Acceptance of a Speech Interface for Biomedical Data Collection

Michael A. Grasso, Ph.D., David Ebert, Ph.D., Tim Finin, Ph.D.
Segue Biomedical Computing, Laurel, Maryland and
Department of Computer Science and Electrical Engineering at the
University of Maryland Baltimore County, Baltimore, Maryland
grasso@cs.umbc.edu

Speech interfaces have the potential to address the data entry bottleneck of many applications in the field of medical informatics. An experimental study evaluated the effect of perceptual structure on a multimodal speech interface for the collection of histopathology data. A perceptually structured multimodal interface, using speech and direct manipulation, was shown to increase speed and accuracy. Factors influencing user acceptance are also discussed.

INTRODUCTION

Data entry has been identified as a key bottleneck in many biomedical applications [1,2,3]. Large volumes of information must be gathered by clinicians and researchers to support patient care and clinical trials. This data must be collected and managed according to specific protocols. Often the situation exists where a clinician is occupied with patient care and cannot document his or her findings until later. This interval of time between the generation of information and its recording can compromise the data collection process. Despite considerable advances in computer architectures over the last 20 years, the keyboard and video display remain the principal means of entering and retrieving data. New human-computer interface modalities are needed which can automate the data collection process at the *source*, where the information is actually generated.

One possibility is to develop interfaces using speech recognition technology. Speech is a natural form of communication that is pervasive, efficient, and can be used at a distance. However, widespread acceptance of speech as a computer interface has yet to occur. The reasons for this include limitations in technology as well as the need for theoretical models which can be used as guidelines for incorporating speech into the user interface. The objective of this paper is to review the application of speech in biomedical interfaces, discuss acceptance issues, and summarize a study of perceptual structure on a multimodal speech interface.

Speech In Biomedical Applications

Speech-driven computer interfaces can address two key concerns in biomedical computer interfaces: the demand for ease of use and constraints on the user's ability to work with the keyboard or mouse. Speech technology is still limited, however, with most successful systems using medium-sized vocabularies with well-defined grammar rules. As described in the literature, the main applications of speech include template-based reporting, natural language processing, multimodal integration of speech with other methods of input, and hands-busy data entry. The first two reflect the need for more intuitive interfaces. The latter two deal with limitations of traditional input using the keyboard or mouse. This topic is covered in greater detail elsewhere [4].

Template-based reporting has been applied to radiology, pathology, endoscopy, and emergency medicine [5,6,7]. The potential advantage is that turnaround time is decreased and accuracy is increased by eliminating the need for dictation and transcription by clerical personnel. An alternative to template-based reporting explored methods that circumvent shortcomings in the current technology while maintaining the flexibility and naturalness of speech [8,9].

Several efforts studied the use of speech interfaces to overcome limitations in traditional input devices. One such system, designed to assist in the collection of stereological data, combined speech input for object identification with a digitizing pad to enter X and Y coordinates [3]. For hands-busy and eyes-busy environments, systems have been developed to input clinical data during dental examinations [10], record information for anesthesiologists during medical procedures [11], and enter findings while reading images during the analysis of bone scintigraphic data [12].

Acceptance

A positive attitude from the user community is often the most critical factor determining the success of a computer application. End-users, frustrated by a system they believe does not enhance or possibly interferes with their work, will most likely abandon that application altogether.

Acceptance can be viewed as a measure of how well a system implements the original requirements or operational goals of the client. It can also been seen as some function of software quality, such as usability, reliability, resilience, or complexity. With respect to speech interfaces, user acceptance is complicated by additional factors such as limitations in current technology. Often expectations of how a speech interface should work are biased by our experience with human-to-human interaction.

A recent effort studied the use of a speech interface to facilitate the collection of cardiovascular data by nurses at the patient's bedside [2]. They reported that as nurses interact with the speech interface over time, the interface becomes more acceptable. Another study showed that a positive attitude toward computers could be a predictor of future use [13].

Initial work by the author includes a feasibility study of a speech interface for the collection of histopathology data [1]. A prototype speech-driven data collection system for histopathology data using only speech input and computer-generated speech responses was developed and tested. It was concluded that this architecture could be considered a viable alternative for hand-free, eyes-free data collection in animal toxicology studies with reasonable recognition accuracy. Based on user interviews, the main problems relating to acceptance were to minimize training requirements and improve audible feedback.

PROBLEM

Perception occurs in the head, somewhere between the observable stimulus and the response, and consists of various kinds of processing that have distinct costs [14]. By understanding and capitalizing on the underlying structure, it is believed that a perceptual system could reduce these costs and gain advantages in speech and accuracy. The dimensions of a structure are integral if they cannot be attended to individually, one at a time; otherwise, they are separable.

