Int. J. Human–Computer Studies (1997) **47**, 591–602

User performance and acceptance of a speech-input interface in a health assessment task

THOMAS W. DILLON
Management Information Systems, College of Business Administration, Western Kentucky University, Bowling Green, KY 42101-3576, U.S.A. email: thomas.dillon@wku.edu

A. F. NORCIO
Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21228, U.S.A. email: norcio@ifsm.umbc.edu

(Received 19 November 1996 and accepted in revised form 12 June 1997)

Novice and expert nurses performed a hands-busy and eyes-busy task using a continuous speech-input interface. This study examined their performance and acceptance of the interface. Three factors were found to improve the user's performance: expertise in the domain, experience with the interface and the use of a small vocabulary. Experience with the interface corresponded with a higher degree of acceptance. Available vocabulary size and level of expertise did not affect acceptance.

1. Introduction

With a limited spoken vocabulary and a well-structured grammar, speech input is successful for a number of "hands-busy" and/or "eyes-busy" tasks (Martin, 1989). This includes a range of applications such as quality control and inspections, stock control, parcel sorting (Martin, 1976; Visick, Johnson & Long, 1984), baggage handling (Nye, 1982; Jones, Frankish & Hapeshi, 1992), meter reading (Markowitz, 1993) and direct speech input to computers for medical and dental procedures (Martin, 1976; Markowitz, 1993). Speech input is credited for improving time on task for each application.

Immediate data entry, a second speech-input success factor, can reduce the number of input errors caused by memory lapse or transcription error (Jones *et al*., 1992; Noyes, Baber & Frankish, 1992). This can translate into significant financial savings and improved data accuracy.

Most studies encourage the integration of speech-input when the user is performing an "eyes-busy" and/or "hands-busy" task (Martin, 1989). However, most speech-input studies do not involve tasks where the user is unable to interact with a computer monitor and keyboard. For example, DeHaemer, Wright and Dillon (1994) made comparisons between manual and speech input and between novice and expert users when performing a spreadsheet task. Results indicated that neither experts nor novices performed more effectively using speech-input.

2. Review of literature

The improvement of user acceptance and performance ensures a system's effectiveness (Simpson, McCauley, Roland, Ruth & Williges, 1985). This is accomplished by gaining

1071-5819/97/100591 + 12$25.00/0/hc970153

an understanding of the speech-input interface characteristics and properly selecting the task vocabulary, feedback and dialogue (Oviatt & Cohen, 1991). In addition, understanding the performance and acceptance differences for domain expertise and interface experience is vital (Norcio & Stanley, 1989). The following literature review introduces three interface design factors: the user's domain expertise and interface experience, and the interface vocabulary design.

2.1. VOCABULARY DESIGN

The variability of word selection by humans is a fundamental fact of human behavior (Furnas, Landauer, Gomez & Dumais, 1987). A speech interface must be designed with this variability in mind to improve user performance and increase user acceptance. This is not to say that a speech interface should include thousands of commands. It is not technically or economically feasible to build an interface capable of dealing with all words and their variations (Kelly & Chapanis, 1977).

In order to narrow the size of a speech-interface vocabulary, the application and the domain must be considered in the design (Shneiderman, 1992). Even the earliest speech interface studies suggest that the spoken commands must be designed using words from previous spoken interactions. In addition to domain specificity, the vocabulary of the speech interface should reflect the user's expectation of an appropriated model of the dialogue (Baber & Stammers, 1989). Transcription conventions have been developed to capture a domain and task specific vocabulary (Rudnicky, 1990; Rudnicky & Sakamoto, 1989). We have used this process, commonly called Wizard of Oz (WOZ), to capture the vocabulary of a nurse performing a cardiovascular assessment (i.e. physical examination).

2.2. DOMAIN EXPERTISE

Dreyfus and Dreyfus (1980) developed a five step process for acquiring expertise: novice, competence, proficiency, expertise and mastery. Since this study involves a nurse performing a nursing task (e.g. cardiovascular assessment), the definition for novice and expert is taken from an adaptation of the Dreyfus and Dreyfus process found in the nursing literature (Benner, 1982). Redesigned and made more specific to nursing tasks, this five stage process contains: novice, advanced beginner, competent, proficient and expert.

