

Advances in Human Factors/Ergonomics, 19B

# Human-Computer Interaction: Software and Hardware Interfaces

*Proceedings of the Fifth International Conference on Human-Computer  
Interaction, (HCI International '93), Orlando, Florida, August 8-13, 1993  
Volume 2*

Edited by

**Gavriel Salvendy**  
Purdue University, West Lafayette, IN 47907, USA

and

**Michael J. Smith**  
University of Wisconsin, Madison, WI 53706, USA



ELSEVIER

Amsterdam - London - New York - Tokyo 1993

## SPOKEN LANGUAGE INTERACTION: EFFECTS OF VOCABULARY SIZE AND EXPERIENCE ON USER EFFICIENCY AND ACCEPTABILITY

Thomas W. Dillon<sup>a</sup>, A.F. Norcio<sup>b</sup>, & Michael J. DeHaemer<sup>c</sup>

<sup>a</sup>Department of Information & Decision Sciences, Franklin P. Perdue School of Business, Salisbury State University, Salisbury, Maryland 21801

<sup>b</sup>Department of Information Systems, University of Maryland, Baltimore, Maryland 21228

<sup>c</sup>Lattanze Human-Computer Interface Laboratory, Sellinger School of Business & Management, Loyola College, Baltimore, Maryland 21210

### Abstract

The purpose of this study is to determine the effects of vocabulary size and interface experience on the performance and acceptance of the user. Subjects performed a hands busy and eyes busy task while interacting with a speaker-dependent connected-speech recognition system with audio output. The time required to perform the task decreased significantly when the user acquired experience with the interface. A large inclusive vocabulary decreased the number of non-recognized words spoken by the user. The results of an interface-acceptance questionnaire reveal that subjects are more accepting of the spoken language interface as they gain experience.

### 1. INTRODUCTION

Speech recognition systems are now being integrated into a number of applications in which the computer user is involved in a hands busy or eyes busy task. For this reason it is important to measure system performance variables such as time of task completion, number of uncorrected errors, and user satisfaction when implementing a Spoken Language Interface (SLI) [1][2][3].

One of the human factors activities identified as necessary for improved user acceptance and efficiency in speech recognition is the selection of a proper vocabulary and the design of systems feedback and dialogue [4]. In the past, studies that examined the use of vocabulary in natural language or speech recognition applications were performed by a simulation, sometimes called the "Wizard of Oz" technique [5][1]. Though limited, simulation studies assisted in providing a foundation for understanding spoken language interaction with a computer system [5]. Recent improvements in speech recognition technology now permit

studies to be  
permit the

The skill  
inexperien  
rigid intera

The pres  
vocabulary  
performanc  
recognition  
Future stud

### 2. METHO

#### 2.1. Subje

The subje  
fourth year  
nursing co  
nursing tec  
of actual nu  
were two  
approximat

#### 2.2. Hardw

A Compa  
audio outpu  
mounted m

Five nurs  
entering pa  
tapes were  
natural voc  
then coded  
grammar.  
interaction  
the vocabu  
chosen are

Two voca  
by the nurs  
word choic  
minimum c  
ordering, a

studies to be performed that examine vocabulary without the use of simulation techniques and permit the design of an SLI.

The skill of an SLI user varies with experience and practice [6]. For example, the inexperienced user requires a rigid interaction style [8] while the experienced user can find rigid interaction to be long, boring, poorly focused, ineffective and sometimes misleading [9].

The present study attempts to define the relationship between the independent variables of vocabulary size and experience with the interface with the dependent variables of task performance and user acceptance. The task performance measures are time on task and recognition accuracy. This study involves nursing students, or novice subjects in the domain. Future studies will include nurses that qualify for Advanced Practice Nurse, or expert nurses.

## 2. METHOD

### 2.1. Subjects

The subjects selected to participate in this study were volunteer student nurses, defined as fourth year nursing students with limited nursing experience. All subjects completed the same nursing course work requirements, which included classes in anatomy and physiology, nursing technologies, and health assessment. All subjects had approximately the same amount of actual nursing experience. Ages of the subjects ranged from 21 to 24 years of age. There were two male and eleven female subjects. Each subject was compensated for the approximate 2 hours required for the study.