For human-computer interfaces, the theory of perceptual structure was extended to show that performance of a unimodal graphical environment improves when the structure of the perceptual space matches the control space of the input device [15]. A two-dimensional mouse and a three-dimensional tracker were used as input devices. Two graphical input tasks with three inputs each were evaluated, one where the inputs were integral (x location, y location, and size) and the other where the inputs were separable (x location, y location, and color). Common sense might say that a three-dimensional tracker is a logical superset of a two-dimensional mouse and therefore always as good and sometimes better than a mouse. Instead, the results showed that the tracker performed better when the three inputs were perceptually integral, while the mouse performed better when the three inputs were separable.

Another effort reported that the most significant factor in predicting the use of integrated multimodal speech and handwriting was contrastive functionality [16]. Here, the two modalities were used in a contrastive way to designate a shift in context or functionality, such as original input versus corrected, data versus command, digits versus text, or digits versus referring description.

Based on these results and the framework of complementary behavior between speech and direct manipulation [17], a research hypothesis was proposed which extended the theory of perceptual structure to multimodal interfaces using speech and direct manipulation.

Our general research hypothesis predicted that the speed, accuracy, and acceptance of the interface would increase if a single input device was used to enter attributes which are perceptually integral and two devices were used to enter attributes which are perceptually separable. For example, consider the following histopathology observation consisting of an organ, site, qualifier, and morphology: *lung alveolus marked inflammation*. It was assumed that the qualifier/morphology relationship was integral, since the qualifier describes the morphology, such as *marked inflammation*. The site/qualifier relationship was assumed to be separable, since the site identifies where in the organ the tissue was taken from, such as *alveolus lung*, not *alveolus marked*. The site/morphology relationship was assumed to be separable for the same reason. Additional background material can be found elsewhere [18].

METHOD

A software prototype was developed with two interfaces to test this hypothesis. The first was a baseline interface that used speech and mouse input in a way that did not match the perceptual structure of the attributes while the second interface used speech and mouse input in a way that best matched the perceptual structure. The software projected images of tissue slides on a computer monitor while subjects entered histopathologic observations in the form of topographical sites, qualifiers, and morphologies. The tissue slides for the experiment were provided by the National Center for Toxicological Research (Jefferson, AK). The vocabulary was based on the Pathology Code Table [19].

Twenty subjects from among the biomedical community participated in this experiment as unpaid volunteers. The sample population consisted of professionals with doctoral degrees (D.V.M., Ph.D., or M.D.), ranged in age from 33 to 51 years old, 11 were male, 9 were female, 15 were from academic institutions, 13 were born in the U.S., and 16 were native English speakers. The majority indicated they were comfortable using a computer and mouse and only one had any significant speech recognition experience. Since the main objective was to evaluate different user interfaces, participants did not necessarily have a high level of expertise in animal toxicology studies, but at a minimum, were familiar with tissue types and reactions.

The independent variables were the interface (baseline, perceptually structured) and the task order (slide group 1, slide group 2). These were counterbalanced between the subjects using a within-groups experiment design. The dependent variables were task completion time, speech errors, mouse errors, diagnosis errors, and user acceptance. Time and errors were tracked by the software prototype. Acceptance was measured with a subjective questionnaire containing 13 bi-polar adjective pairs used in previous human-computer interaction studies [2, 20]. Users rated each question on a scale of 1 to 7.

RESULTS

Speed and accuracy increased when a single modality was used to enter attributes which were integral and two modalities were used to enter attributes which were separable. Task completion time improved 41.5 seconds or 22.5% using the perceptually structured interface ($t(19) = 4.791$, $p < .001$, two-tailed). The results are summarized by task in Figure 1. ANOVA was used to show that interface order and task order had no significant effect on the results. Speech recognition errors were reduced by 36%, which was significant (paired $t(19) = 2.924$, $p < .01$, two-tailed). Error rates by task are shown in Figure 2. Mouse errors increased slighlty and diagnosis errors decreased slightly for the baseline interface, but where not significant ($p = .733$, $p = .858$, respectively).