This study observed only the novice (a fourth year nursing student) and expert [a nurse who meets the criteria of Advanced Practice Nurse (APN)] stages. The APN criteria are as follows: (1) the ability to choose a diagnostic and treatment process and (2) to have specialized knowledge and skill in dealing with a human response that may cut across multiple health problems (Calkin, 1984). In practice, an APN has one of the following specialities: Clinical Nurse Specialist (CNS), nurse practitioner, nurse midwife or nurse educator/administrator.

2.3. EXPERIENCE WITH A SPEECH INTERFACE

The skill of a speech-interface user varies with experience and practice (Leggett & Williams, 1984). The needs and abilities of an inexperienced user are quite different from those of an experienced user (Norcio & Stanley, 1989). For example, an inexperienced user needs a request or prompt that provides a model for the user's spoken utterance (Zoltan-Ford, 1991) and a system response or feedback method that informs the user if the utterance is correct or contains errors (Poock, Martisa & Roland, 1983). The inexperienced user also wants a rigid interaction style that provides error-free data entry (Morrison, Green, Shaw & Payne, 1984) and guides the user through the step-by-step interactive process (Schurick, Williges & Maynard, 1985). This rigid interaction style, though less natural, increases the comfort and usability of the speech interface for an inexperienced user. An experienced user does not need this structure and rigidness. The experienced user may find the rigid interface long, boring, poorly focused, ineffective and sometimes misleading (Brajnik, Guida & Tasso, 1990).

3. Purpose of the study

This study has three general research questions.

1. Will an experienced subject perform faster and be more accepting of the speech-input interface than an inexperienced subject? We expected to find a practice effect after a subject interacted with the speech-input interface.
2. Will a subject perform faster and be more accepting of a large vocabulary? A larger vocabulary, providing a user with synonyms and alternate words, should increase the recognition of spoken utterances and decrease the performance time.
3. Will an expert perform faster and be more accepting than a novice when interacting with the interface? An expert, with practice and experience, performs a task faster than a novice.

The remainder of this paper presents the methodology, the results and the discussion. The results provide confidence for the use of a speech-input interface in the nursing environment (Dillon, McDowell, Norcio & DeHaemer, 1994) and will be useful for human-factors engineers and system designers.

4. Methodology

4.1. THE SPEECH INTERFACE

The speech-input interface actually contains two different interface designs: (1) a speech-input interface that contained 101 words in assorted grammar designs based on the Wizard of Oz (WOZ) process (Baber & Stammers, 1989; Rudnicky, 1990; Rudnicky & Sakamoto, 1989) and (2) a speech-input interface that contained only 69 words and a restricted grammar. As can be seen by the samples presented in Tables 1 and 2, there was minimal difference in the complexity of the two grammars.

For the WOZ process, five expert nurses were asked to verbally record assessment items while performing a cardiovascular assessment on a non-responding patient. These recordings were transcribed to capture the vocabulary and grammar (Rudnicky, 1990; Rudnicky & Sakamoto, 1989). All vocabulary words from the five expert nurses were used in the large vocabulary.

To create the small vocabulary the five transcripts were analysed for similarities. If a given vocabulary word was spoken by at least three of the expert nurses, it was retained

TABLE 1
Large vocabulary assessment items

respirations .RESP and regular
respirations are .RESP and regular
respirations are .RESP good volume
respirations are .RESP per minute and regular
respirations are .RESP per minute rhythm regular
pulse is .PULSE and regular
radial pulse is a rate of .PULSE with good volume
pulse is .PULSE and regular good volume
pulse .PULSE per minute rhythm regular
pulse is .PULSE per minute and regular
blood pressure .PRESSURE1 over .PRESSURE2
.PRESSURE1 over .PRESSURE2 .OUTCOME blood pressure
blood pressure is .OUTCOME .PRESSURE1 over .PRESSURE2
carotid pulse .OUTCOME
carotid pulse .PULSE strong and regular
.OUTCOME bruits heard over carotid arteries

TABLE 2
Small vocabulary assessment items

respirations .RESP and regular
respirations are .RESP and regular
respirations are .RESP per minute and regular
pulse is .PULSE and regular
radial pulse is a rate of .PULSE and regular
pulse is .PULSE per minute and regular
blood pressure .PRESSURE1 over .PRESSURE2
.PRESSURE1 over .PRESSURE2 .OUTCOME blood pressure
carotid pulse .OUTCOME

and included with its grammar in the small vocabulary. All other vocabulary words were discarded.