### 2.2. Hardware and Software

A Compaq DeskPro 386s with a Verbex 6000 connected-speech recognition board with audio output acted as the spoken language interaction system. The system included a head-mounted microphone.

Five nurses performed a "Wizard Experiment" (talked through a physical assessment as if entering patient data into a natural-language voice-processing computer system)[10]. Audio tapes were made of the "Wizard Experiment." The tapes were transcribed to acquire the natural vocabulary spoken by the nurses while performing the task. These transcripts were then coded into a spoken language interface using the Verbex Voice System's development grammar. By using a "Wizard Experiment" we feel that we were able to capture the users interaction style, structure, and most of all, vocabulary [11] (See Table 1 for an example of the vocabulary and grammar). As recommended in the literature, the words which were chosen are tailored specifically for the application [12].

Two vocabulary sets were created for the SLI. The first contains all of the words spoken by the nurses. This results in a large vocabulary of 103 words that accommodates alternative word choices. The second vocabulary set contains 74 words or only those spoken by a minimum of three of the five nurses in the "Wizard Experiment," The grammar, word ordering, and feedback remained the same for both vocabulary sets.

Table 1

---

respirations RATE regular  
 respirations are RATE regular  
 respirations are RATE and regular  
 respirations are RATE per minute and regular  
 respirations are RATE per minute rhythm regular

---

#### Sample Vocabulary and Grammar

### 2.3. Experimental Design

The experimental design, a split-plot design, consisted of one between-groups factor (vocabulary) and one repeated-measures factor (experience). The vocabulary size, determined by the "Wizard Experiment," provided the available vocabulary for both levels of the study. The first vocabulary level contained 103 words or all of words spoken by the five professional nurses during the "Wizard Experiment." The second level contained 74 words, or only those words spoken by a minimum of three of the five "Wizard Experiment" participants or 60%. The two levels of the independent variable experience, were determined by each subjects first and fourth physical assessment task, each of which provide an interaction with the SLI.

Task completion time, number of non-recognitions, number of mis-recognitions, and items skipped were the performance variables collected during the study. Task completion time measured from when the subject began to assess the patient until all data was entered into the SLI. Number of non-recognitions (when the subject uttered a word not in the vocabulary or a word in the vocabulary that was not recognized by the system), number of mis-recognitions (when a subject uttered a proper word or phrase from the vocabulary and the speech recognition system recognized the utterance incorrectly), and items skipped were gathered by observation, review of the input file, and analysis of audio tapes recorded during the study.

In addition, subjective satisfaction with the SLI and attitude toward computers was assessed using two separate questionnaires.

### 2.4. Procedure

Subject's sessions were conducted individually. Prior to participating in the study, subjects completed a pre-experimental computer-attitude questionnaire that contained the following bipolar pairs: personal/impersonal, simple/complicated, helpful/hindering, systematic/random, easy/difficult, forgiving/unforgiving, obedient/bossy, cooperative/obstinate, unthreatening/threatening, intelligent/simple-minded, pleasing/disgusting, flexible/inflexible, satisfying/frustrating, calming/anxiety-provoking, and obliging/demanding [13].

Subjec  
 interface  
 were ran  
 and seve  
 first by i  
 to the sp  
 isolated  
 word phi  
 and conn

For the  
 a patient.  
 or reques  
 including  
 For simp  
 sphygmo  
 cardiovas  
 exam, dat  
 Since no  
 beep when  
 a spoken  
 items five  
 the SLI to

To study  
 first and fo  
 scales of  
 acceptabili  
 pleasing/in  
 comfortabl  
 useful/usef  
 interface,  
 administer

## 3. RESULT

### 3.1. Task

An analy  
 experience  
 interface v  
 gained exp  
 effect of v  
 of experien  
 3.36, p <

Subjects were told that the purpose of the study was to test a newly developed computer interface that permits nurses to enter patient data into a computer system by talking. Subjects were randomly assigned to one of the two vocabularies, six to the large inclusive vocabulary and seven to the small significant vocabulary [14]. Then each subject enrolled voice models first by isolated, followed by connected speech. During enrollment, subjects were introduced to the spoken language vocabulary and grammar that would be used during the study. Each isolated word in the vocabulary was trained a minimum of two times. Then all connected-word phrases were also trained. After the training passes were completed for both isolated and connected speech, subjects performed a practice trial of the spoken language interface.

