they were done with each test case.

In phase 2, Whisper Prompt 09.A performed fairly well, other than the Error and the No Match Rates. Most of the comments had to do with confusing direction, rather than with the interactor itself. From talking to some of the users, it seemed like having no pause between the end of the question and the whisper prompt was quite confusing. Some people thought this was another list interactor and tried to say OK after items in the whisper prompt. Some users tried to say the response during the whisper prompt and

computer refused to understand them. Many did not quite know what to do, and waited until the end of the whisper prompt to supply their response. We suspect that this was the reason for such high Error and No Match Rates. Having the whisper prompt play at a lower volume or at a softer tone might have helped user to understand things better, but unfortunately, IE does not support the desired features.

## Changes

The only change made to interactor 09 was to adding of the verbal confirmation, which lowered the Time per Iteration rate by almost 3 seconds.

In response to the sponsor's request we created a new interactor - Whisper Prompt 09.A, which took away the pause between the initial `question` prompt and the following example prompt from Interactor 09. The results from testing this new interactor showed that having no pause did not make any improvement over the original Whisper Prompt interactor and might actually be more difficult and confusing. The only change needed for interactor 09.A, which might make it better than Whisper Prompt 09 is to lower volume of the whisper prompt, which we were unable to implement at current time.

# *Interactor 10 (Confirmation and Correction Dialog)*

Confirm the responses given in an html form. Read back each response in the html form and allow the user to say 'no' to correct a response – when the user says no, the voice interactor corresponding to that field is activated to get a new response. When the correction is finished, continue confirming with the next html form element.

This interactor was not implemented for either phase 1 or phase 2.

# *Interactor 11 (Timeout Adjuster)*

This analysis is based on 26 valid tests of interactor 11, for which 25 returned preference metrics.

This interactor was implemented for phase 1 but not for phase 2. It was decided that it performed well in phases 1 and didn't provide enough benefit to include in phase 2. This analysis is from phase 1.

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below. The preference metrics used were:

Clarity – How well did you understand what the computer wanted you to do?
Effectiveness – How well were you able to accomplish your task?
Ease Of Use – How easy was this program to use?
Here is a summary of the measurements:

| Metric | Mean Score | Target | # above 4 | # below 4 |
|---|---|---|---|---|
| Clarity | 4.4 | 4.5 | 20 | 5 |
| Effectiveness | 4.54 | 4.5 | 20 | 5 |
| Ease of Use | 4.54 | 4.5 | 21 | 4 |

All the metrics made their target value.

# Performance Metrics

For this interactor, we used the following performance metrics:

Here is a summary of the measurements:

| Metric | Mean Score | Target | Median | Standard Deviation |
|---|---|---|---|---|
| Time per Iteration | 10750.56 | 11000 | 11201 | 3013 |
| Error Rate | 0.124 | 0.10 | 0.0 | 0.193 |
| No Match Rate | 0.027 | 0.05 | 0.0 | 0.092 |

The mean Time per Iteration was less than the target time, suggesting that many of the tests finished within the expected time.

The Error Rate is is slightly higher than expected. A possible improvement would be to add a verbal confirmation unit to our test.

The No Match Rate was well under the target rate, indicating that our responses were usually given in the expected format.

# Comments

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- Comment

**Least Liked Features**

These comments were given as the features of the interactors that the user liked least. Comments are paraphrased to protect the user's identity:

- Comment

**Desired Features**

These comments were given as the features the user would like to see in the interactors. The comments are paraphrased to protect the user's identity:

- Comment

**General Comments**

These are general comments that the users gave. The comments are paraphrased to protect the user's identity:

- Comment

## Analysis

Other than the error rate, this interactor performed very close to how we expected. Other tests will determine if the system was set correctly, but this error could be mitigated by confirming the given response with the user.

## Changes

The changes are given in priority order:

**Change 1**

**What:** Confirm with the user's response, by repeating the response, and asking for confirmation from the user. The user could say "yes" or "ok," or another affirmative response.

**Why:** Error rate slightly lower than expected. Need to verify that the system understood what the tester was saying.

**How to Measure:** Should cause a drop in the error rate if successful.

# *Interactor 12 (Multiple Related Fields)*

This analysis is based on 26 valid tests of interactor 12, for which 25 returned preference metrics.

This interactor was not tested in phase 2. This analysis is from phase 1 only.

## Preference Metrics

For this interactor, most preferences should be rated "good" (4) or "excellent" (5). We would like to minimize the number of metric scores of "average" (3) or below. The preference metrics used were:
Clarity – How well did you understand what the computer wanted you to do?
Effectiveness – How well were you able to accomplish your task?
Ease Of Use – How easy was this program to use?
Here is a summary of the measurements:

| Metric | Mean Score | Target | # above 4 | # below 4 |
|---|---|---|---|---|
| Clarity | 4.35 | 4.5 | 16 | 5 |
| Effectiveness | 4.04 | 4.5 | 15 | 7 |
| Ease of Use | 4.08 | 4.5 | 16 | 7 |

All of the metrics missed their targets, although not by much. The high number of excellent marks and the median of 5 for all the metrics suggest that we were closer to target value than the mean scores showed. If we were to change this interactor, we would need to improve the way computer gives instructions, and possibly give a better scenario of use.