Since both vocabularies were obtained from transcripts created by expert nurses, both contained words that were familiar and frequently used in the profession (Baber & Stammers, 1989; Shneiderman, 1992). Both vocabularies also contained the numbers 1–199. The feeback, an auditory beep, was the same for each of the two interfaces.

The speech-input interface consisted of a *Verbex Series 6000 Speech Commander* card installed in a Gateway 2000 486DX/33 and a *Profile I* microphone with earhook and feedback speaker. The speech-input application was written in *Listen for Windows.*

4.2. SUBJECTS

Fourteen expert subjects and 14 novice subjects participated. The expert subjects were all Clinical Nurse Specialists (CNS) in trauma/critical care. Each had a master of science degree in nursing or was to graduate with a master of science degree in nursing in less than one month. All had more than eight years nursing experience, with a maximum of 21 years. There were 3 males and 11 females. Expert subjects ranged in age from 30 to 42.

The novice subjects were all nursing students in the final semester of a bachelor of science in nursing program. All had some experience gathering assessment data from a student internship, practicum or volunteer service. Novice subjects ranged in age from 22 to 35. There were 2 males and 12 females.

Prior to interacting with the speech-input interface, each subject enrolled voice models (i.e. created voice models by training) first by isolated, followed by connected speech for one of the two vocabularies. During enrollment, the subject was introduced to the spoken vocabulary and grammar of the speech-input interface. Each isolated word and connected word phrase in the vocabulary was trained a minimum of two times. After training was completed, the nurse-subject performed a practice assessment with the speech-input interface.

4.3. EXPERIMENTAL DESIGN

Three independent variables were manipulated in this experiment. They were user *experience* with the speech-input interface, *vocabulary size* and user *expertise* in the task domain.

Experience. After one practice session, an *inexperienced* subject entered patient data with the speech-input interface. Experience with the interface required that the subject interact with the speech-input interface two additional practice times, approximately one hour. After two practice interactions the *experienced* subject entered data a fourth and final time.

Vocabulary size. Two vocabularies of varying sizes were compared. The large vocabulary included alternative words and synonyms (101 words). The small vocabulary had minimal word choices for each item (69 words).

Expertise. Novice or expert nurses entered data into an information system with the speech-input interface.

In summary, this is a split plot design consisting of two between-subjects factors (i.e. vocabulary and expertise) and one repeated measure within-subjects factor (i.e. experience with the interface).

Dependent variables. The performance variable, task completion time, was recorded by the experimenter while the user was interacting with the interface. Task completion time was measured when the nurse-subject began to assess the patient until all 21 data items were entered. Performance time was recorded for the first and final assessments.

An acceptance questionnaire also was completed after the first and fourth assessment. The instrument used was a set of 12 bipolar adjective rating scales of seven intervals each and a concluding 13th seven-interval scale for overall acceptability (Casali, Williges & Dryden, 1990). Scale items for the acceptance questionnaire are listed in Table 3. The response of the 12 scales were used to create an acceptability index (AI) that was defined

TABLE 3
Adjective pairs used in the acceptance questionnaire. (The acceptability index did not include acceptable/unacceptable.)

Fast	Slow
Accurate	Inaccurate
Consistent	Inconsistent
Pleasing	Irritating
Dependable	Undependable
Natural	Unnatural
Complete	Incomplete
Comfortable	Uncomfortable
Friendly	Unfriendly
Facilitating	Distracting
Simple	Complicated
Useful	Useless
Acceptable	Unacceptable

as the sum of the scale responses for the 12 bipolar scales for each subject under each treatment condition.

4.4. THE TASK

The application chosen for the *hands-busy* and *eyes-busy task* was the performance of a cardiovascular assessment on a non-responding patient. The patient laid quietly and did not respond to commands or requests from a nurse-subject gathering and entering 21 assessment items. Assessment items included 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.

Traditionally, a cardiovascular assessment is performed in one of the following two ways.

(1) Two nurses, a data gatherer and a data recorder, perform the assessment. The gatherer locates each assessment item while verbally "calling" each item to the recorder. The recorder writes each assessment item down to be later entered into the computer system.

(2) One nurse, acting as both gatherer and recorder, locates approximately five or more assessment items and records them. Items are entered into the computer system either immediately or at a later time. The process is repeated until all items are gathered and recorded.

