Usability Evaluation of Prompt Clarity for a Language Selection Task

TR 29.3819

August 13, 2004

Barbara Millet
James R. Lewis

IBM Pervasive Computing
Boca Raton, Florida

Abstract

This report describes the method and results of a usability evaluation of a prototype speech recognition IVR application. Twelve participants, 6 of which were considered Expert users, completed two tasks: Select a Language (Task 1) and Purchase a Service Contract (Task 2). Task 2 served as a dummy task, as the primary purpose of the study was to assess the clarity of the language selection prompt ("Select English o seleccione Espanol"). All participants successfully completed Task 1 and 11 of 12 participants successfully completed Task 2. Results showed that user skill level had a marginal significant effect on the time needed to successfully complete Task 1 ($p = 0.075$) and on participant satisfaction ($p = 0.063$) in completing the both tasks. No statistically significant difference was observed between user groups in time to complete Task 2. Overall, few usability problems occurred, none of which were severe, and responses to the ASQ indicated that participants were highly satisfied with the system.

ITIRC Keywords

Voice systems
Speech systems
Interactive Voice Response (IVR) systems
Speech user interface
Prompt clarity
Turn-taking prompts

Contents

Introduction

Interactive Voice Response (IVR) applications are typically comprised of statements and prompts that relay information to the user about the task at hand. Prompts, in specific, are turn-taking cues that should provide the user with information that "cause the user to speak" and that "convey to the user what may be spoken" (Balentine, 1999). To ensure ease of use and overall user satisfaction, it is imperative that system dialogues not be ambiguous.

The objective of this study was to assess the clarity of a language selection prompt planned for use in a speech recognition IVR system. Language selection was the first turn taking prompt in the application ("Select English o seleccione Espanol.") This prompt requires that the user select a language with which to navigate through the system. The goal of this evaluation was to determine how well the prompt met this objective.

Method

Participants

Ten IBM employees (7 men and 3 women) and two contractors (1 man and 1 woman) participated in this study. Five of the 12 participants' native language was English. The participants' ages ranged from 20 to 49 years old. All participants had at least some college education, had identified themselves as "very skilled" computer users with more than 5 years of computer experience. Six participants specified previous experience with speech recognition software and all involved indicated experience in using speech recognition systems by telephone.

Materials and Equipment

A prototype of the application was developed using the IBM Voice Tool Kit for WebSphere® Studio version 5.0, starting with the Call Flow Builder, followed by direct modification of the VoiceXML code. A voice talent was employed to record the audio files. The VoiceXML code and audio files were then placed on a voice server, which in turn allowed connectivity to the system by telephone. Each test session was video recorded to capture voice inputs as well as any non-verbal gestures produced by the user. Additionally, a phone tap was used to record all dialogue generated by the system.

Procedure

All participants were tested, individually, in the Human Factors lab in IBM's Boca Raton facility. Testing occurred on June 22- 24, with each test session taking no more than 15 minutes. At the start of each test session, a background questionnaire (see Appendix A) was provided to all participants. Immediately following, the test user was presented with a task scenario and the test task (see Appendix B).

Based upon the task scenario provided, participants understood the task to be the purchase of a service contract using the IVR system. Note that, this test evaluated both the participant's ability to correctly Select a Language (Task 1) and to Purchase a Service Contract (Task 2). The second task (purchasing a service contract) served as a dummy task masking the actual task of interest since the user needed to select a language before reaching the Main Menu.

While the participant executed the tasks, the experimenter logged the participant's actions. If the participant completed the tasks by providing the correct inputs, resulting in the correct outputs, then he or she completed the task successfully. For Task 1, a correct input was saying "English" or "Espanol" or pressing 1 or 2. Similarly, for Task 2, a correct input for buying a service contract would be to say "Make a Purchase" or pressing 2 at the Main Menu and then saying "Service Contract" or pressing 4. If the user progressed to the global introduction of the system (by selecting English) or if the user was transferred directly to a Spanish speaking agent (by selecting Spanish), then user produced the correct output for Task 1[1]. To produce the correct output for Task 2, the user must be transferred to a Service Contract Specialist. The navigation strategy to produce these outputs is depicted in Figure 1.