For the hands busy and eyes busy task, subjects performed a cardiovascular examination on a patient. The patient was a volunteer that laid quietly as if unable to respond to commands or requests from the nurse-subject. Twenty-one data items were gathered by the subjects including respirations and blood pressure, and various pulses, impulses, and heart sounds. For simplicity, data items were limited to cardiovascular response/results gathered by sphygmomanometer (blood pressure cuff) and stethoscope. An adaption of a standard cardiovascular examination guide sheet was provided for each subject. While performing the exam, data was entered into the spoken language interface by a head mounted microphone. Since no visual feedback was available to the subject, feedback was provided by an audio beep when data was received and recognized by the system. If the system did not recognize a spoken utterance, subjects were instructed to attempt to get recognition by repeating all items five times. Each subject performed four complete cardiovascular examinations utilizing the SLI to record response/results.

To study user acceptance, a subjective satisfaction questionnaire was completed after the first and fourth examination. The instrument used was a set of twelve bipolar adjective rating scales of seven intervals each and a concluding thirteenth seven-interval scale for overall acceptability [1]. Scale items were: fast/slow, accurate/inaccurate, consistent/inconsistent, pleasing/irritating, dependable/undependable, natural/unnatural, complete/incomplete, comfortable/uncomfortable, friendly/unfriendly, facilitating/distracting, simple/complicated, useful/useless. After subjects completed the enrollment procedure, a practice trail with the interface, and four examinations for the study, a second computer-attitude questionnaire was administered.

### 3. RESULTS

#### 3.1. Task Completion Time

An analysis of variance was run on the task completion times using vocabulary size and experience with the interface as the main effects. The main effect of experience with the interface was significant,  $F(1,11) = 18.78$ ,  $p < 0.0012$ . This suggested that as a subject gained experience with the interface, the time to complete the task decreased. The main effect of vocabulary size was not significant,  $F(1,11) = 1.40$ ,  $p < .2619$ . The interaction of experience and vocabulary was also not significant for task completion time,  $F(1,11) = 3.36$ ,  $p < .0942$ .

### 3.2. Non-Recognitions of the Spoken Vocabulary

The mean number of spoken language phases that were not recognized by the SLI with the large inclusive vocabulary were 5.25. This was substantially lower than the mean number of spoken language phrases not recognized by the smaller significant vocabulary, 9.642. An analysis of variance was run on the number of non-recognitions using vocabulary size and experience with the interface as the main effects. The main effect of vocabulary size was significant,  $F(1,11) = 16.24$ ,  $p < 0.002$ . The performance of subjects using the vocabularies of 103 words and 74 words was significantly different. Those with the large inclusive vocabulary had far fewer non-recognized phrases. The main effect of experience ( $F(1,11) = 1.07$ ,  $p < 0.3235$ ) and the interaction of vocabulary and experience were not significant ( $F(1,11) = 1.96$ ,  $p < 0.1887$ ).

### 3.3. Mis-recognitions and Skips

The number of mis-recognized phrases was counted along with the number of times a subject would skip a response or result on the physical assessment. An analysis of variance displayed no significance for either mis-recognitions or skips. These two items are more an evaluation of the speech recognition system than an analysis of the effects of vocabulary size or level of experience. For this reason they will not be discussed in the conclusions of the paper.