Figure 1: Mean Task Completion Times

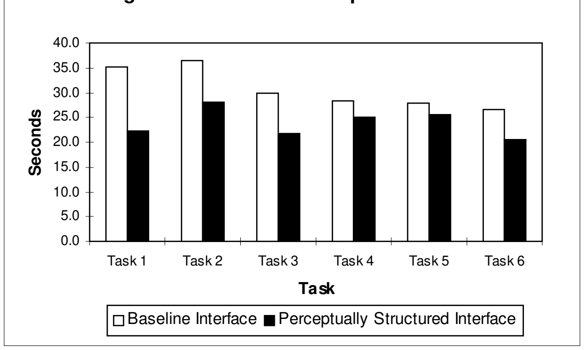


Figure 2: Mean Speech Error Rates

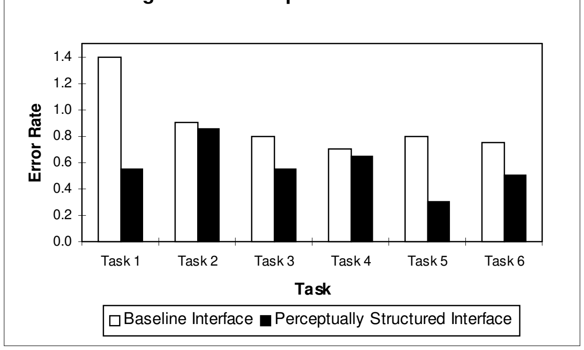
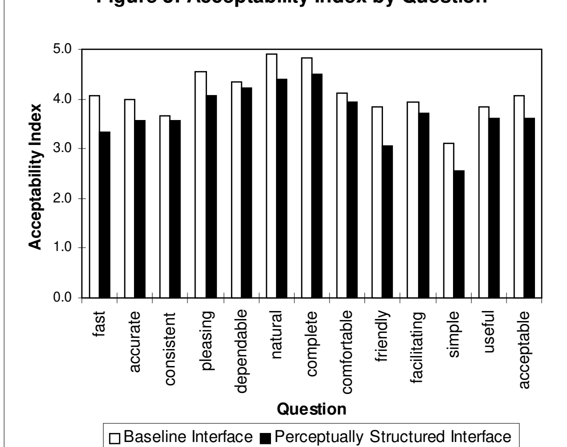


Figure 3: Acceptability Index by Question



An acceptability index (AI), based on the average ranking for each of the 13 bi-polar adjective pairs, showed an improvement of 2.4%. However, one subject's score was more than 2 standard deviations outside the mean AI. With this outlier removed, the perceptually structured interface showed a 6.7% improvement. A 2x13 ANOVA with repeated measures, to compare the 2 interfaces for the 13 questions, was significant ($p < .05$). The results were summarized by question in Figure 3, where a lower AI was indicative of greater acceptance.

DISCUSSION

The results of this experiment support the hypothesis when using a multimodal interface on multidimensional biomedical tasks. Task completion time and the number of speech errors were reduced when a single modality was used to enter attributes which were integral and two modalities were used to enter attributes which were separable. Results from mouse and diagnosis errors were not significant. This was most likely because very few mouse errors were recorded. Also, since each subject was allowed to review the slides before the test, the effect of perceptual structure on the ability to apply domain expertise was not measured.

From the subjective questionnaire, subjects felt the perceptually structured interface was faster and more accurate. This was substantiated by quantitative data on time and speech errors. Subjects also felt the perceptually structured interface was more consistent, pleasing, dependable, natural, complete, comfortable, friendly, facilitating, simple, useful, and acceptable.

Pearson's correlation coefficients were used to identify other factors which might influence acceptance. A positive correlation was observed between acceptance and the number of speech errors ($p < .01$), highlighting the importance of increasing recognition accuracy. The reduction of speech errors is typically viewed as a technical problem. However, this effort successfully reduced the rate of speech errors by applying user-interface principles based on perceptual structure. Similar to this, another study reported a reduction in spoken disfluencies by using more structured interfaces [21].

A significant positive correlation was also observed between the increased acceptance and decreased diagnosis errors ($p < .01$). Diagnosis errors were assumed to be inversely proportional to the domain expertise of each subject. What this finding suggests is that the more domain expertise a person has, the more he or she is likely to embrace the computer interface. This also suggests that including domain knowledge into the user interface would be advantageous.

No correlation was observed between acceptance and task completion time ($p > .05$). This occurred, even though the subjects believed the perceptually structured interface was faster, and quantitative data corroborated this as well. Thus, finding no relationship between acceptance and time suggests that overall user acceptance is predominantly influenced by something other than speed.

CONCLUSION

This effort applied the theory of perceptual structure to improve the speed, accuracy, and acceptance of a speech-driven biomedical computer interface. The perceptually structured interface significantly reduced task completion time and the number of speech errors. A moderate increase in user acceptance was also observed. User acceptance was influenced more by accuracy than speed. In addition, factors unrelated to the software itself affected acceptance, such as the level of domain expertise. In light of the need for increased automation for biomedical data collection, a better understanding of these issues is essential before the widespread acceptance of speech as a user interface can occur.