When performing the assessment with the speech-input interface, data was gathered and entered at the same time by a nurse wearing a head mounted microphone that rested comfortably on one of the subject's ears. Since no visual feedback was available to the

TABLE 4
The experimental procedure session

Enrollment of voice models/training
Practical trial
Assessment 1
Acceptance questionnaire
Assessment 2
Assessment 3
Assessment 4
Acceptance questionnaire

nurse-subject, feedback was provided by an audio beep. If the speech interface did not recognize a spoken utterance (e.g. did not provide a feedback beep), the nurse-subject was instructed to attempt to seek recognition by repeating the utterance or a similar utterance within the vocabulary a maximum of five times. During a pilot study we found the speech-input interface recognized close to 100% and that non-recognitions occurred most often when the subject attempted to enter unavailable vocabulary words.

Each session was conducted individually in a laboratory setting (Table 4). The nurse-subject was told that the purpose of the study was to test a newly developed computer interface that permitted a nurse to enter patient data into an information system *accurately* by talking. Accuracy of the data was emphasized.

5. Results

Less than one in 21 response results were not recognized by the speech recognition system, giving the interface a task input item accuracy between 95% and 100%. Non-recognitions normally occurred when a subject uttered words not found in the interface vocabulary or acceptable words with an incorrect grammar. An occasional word may have been mis-recognized by the speech recognition system. Since feedback was provided by an auditory beep, subjects were not aware of mis-recognitions.

Results are presented first for performance time (in seconds), and then for acceptance (AI). A summary of descriptive statistics, including mean performance and acceptance, standard deviation and level of significance are presented in Table 5.

5.1. PERFORMANCE

There were three overall performance effects: experts performed faster than novices, small vocabulary users performed faster than those with a large vocabulary and subjects with experience performed faster.

The repeated-measures ANOVA of the split-plot design identified a significant within-subjects main effect for experience (i.e. a change in mean performance time from 265 to 207 s) ($F(1, 24) = 66.29$, $p < 0.0001$). Experienced subjects performed faster than inexperienced subjects. This confirmed the performance aspect of Research question 1.

TABLE 5
Mean performance (Time) and acceptance (AI), standard deviation and level of significance for experience, vocabulary size and expertise

	Experience			Vocab. size			Expertise		
	Inexp.	Exper.	*p*	Small	Large	*p*	Novice	Expert	*p*
Performance time (s)	265	207	0.0001	217	255	0.033	254	219	0.049
Stand. Dev.	56.6	47.2		47.9	64.4		55.7	58.7	
Acceptance (AI)	33.57	28.43	0.0001	31.04	30.96	0.763	27.75	34.25	0.111
Stand. Dev.	9.7	6.8		7.7	9.7		7.4	8.7	

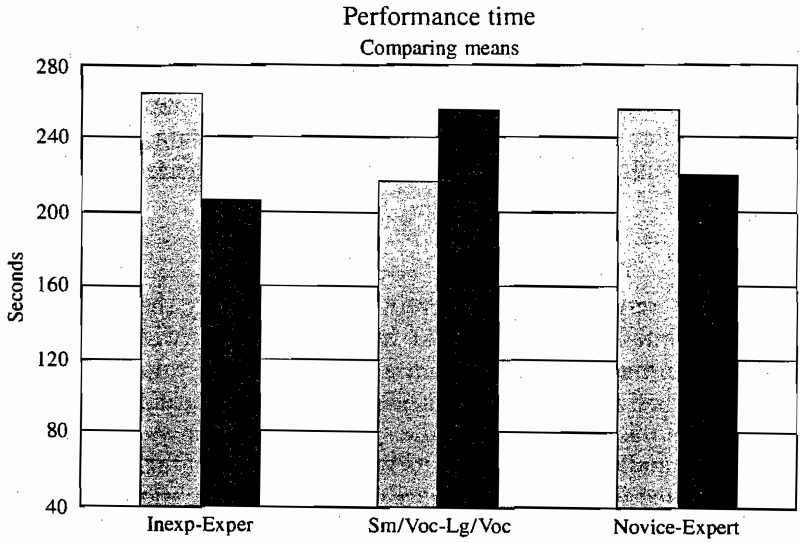

FIGURE 1.