### 3.4. Subjective Ratings

Two subjective ratings were used to assess user satisfaction and acceptance. The user satisfaction questionnaire consisted of an attitude scale of fifteen bipolar adjective pairs. The mean responses to each subjects pre- and post-experimental questionnaire did not differ statistically for vocabulary size ( $F(1,11) = 0.02$ ,  $p < .8974$ ), experience with the interface ( $F(1,11) = 0.09$ ,  $p < .7759$ ), or the interaction of vocabulary and experience ( $F(1,11) = 0.03$ ,  $p < .8765$ ). This may be seen as a positive reaction to the SLI since subjects did not react negatively to the interface. The second questionnaire was administered following the first and fourth physical assessments. The responses of the twelve bipolar rating scales were used to create an acceptability index (AI). The AI was defined as the sum of the scale responses. An analysis of variance for the AI displayed significance for the main effect of experience,  $F(1,11) = 9.16$ ,  $p < 0.0115$ . No effect was found for vocabulary size ( $F(1,11) = 0.40$ ,  $p < .5396$ ) or the interaction of vocabulary and experience ( $F(1,11) = 2.93$ ,  $p < 0.1149$ ). These findings indicate that as users interact with the interface, they find the interface to be more acceptable.

## 4. CONCLUSIONS

The purpose of this study was to examine the effects of vocabulary size and experience on user efficiency and acceptability. The results show, that as a user gains experience with the spoken language interface task completion time decreases and user acceptance increases. In addition, a large inclusive vocabulary reduces the number of non-recognized word utterances, thus improving user efficiency.

5. ACK

This r  
Executi  
recogni

6. REF

1. C

a

a

2. F

V

A

3. S

C

I

4. F

s

I

5. P

a

6. I

n

7. M

c

8. I

E

i

9. S

F

t

10. I

11. I

12.

13.

## 5. ACKNOWLEDGMENTS

This research was partially funded by a grant from the David D. Lattanze Center for Executive Studies in Information Systems, Loyola College in Maryland. The speech recognition system was provided by Verbex Voice Systems, Inc. Edison, New Jersey.

## 6. REFERENCES

1. Casali, S.P., Williges, B.H., & Dryden, R.D. (1990). Effects of recognition accuracy and vocabulary size of a speech recognition system on task performance and user acceptance. *Human Factors*, 32, 183-196.
2. DeHaemer, M.J. Wright, G., & Dillon, T.W. (1992). Performance effectiveness with voice input for beginning spreadsheet users. *Proceedings of AVIOS '92 Voice I/O Applications Conference*, 179-185.
3. Simpson, C.A., McCauley, M.E., Roland, E.F., Ruth, J.C., & Williges, B.H. (1985). Systems design for speech recognition and generation. *Human Factors*, 27, 115-141.
4. Frankish, C., Jones, D., & Hapeshi, K. (1992). Decline in accuracy of automatic speech recognition as a function of time on task: Fatigue or voice drift?. *International Journal of Man-Machine Studies*, 36, 797-816.
5. Fraser, N.M. & Gilbert, G.N. (1991). Simulating speech systems. *Computer Speech and Language*, 5, 81-99.
6. Leggett, J. & Williams, G. (1984). An empirical investigation of voice as an input modality for computer programming. *International Journal of Man-Machine Studies*, 21, 493-520.
7. Morrison, D.L., Green, T.R.G., Shaw, A.C., & Payne, S.J. (1984). Speech controlled text-editing: Effects of input modality and command structure. *International Journal of Man-Machine Studies*, 21, 49-64.
8. Brajnik, G., Guida, G., & Tasso, C. (1990). User modeling in expert man-machine interfaces: A case study in intelligent information retrieval. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-20, 166-185.
9. Rudnicky, A.I. & Sakamoto, M.H. (1989). Transcription conventions and evaluation techniques for spoken language systems research. No. CMU-CS-89-194, Carnegie Mellon University School of Computer Science.
10. Rudnicky, A.I. (1990). The design of spoken language interfaces. No. CMU-CS-90-118, Carnegie Mellon University School of Computer Science.
11. Michaelis, P.R., Chapanis, A., Weeks, G.D. & Kelly, M.J. (1977). *IEEE Transactions on Professional Communication*, PC-20, 214-221.
12. Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34, 527-547.
13. Kelly, M.J. & Chapanis, A. (1977). Limited vocabulary natural language dialogue. *International Journal of Man-Machine Studies*, 9, 479-501.