[1] If the user provided "Spanish" as the input for task 1, then this would have resulted in a correct output for both Task 1 and Task 2 because the system was not designed for self service in Spanish.

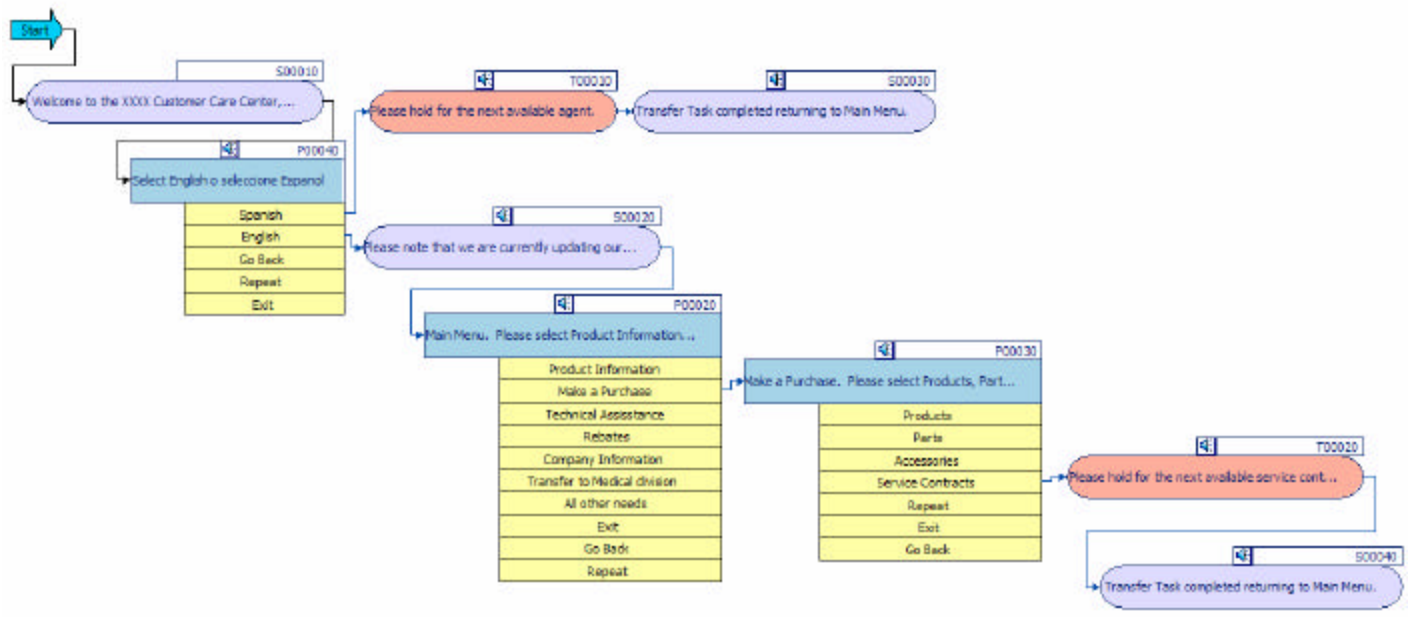

Figure 1. Diagram of the navigation strategy required to produce the correct outputs for the test tasks.

Upon task completion, the test users completed the After Scenario Questionnaire (ASQ)[2]. Additionally, to inquire further about the user's experience interacting with the system the evaluator asked each participant an additional five questions (see Appendix C).

Data Analysis

The measures of usability in this study were:

1. Time to complete the tasks (measured for Task 1 as the time from when the language selection prompt began until the participant provided a correct input resulting in the correct output and for Task 2 as the time from when the Main Menu prompt began until the correct output was produced).

2. ASQ scores for Task 2[3].

3. Successful Task Completion Rates (ranging from 0, for unsuccessful completion of a task), to 1 (for a successful completion).