Acknowledgments

The authors wish to thank to Judy Fetters and Alan Warbritton from the National Center for Toxicological Research for providing tissue sides and other assistance with the software prototype. The authors also thank Lowell Groninger, Greg Trafton, and Clare Grasso for help with the experiment design. Finally, the authors thank those who graciously participated in this study from the University of Maryland Medical Center, the Baltimore Veteran Affairs Medical Center, the Johns Hopkins Medical Institutions, and the Food and Drug Administration.

References

1. Grasso MA, Grasso CT (1994). Feasibility Study of Voice-Driven Data Collection in Animal Drug Toxicology Studies. Computers in Biology and Medicine, 24(4):289-294.
2. Dillon TW, McDowell D, Norcio AF, DeHaemer MJ (1994). Nursing Acceptance of a Speech-

Input Interface: A Preliminary Investigation. Computers in Nursing, 12(6):264-271.

3. McMillan PJ, Harris JG (1990). Datavoice: A Microcomputer-Based General Purpose Voice-Controlled Data-Collection System. Computers in Biology and Medicine, 20(6):415-419.
4. Grasso MA (1995). Automated Speech Recognition in Medical Applications. M.D. Computing, 12(1):16-23.
5. Klatt EC (1991). Voice-Activated Dictation for Autopsy Pathology. Computers in Biology and Medicine, 21(6):429-433.
6. Massey BT, Geenen JE, Hogan WJ (1991). Evaluation of a Voice Recognition System for Generation of Therapeutic ERCP Reports. Gastrointestinal Endoscopy, 37(6):617-620.
7. Hollbrook JA (1992). Generating Medical Documentation Through Voice Input: The Emergency Room. Topics in Health Records Management, 12(3):58-63.
8. Issacs E, Wulfman CE, Rohn JA, Lane CD, Fagan LM (1993). Graphical Access to Medical Expert System: IV. Experiments to Determine the Role of Spoken Input. Methods of Information in Medicine, 32(1):18-32.
9. Wulfman, C. E., Rua, M., Lane, C. D., Shortliffe, E. H., Fagan, L. M. (1993). Graphical Access to Medical Expert System: V. Integration with Continuous-Speech Recognition. Methods of Information in Medicine, 32(1):33-46.
10. Feldman CA, Stevens D (1990). Pilot Study on the Feasibility of a Computerized Speech Recognition Charting System. Community Dentistry and Oral Epidemiology, 18:213-215.
11. Smith NT, Brian RA, Pettus DC, Jones BR, Quinn ML, Sarnat L (1990). Recognition Accuracy with a Voice-Recognition System Designed for Anesthesia Record Keeping. Journal of Clinical Monitoring, 6(4):299-306.
12. Ikerira H, et al. (1990). Analysis of Bone Scintigram Data Using Speech Recognition Reporting System. Radiation Medicine, 8(1):8-12.
13. Thomas BS, Delaney CW, Weiler K (1992). The Affective Outcomes of Course Work on Computer Technology in Nursing. Journal of Nursing Education, 31(4):165-170.
14. Garner WR (1974). The Processing of Information and Structure. Lawrence Erlbaum, Potomac, Maryland.
15. Jacob RJK, et al. (1994). Integrality and Separability of Input Devices. ACM Transactions on Computer-Human Interaction, 1(1):3-26.
16. Oviatt SL, Olsen E (1994). Integration Themes in Multimodal Human-Computer Interaction. In Proceeding of the International Conference on Spoken Language Processing, volume 2, pp. 551-554, Acoustical Society of Japan.
17. Cohen PR (1992). The Role of Natural Language in a Multimodal Interface. In Proceedings of the ACM Symposium on User Interface Software and Technology, Monterey California, pp. 143-149, ACM Press, November 15-18.
18. Grasso M.A (1997). Speech Input in Multimodal Environments: Effects of Perceptual Structure on Speed, Accuracy, and Acceptance. University of Maryland Baltimore County, Doctoral Dissertation.
19. Pathology Code Table Reference Manual, Post Experiment Information System (1985). National Center for Toxicological Research, TDMS Document #1118-PCT-4.0, Jefferson, Ark.
20. Casali SP, Williges BH, Dryden RD (1990). Effects of Recognition Accuracy and Vocabulary Size of a Speech Recognition System on Task Performance and user Acceptance. Human Factors, 32(2):183-196.
21. Oviatt SL (1996). Multimodal Interfaces for Dynamic Interactive Maps. In Proceedings of the Conference on Human Factors in Computing Systems (CHI'96), ACM Press, New York, pp. 95-102.