The between-subjects ANOVA of the split-plot design found significant main effects for vocabulary (i.e. performance time of 217 s for the small vocabulary and 255 for the large) ($F(1, 24) = 5.11$, $p < 0.033$) and expertise (i.e. a mean performance time of 254 for novices and 219 for experts) ($F(1, 24) = 4.32$, $p < 0.049$) (see Figure 1).

Performance research question 2 was rejected. The performance of subjects with the large vocabulary was significantly slower than the performance of subjects with a small vocabulary. Performance research question 3 was accepted. Expert subjects performed faster than novice subjects.

There was no between-subjects interaction for vocabulary by expertise [$F(1, 24) = 0.03$, $p < 0.865$]. No effect was found for the within-subjects interactions of vocabulary by experience [$F(1, 24) = 0.46$, $p > 0.506$], expertise by experience [$F(1, 24) = 0.52$, $p > 0.476$] or vocabulary by expertise by experience [$F(1, 24) = 0.51$, $p > 0.482$].

5.2. USER ACCEPTANCE

The acceptability index (AI), the sum of the scale responses for the 12 bipolar scales for each subject under each treatment condition, was created with a range of 12–84. Twelve was the highest score of acceptability and 84 was the lowest. Because bipolar rating scales do not necessarily provide interval scale data, all analyses of the subjective rating scale data employed only non-parametric statistical tests (Casali *et al.*, 1990).

To validate the assessment data from both acceptance questionnaires, a Spearman rank correlation coefficient was computed for each response of the bipolar scales with the overall acceptable/unacceptable scale (Siegel & Castellan, 1988). The questionnaire results after the first assessment showed that all 12 of the scales were highly correlated ($r_s > 0.4180$, $p < 0.021$) and thus were considered attributes of acceptability.

The results for final cardiovascular assessment showed that 11 of the 12 scales were highly correlated ($r_s > 0.4178$, $p < 0.021$) with the acceptable/unacceptable scale. Item three (consistent/inconsistent) did not obtain significance ($r_s > 0.2678$, $p < 0.103$). Because of the strength of the questionnaire (i.e. significant results from a pilot study and references in the HCI literature, Casali *et al.*, 1990), item three was sustained as an attribute and included as a component of the AI.

The acceptance scores (AI) were subjected to two non-parametric data analysis techniques (Siegel & Castellan, 1988). For the within-subjects analysis of experience,

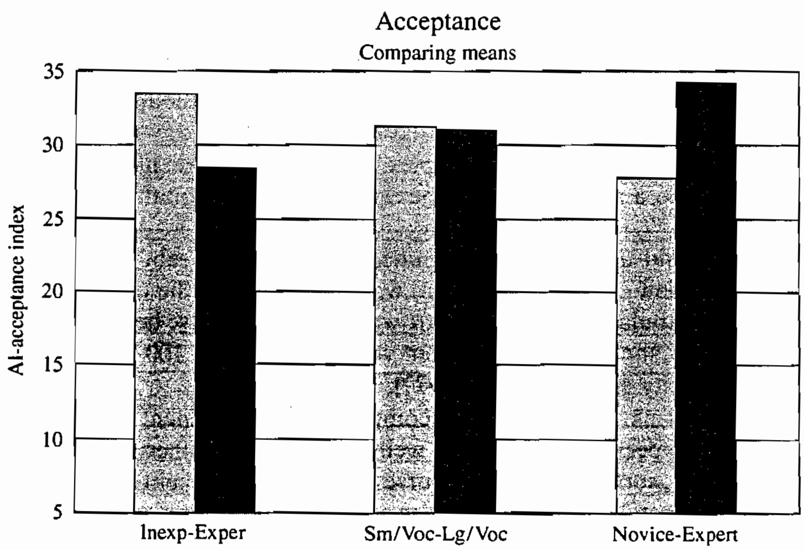

FIGURE 2.

a Wilcoxon Matched-Pairs Signed-Ranks Test identified a significant difference in user acceptance from the first (mean AI of 33.57) to the final assessment (mean AI of 28.43) ($p = 0.0001$) (see Figure 2). Research questionnaire 1 was accepted when subjects found a higher level of acceptance after gaining experience with the interface.

A Kolmogorov–Smirnov two-sample test did not find a significant difference for the between-subjects variables of vocabulary size ($p = 0.763$) or expertise ($p = 0.111$). However, there was a trend that novices (mean AI of 27.75) found speech input to be more acceptable than experts (mean AI of 34.25). We expected to find experts more accepting of the speech interface than novices. This was not visible. Vocabulary size showed no detectible difference in acceptance ratings (i.e. small vocabulary AI mean of 31.04 and a large vocabulary AI mean of 30.96).