Time and preference data were analyzed using Microsoft® Excel (ANOVA with $\alpha = 0.05$ for significance, $\alpha = 0.10$ for marginal significance). Success and error rates were analyzed with 95% binomial confidence interval (Lewis, 1996).

[2] Since the test user was not informed that the event of interest was the successful completion of the Language Selection task prior to completing the ASQ, the ASQ is assumed to reflect the users experience in completing the task scenario.

[3] The items in the ASQ were scored following the 7-point scale scoring method, in which low scores correspond to more favorable ratings. The ASQ score for a participant's satisfaction with the system was obtained by averaging the scores from the three items. If a participant marked N/A for an item, the remaining items were averaged to obtain the ASQ score (Lewis, 1995).

Results

Data was collected for this study with 12 participants (N = 12). Of these participants, 6 were identified as Expert users, given that they had indicated previous experience using speech recognition software on a computer, and 6 were categorized as Novice users (as they had no previous experience with speech recognition software on a computer). The overall mean times to successfully complete Task 1 and Task 2 were 10.9 and 47.5 seconds with standard deviations of 2.8 and 16.3, respectively (data provided in Appendix D).

For Task 1, only one participant generated a "Help" prompt to complete the task. The participant was commenting, out loud, about his uncertainty in providing a valid input and received a support statement as a result of a no-match input. Nonetheless, all participants completed Task 1 successfully (0% error). A 95% binomial confidence interval for this error percentage ranged from 0.0 to 26.5%.

For Task 2, five of 12 participants initially requested a repeat of menu options prior to selecting an option from the Main Menu. Therefore, it can be expected (with a 95% confidence) that a minimum rate for providing repeat as input for this prompt would be 15.2% and as high as 72.3%. Similarly, one of 12 participants generated a help prompt from the Main Menu prior to providing the correct input for the Purchase a Service Contract task. This indicates an observed rate for selecting "Help" of 8.3%, with a 95% confidence interval from 0.2 to 38.5%.

In completing the second task, eight of 12 participants reached the target final path (i.e. that of being transferred to a Service Contract Specialist) by selecting "Make a Purchase" from the Main Menu. This observation yields an observed success rate of 66.7%, which with 95% confidence can be as low as 34.9% or as high as 90.1%. Three participants selected "All Other Needs" with an observed rate of 25.0%, with a 95% confidence interval from 5.5 to 57.2%. Because the agent reached by requesting "All Other Needs" would transfer a caller to the appropriate agent, this is also a successful, albeit less efficient, task completion. Using this relaxed criterion, 11 of 12 participants successfully completed Task 2 (with a 95% binomial confidence interval ranging from 61.5 to 99.8%). Unexpectedly, one participant provided "Rebates" as the input for the Main Menu prompt. Tables 1-5, presented below, provide the data (i.e. means, standard deviation, success rate and 95% confidence intervals) collected for Task 1 and Task 2.

Table 1.
Summary of Data

				95% Binomial CI	
Task #	Task Description	Mean Completion Time (secs)	Success Rate	Lower Limit	Upper Limit
1	Language Selection	10.9	100.0%	73.5%	100.0%
2	Purchase a Service Contract (by Selecting "Make a Purchase" from the Main Menu.)	47.5	66.7%	34.9%	90.1%
	Purchase a Service Contract (by Selecting "Make a Purchase" or "All Other Needs" from the Main Menu.)	44.1	91.7%	61.5%	99.8%

Table 2.
Task 1: Language Selection

	Time to Successfully Complete Task (secs)						Mean	STD DEV	95% CI
Expert	9	10	11	9	9	9	9.5	0.8	±0.7
Novice	17	16	10	10	12	9	12.3	3.4	±2.7
All							10.9	2.8	±1.6

Table 3.
Task 2: Purchase a Service Contract*

	Time to Successfully Complete Task (secs)						Mean	STD DEV	95% CI
Expert		53	68		44		55.0	12.1	±9.7
Novice	34		66	33	58	24	43.0	18.0	±14.4
All							47.5	16.3	±9.2

*Participants that selected "Make a Purchase" from the Main Menu.