## Performance Metrics

For this interactor, we used the following performance metrics:

Time per Iteration – The amount of time (in milliseconds) it took to run the interactor to completion once. From pilot tests, we suspect that 11 seconds (24000 milliseconds) is a good target value for this metric.

Error Rate – The percentage of the time that the user selected an incorrect option – taken as a percentage of the total number of choices. We would like to keep the error rate below 10%

No Match Rate – The percentage of the time that the user said something that the interactor did not understand – taken as a percentage of the total number of times the interactor expected a response. We would like to keep the No Match Rate below 5%

No Response Rate - The percentage of the time that the user did not say anything – taken as a percentage of the total number of times the interactor expected response. Note that we measured the rate of "no responses" only to show the correlation between the No Match Rate and no response.

Here is a summary of the measurements:

| **Metric** | **Mean Score** | **Target** | **Median** | **Standard Deviation** |
|---|---|---|---|---|
| Time per Iteration | 25957 | 24000 | 23379 | 5510 |
| Error Rate | 0.24 | 0.10 | 0.0 | 0.23 |
| No Match Rate | 0.06 | 0.05 | 0.0 | 0.09 |
| No Response Rate | 0.24 | 0.05 | 0.2 | 0.09 |

Time per Iteration was off by about 2 seconds, and the median suggests that the average overestimates the amount of time taken for most users – a slight change in the interactor may be indicated.

The Error Rate is nearly two times what we would like – although the median suggests that most users encountered no errors, and that a few users who had problems are skewing the average. Clearly, though, we must make the interactor more accurate for those users who had problems.

The No Match Rate was just above what we would like. The median indicates that most users did not encounter any problems, however, there is a fairly strong correlation between those users who encountered bad response problems, and those who had a non-zero error rate (10 of the 10 non-zero error rates also encountered bad response problems). This indicates that most of the time people did not know what to say or ran out or time. Expanding the vocabulary of the interactor, giving precise directions, and increasing WAIT FOR RESPONSE time may have a significant impact on the overall error rate.

# Comments

The following comments addressed this interactor:

**Best Liked Features:**

These comments were given by the user as the features of the interactors they liked the best. Comments are paraphrased to protect the user's identity:

- There were no comments for this interactor

**Least Liked Features**

These comments were given as the features of the interactors that the user liked least. Comments are paraphrased to protect the user's identity

- The pause between choices was so long that I didn't know if there were more options to come
- The choices were read somewhat slowly
- I did not have enough time to say the date. By the time I looked back at the directions, the computer stopped listening.

**Desired Features**

These comments were given as the features the user would like to see in the interactors. The comments are paraphrased to protect the user's identity:

- I don't want to feel rushed when answering questions. Seeing how this technology can be applied in real life can help with understanding of the timing of the answers.

# Analysis

Most of the people rated this interactor highly. At the same time the error and no response rate was quite high. This is unusual since people tend to get frustrated and rate things low when something does not work correctly. The metrics suggest that most problems occurred when people did not know what to say or ran out of time to say it. The main improvement needed is in diminishing the rate of errors and no response.

## Changes

The changes are given in priority order:

**Change 1**

**What:** Increase listening time to allow the user more time to figure out the date to supply.

**Why:** High Error and No Match Rate correlated with No response rate may indicate that the user did not know what to say or did not have enough time to say it. This was also suggested by some comments.

**How to Measure:** Should cause a drop in the No Match Rate, a drop in the error rate, and a no response rate if successful.

**Change 2**

What: Add verbal confirmation of user choices.

# Limitations of Testing

These usability tests were performed in a short time frame as a student project, with limited access to resources. For this reason and others, there are several limitations that may affect conclusions that can be drawn from these tests:

- We are testing interactors, which are tools used to build an interface, and not a specific interface itself. Certain aspects of a real user interface – such as user-supplied prompts and grammars - could have a significant impact on the overall usability of any real interface. Therefore, the prompts and user-defined grammars used in these tests provide a minimum standard for prompts and user-defined grammars – any implementation of this interface should separately determine the usability of any user-supplied prompts and grammars.
- Testing was scripted – users were told which items to select when exercising each interactor. We saw this as a tradeoff – we wanted to make sure that the limited number of users we tested would sufficiently test each component of each interactor. However, scripted testing means that the usability of certain features (such as supplied context-sensitive help prompts) might not have been fully examined. We decided that this was the lesser of two evils, since many of the affected features (like help prompts) are user-supplied and, as noted, any user of the interactors should separately determine the usability of supplied prompts and grammars.
- Sampling was not truly random, nor was the sample size as large as we would like. We picked subjects on the basis of their willingness to volunteer and meet certain minimum criteria (English speaker, no thick accent, no computer experts). As such, the applicability of the test results to the population at large may be legitimately questioned.
- The interactor usability was determined only in English, using speakers without thick accents. Separate tests would need to determine interactor usability for non-English speakers and non-native English speakers.
- We did not have access to a traditional usability testing environment, so our observations on subjects were limited to what could be determined from being in the room during testing.
- Tests used a multimodal interface, to try to isolate the voice interface from the testing mechanism for user evaluation – tests using a voice-only test harness might produce different results.
- Timing issues limited the number and scope of changes that could be performed between the first and second rounds of usability tests.