6. Discussion

6.1. PERFORMANCE

The time required to complete the task was significantly affected by the user's level of experience with a speech-input interface. Performance time decreased on average 22% from the first assessment to the final assessment. Experienced subjects performed 58 s faster than inexperienced subjects.

There was a practice or learning effect with speech input. All subjects entered data in a more traditional way during the first assessment (i.e. found three to five assessment items, stopped the assessment, then entered the assessment data), but altered the data entry process by the fourth assessment to utilized speech input (i.e. found the assessment item and entered it by speech immediately). We found this process of immediate data entry to be a very important observation (Jones *et al.*, 1992; Noyes *et al.*, 1993).

The available vocabulary size had a significant influence on the time needed to complete a task. The large vocabulary increased (worsened) the performance time by 22%. Possibly, the large vocabulary, with alternative word choices and synonyms, increases the number of words (i.e. commands) employed by the user and adds to the complexity of the interface. As demonstrated by Wickens, Sandry and Vidulich (1983), when systems become increasingly complex the user is more likely to incorporate a complex system into a mental model. The creation of this mental model increases the time on task.

We expected to find a much more significant effect for vocabulary size. After exit interviews and a review of our experiment, we theorized that the difference in the size between the two vocabularies (i.e. 69 and 101 words) may not have been sufficiently large. Further experiments should involve larger vocabularies (e.g. 100 and 200 words) as recommended by Kelly and Chapanis (1977).

A significant performance difference for completing an assessment between novice and expert nurses was expected. Experts are well practiced at gathering assessment data (Calkin, 1984). In fact, some of the expert nurses were able to gather two assessment items at the same time (e.g. take a patient's pulse while also counting respirations). On the other hand, novices were only able to gather one assessment item at a time, increasing performance time.

Expert subjects were probably better able to adapt to the speech-input interface. For an expert nurse, performing an assessment requires very little concentration (Calkin, 1984). Including the interface involved only a small increase in effort.

6.2. USER ACCEPTANCE

The user's level of experience with the interface proved to have a significant effect on the user's subjective judgements of whether or not the device was an acceptable means of data entry for the cardiovascular assessment task. Experienced users found the interface more acceptable.

The results of this study support the inference that experience with speech input has an effect on the acceptance and performance of users performing a hands-busy/eyes-busy task. The continued use of the interface leads to improvements in performance and greater acceptability ratings.

A user's expertise has an influence on performance time. However, regardless of performance (i.e. task completion time), expertise did not affect acceptance. Close examination of the acceptance scores revealed that novices found the interface highly acceptable for both assessments. Therefore, such high initial acceptance scores provided little room for an increase, even as the novices gained experience and became more comfortable with the interface.

The available vocabulary did not have a major effect on the acceptability rating for the interface. Despite the trend that the small vocabulary increased performance, users did not rate the two available vocabularies differently. This may be based on the domain specificity of the vocabulary and the expectations of the users (Baber & Stammers, 1989). In fact, several nurses who used the large vocabulary did express that an even larger vocabulary with "more explicit medical terms" would have been more favorable.

The authors would like to thank Michael J. DeHaemer for his editorial contributions and theoretical guidance.

References

BABER, C. & STAMMERS, R. B. (1989). Is it natural to talk to computers? An experiment using the "Wizard of Oz" technique. In E. D. MEGAW, Ed. *Contemporary Ergonomics: Proceedings of the Ergonomic Society's Annual Conference*, pp. 234–239. London: Taylor & Francis.

BABER, C., USHER, D. M., STAMMERS, R. B. & TAYLOR, R. G. (1992). Feedback requirements for automatic speech recognition in the process control room. *International Journal of Man–Machine Studies*, **37**, 703–719.

BENNER, P. (1982). From novice to expert. *American Journal of Nursing*, **82**, 402–407.

BRAJNIK, G., GUIDA, G. & TASSO, C. (1990). User modeling in expert man–machine interface: a case study in intelligent information retrieval. *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-20**, 166–185.

CALKIN, J. D. (1984). A model for advanced nursing practice. *The Journal of Nursing Administration*, January, 24–30.

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.