Table 4.
Task 2: Purchase a Service Contract**

	Time to Successfully Complete Task (secs)						Mean	STD DEV	95% CI
Expert	40	53	68	23	44		45.6	16.6	±13.3
Novice	34	42	66	33	58	24	42.8	16.1	±12.9
All							44.1	15.6	±8.8

**Participants that selected "Make a Purchase" or "All Other Needs" from the Main Menu.

Table 5.
ASQ Scores for Task Scenario

	Score						Mean	STD DEV	95% CI
Expert	2.0	1.3	1.3	2.0	2.3	2.0	1.8	0.4	±0.3
Novice	1.0	1.0	1.3	2.0	1.3	1.5	1.4	0.4	±0.3
All							1.6	0.4	±0.3

There were no statistically significant differences (with $\alpha = 0.05$) between user groups on the time to successfully complete Task 1 ($F(1,10) = 3.96$, $p = 0.075$) or Task 2 ($F(1,9) = 0.08$, $p = 0.786$). Similarly, there was no statistically significant difference between groups in participant satisfaction ($F(1,10) = 4.39$, $p = 0.063$). However, marginal significance (with $\alpha = 0.10$) was detected between user groups on the time to complete Task 1 and in participant satisfaction when using the system.

Discussion and Recommendations

The purpose of this investigation was to evaluate the clarity of a language selection prompt to be used in a speech recognition IVR system. In doing so, it was discovered that few usability problems occurred, none of which were severe. Additionally, responses to the ASQ indicated that participants were highly satisfied with the system.

The key findings and recommendations from this study are as follows:

- For the Language Selection task, all participants were able to complete the task, but 5 participants displayed facial expressions suggesting confusion with the prompt wording. Additionally, one participant generated a no-match help prompt in trying to complete this task.

  Recommendation: *Change the first word of the prompt from "Select" to "Say."*

- Eleven of 12 participants successfully complete Task 2. Eight of these participants selected "Make a Purchase" from the Main Menu, while three participants chose "All Other Needs".

- Five of 12 participants repeated the list of Main Menu options prior to making a selection. This is probably due to the participants' unfamiliarity with the application, is likely to only occur with initial system use, and therefore has no system design implications.

- At the end of the experiment, most participants suggested that there were too many options in the Main Menu. However, had the correct option been more intuitive then the participants may not have needed to hear all the options (i.e. and may have barged-in) after hearing the correct selection.

- Based upon the statistical analyses, user skill level had a marginal significant effect on the time needed to successfully complete Task 1 and on participant satisfaction in completing the task scenario. Furthermore, there was no statistically significant difference between user groups in time to complete Task 2. On average, the Expert group was somewhat faster in completing Task 1 and slightly more critical of the system. Compared to the Expert group, the Novice users were at an initial disadvantage in using the system because they lacked experience with speech recognition software. As the Novice user progressed to Task 2 and gained more familiarity with the system the difference between groups, in time to complete the task, was no longer detectable. Similarly, the observed marginal difference in user satisfaction (with Novice users indicating a greater satisfaction with the system) is a likely consequence of different user exposure to speech recognition systems.

- Although there were many bilingual participants, all participants' selected English as the language in which to navigate within the system.

- Two participants provided variations of the required inputs in the completion of Task 2. Participant 8 provided "Purchase" instead of "Make a Purchase," while Participant 11 provided "Service" instead of "Service Contract."

  Recommendation: *Equip the application with flexible grammars to accommodate various inputs for each option. For example, include inputs that begin with "Purchase" to select the making a purchase option from the main menu and "Service" or "Contracts" for selecting the purchase of a service contract.*

References

Balentine, Bruce and David P. Morgan (2001). *How to Build a Speech Recognition Application* (2nd edition). San Ramon, California: EIG Press.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*, 57-78.