# Lessons Learned

We have split lessons learned into lessons about SALT and the IE implementation of SALT, lessons about human behavior revealed by the test, and administrative lsssons about running a usability test on a large University campus.

## *Technical Lessons Learned*

- Background noise tends to come through as bad responses, especially when a prompt is being read and a listener is active at the same time – better handling of the no match, or improvement of the default "bargein" behavior is indicated (our solution was to use headsets). To the user, this may seem like 'running out of time too soon'
- Grammars that are looking for a particular response (like the 'yes' grammar) should include ANY word the user likely say – having a grammar of only 'yes' responses results in a lot of false positives when the user says 'no'. Adding 'no' to the grammar then parsing the answer increases the probability of understanding what the user said. There is on caveat where using words that are phonetically similar such as 'ya', 'yeah', 'neah' seemed to confuse the interactor and increase no match responses.
- "Multiple" mode LISTEN objects in IE have a perceptible threshold between answers – user responses are sometimes lost if they are provided too quickly. Most users' speech matched the prompt speed, but some provided their responses too quickly.
- Monosyllabic responses result in a higher percentage of 'no match' errors
- PromptQueue is not implemented in IE – neither are DTMF tags, or n-best lists
- The SSML tags appear not to be implemented in IE – particularly, examples of using SSML Emphasis tags as published on Microsoft's website have no effect in IE.
- We were not able to figure out how to play sounds using the Prompt object in IE – examples from the SALT 1.0 Specification did not work.
- It is necessary to have a good idea of what you are trying to measure before you run a usability test on any piece of software.
- Fine gradations of confidence level do not appear to have a significant effect (.85 vs. .8). Most successful recognitions have a confidence level between .8 and .9
- Setting the "reject" parameter in a Listen object to a low level (below 0.6) does not result in actually lowering the threshold for recognition – raising the "reject" level, though, does have the desired effect.
- The voice recognition engine in the SDK will return a -7 error if it was unready to begin recognition or was unable to load all page components before a listen executed.
- In IE, it takes some time for the SDK object to load, and it is sometimes unloaded if it is unused for a while – this can cause problems with the first test in a series – it is, therefore, a good idea to run a SALT 'warm up' test at the beginning of each test session before executing tests where results are recorded.

## *Human Behavior*

- Users like verbal confirmation of choices – verbal confirmation was a uniformly successful addition to all of the interactors
- Users are about evenly split on how they prefer selecting items from a short, fixed list of options – about half prefer saying the responses, and the other half prefer making selections as the list is read.
- When selecting from a fixed list of options, many users want to say the name of an item, even when told to say something else (like 'yes' or 'ok')
- 'That One' is a particularly bad alternative for yes – this made MOST users try to say the name of the item, rather than the words 'that one'.
- Confirming 'yes' and 'no' responses seems to have limited usefulness – we did not find a significant error rate when testing 'yes' and 'no' by themselves, and the error rate for confirmation was extremely high. The error rate for confirmation remained high regardless of whether we used speech or a multi-modal dialog, and whether we asked them what they had spoken, or re-asked the question. People seem to be quite confused when asked to confirm a yes/no choice – in our estimation, the extra confusion is not worth the extremely slight gain in overall accuracy.
- Users prefer having a substantial gap between a prompt and a list of examples designed to help them make a selection – users seem reluctant to speak until all prompts are finished. Changing the volume of the prompt may have some effect on this.
- Testing similar interactors together at the same time proved to be somewhat problematic for some users – many expected each interactor to behave identically. When testing similar interactors, one should either use a completely different testing scenario to minimize user expectations of parallelism, or give the user explicit instructions of what to expect.
- Users responded better to being asked for optional textual comments throughout a test than to being required to give textual comments at the end of testing. Using the first method, we gathered many more comments, and, in general, comments were more specifically targeted without having to ask more specific questions.
- Contextual instructions worked much better than instructions at the beginning of a test. Users only want to know what they need to know to complete the task they are engaged in – any additional information only serves as a distraction for most users.

## *Administrative Lessons*

- Many Human Subjects review boards are particularly sensitive to information that is published – if you are performing a study that will publish its results, running your project by Human Subjects at least two months in advance of any testing will save you many headaches down the road.
- Some Human Subjects review boards are split about whether publishing on the web requires as much scrutiny as publishing in print.
- Acquiring volunteers at the time of testing proved to be a much more effective method than pre-scheduling them in advance – concentrating on older volunteers

and offering more money as an inducement might make pre-scheduling more effective.

- A testing location central to campus proved to be very important – it is fairly easy to recruit students as tests are going on, if they don't have far to go to get to the test location.