DEHAEMER, M. J., WRIGHT, G. & DILLON, T. W. (1994). Automated speech recognition for spreadsheet tasks: performance effects for experts and novices. *International Journal of Human–Computer Interaction*, **6**, 299–318.

DILLON, T. W., MCDOWELL, T., NORCIO, A. F. & DEHAEMER, M. J. (1994). A preliminary investigation: nursing acceptance of a speech-input interface. *Computer in Nursing*, **12**, 264–271.
DREYFUS, S. E. & DREYFUS, H. L. (1980). A five-stage model of the mental activities involved in directed skill acquisition. Unpublished manuscript (Supported by the US Air Force, Office of Scientific Research under contract f49620-C-0063), Berkeley: University of California, Operations Research Center.
FURNAS, G. W., LANDAUER, T. K., GOMEZ, L. M. & DUMAIS, S. T. (1987). The vocabulary problem in human–system communication. *Communications of the ACM*, **30**, 964–971.
JONES, D. M., FRANKISH, C. R. & HAPESHI, K. (1992). Automatic speech recognition in practice. *Behaviour & Information Technology*, **11**, 109–122.
KELLY, M. J. & CHAPANIS, A. (1977). Limited vocabulary natural language dialogue. *International Journal of Man–Machine Studies*, **9**, 479–501.
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.
MARKOWITZ, J. (1993). The power of speech. *AI Expert*, January, 29–33.
MARTIN, G. L. (1989). The utility of speech input in user–computer interfaces. *International Journal of Man–Machine Studies*, **30**, 355–375.
MARTIN, T. B. (1976). Practical applications of voice input to machines. *Proceedings of the IEEE*, **64**, 487–501.
MCCAULEY, M. E. (1984). Human factors in voice technology. In F. A. MUCKLER, Ed., *Human Factors Review*, Chap. 4, pp. 131–166. New York: Media Dimensions.
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.
NORCIO, A. F. & STANLEY, J. (1989). Adaptive human–computer interfaces: a literature survey and perspective. *IEEE Transactions on Systems, Man and Cybernetics*, **19**, 399–408.
NOYES, J. M., BABER, C. & FRANKISH, C. R. (1992). Using automatic speech recognition in industrial applications. *Journal of the American Voice Input/Output Society*, **12**, 51–68.
NYE, J. M. (1982). Human factors analysis of speech recognition systems. *Speech Technology*, **1**, 50–57.
OVIATT, S. L. & COHEN, P. R. (1991). Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language*, **5**, 297–326.
POOCK, G. K., MARTISA, B. J. & ROLAND, E. F. (1983). *The effects of feedback to users with voice recognition equipment*. Report No. NP55-83-003. Monteray, CA: Naval Postgraduate School.
RUDNICKY, A. I. (1990). *The design of spoken language interfaces*. No. CMU-CS-90-118. Pittsburgh: Carnegie Mellon University, School of Computer Science.
RUDNICKY, A. I. & SAKAMOTO, M. (1989). *Transcription conventions and evaluation techniques for spoken language systems research*. No. CMU-CS-89-194. Pittsburgh: Carnegie Mellon University, School of Computer Science.
SCHURICK, J. M., WILLIGES, B. H. & MAYNARD, J. F. (1985). User feedback requirements with automatic speech recognition. *Ergonomics*, **28**, 1543–1555.
SHNEIDERMAN, B. (1992). *Designing the User Interface: Strategies for Effective Human–Computer Interaction*. 2nd. Addison-Wesley, Reading, Massachusetts.
SIEGEL, S. & CASTELLAN, N. J., Jr. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd edn.). New York: McGraw-Hill.
SIMPSON, C. A., MCCAULEY, M. E., ROLAND, E. F., RUTH, J. C. & WILLIGES, B. H. (1985). System design for speech recognition and generation. *Human Factors*, **27**, 115–141.
VISICK, D., JOHNSON, P. & LONG, J. (1984). The user of simple speech recognisers in industrial applications. In B. SHACKEL, Ed. *Human–Computer Interaction INTERACT '84*, pp. 209–213. Amsterdam: Elsevier.
WICKENS, C. D., SANDRY, D. L. & VIDULICH, M. (1984). Compatibility and resource competition between modalities of input, central processing and output. *Human Factors*, **25**, 227–248.
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.

Paper accepted for publication by Associate Editor, Dr David Hill.