Lewis, J. R. (1996). Binomial confidence intervals for small sample usability studies. In G. Salvendy and A. Ozok (Eds.), *Advances in Applied Ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics -- ICAE '96* (pp. 732-737). Istanbul, Turkey: USA Publishing.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*, 463- 488.

Appendix A. Background Questionnaire

Participant Name:_____
Participant ID: XXX _____
Group: Internal/ External__
System: XXXXX IVR_____

Background Questionnaire

1. I am:
   a. Male b. Female

2. My age is:
   a. Less than 20 years old b. 20-29 years old
   c. 30-39 years old d. 40-49 years old
   e. 50-59 years old f. over 59 years old

3. My education level is:
   a. High school graduate b. Vocational/Technical graduate*
   c. Some college d. Bachelors degree*
   e. Masters degree* f. Doctoral degree*
   g. Other _____

   **Specialty Area* _____

4. My native language is: _____

5. I have used computers for:
   a. more than 5 years b. 1-5 years
   c. less than 1 year d. I have never used a computer.

6. I have used computers at:
   a. home and work b. work only
   c. home only d. I have never used computers

   What type of computer(s) do you use? _____

7. When I use a computer, I typically use it for: (*circle all that apply*)
   a. word processing b. spreadsheets
   c. graphics/paint/draw d. video games
   e. other (*please describe*) _____

8. My typing speed is about _____ words per minute.
   I do NOT type ___.

9. I have used speech recognition software on a computer?
   a. Yes b. No

   If yes, used it for what? _____
   If yes, what kind? _____
   If yes, how long ago? _____

10. I have used speech recognition systems by telephone?
    a. Yes b. No

    If yes, used it for what? _____
    If yes, what kind? _____
    If yes, how long ago? _____

11. When I was a child (*birth -10 years old*) I lived in the following location(s):
    a. _____
    b. _____
    c. _____

12. I enjoy working with computers?
    a. Yes b. No

13. On a scale of one to ten how skilled are you using computers?

Very Skilled					*Average*				*Not Skilled*
10	9	8	7	6	5	4	3	2	1

Appendix B: Task Description and ASQ

Automated Phone System Prototype

Participant Name: _____
ID #: XXX
System: XXXX IVR
Group: Internal/ External

Task Scenario:

You have purchased a product from a Manufacturer and are now interested in buying a service contract.

Task:

Call the Manufacturer's Customer Care Center at XXX-XXX-XXXX and use the automated system to buy a service contract.

After Scenario Questionnaire:

For each of the statements below, circle the rating of your choice.

1. Overall, I am satisfied with the ease of completing this task.

STRONGLY								STRONGLY
AGREE	1	2	3	4	5	6	7	**DISAGREE**

2. Overall, I am satisfied with the amount of time it took to complete this task.

STRONGLY								STRONGLY
AGREE	1	2	3	4	5	6	7	**DISAGREE**

3. Overall, I am satisfied with the support information provided when completing this task.

STRONGLY								STRONGLY
AGREE	1	2	3	4	5	6	7	**DISAGREE**

Appendix C. After Scenario Evaluator Questions

1. How did you feel in using the system? (What are your thoughts on using the system?)
2. What do you remember about the Introduction/Welcome prompts?
3. Can you recall what the first thing you were asked to do was?
4. Do you feel that the prompts provided you with all or most of the information you needed to perform the task?
5. Do you think the support statements were helpful?

Appendix D. Evaluation Data

		Time to Complete Task (secs)	
Participant	ASQ Score	Task 1	Task 2
1	2.0	9	40
2	1.3	10	53
3	1.3	11	68
4	2.0	9	23
5	1.0	17	34
6	1.0	16	42
7	1.3	10	66
8	2.3	9	44
9	2.0	10	33
10	2.0	9	43*
11	1.3	12	58
12	1.5	9	24

* Participant did not complete the task successfully (based on relaxed criterion).

Trademarks

IBM, the IBM logo, and WebSphere are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both.

Microsoft is trademark of Microsoft Corporation in the United States, other countries, or both.

Other company, product or service names may be trademarks or service marks of others.