The Effect of Perceptual Structure on Multimodal Speech Recognition Interfaces

Michael A. Grasso, Ph.D., David Ebert, Ph.D., Tim Finin, Ph.D.
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, Maryland USA
grasso@cs.umbc.edu

Abstract

A framework of complementary behavior has been proposed which maintains that direct manipulation and speech interfaces have reciprocal strengths and weaknesses. This suggests that user interface performance and acceptance may increase by adopting a multimodal approach that combines speech and direct manipulation. This effort examined the hypothesis that the speed, accuracy, and acceptance of multimodal speech and direct manipulation interfaces will increase when the modalities match the perceptual structure of the input attributes. A software prototype which supported a typical biomedical data collection task was developed to test this hypothesis. A group of 20 clinical and veterinary pathologists evaluated the prototype in an experimental setting using repeated measures. The results of this experiment supported the hypothesis that the perceptual structure of an input task is an important consideration when designing a multimodal computer interface. Task completion time, the number of speech errors, and user acceptance improved when interface best matched the perceptual structure of the input attributes.

Introduction

For many applications, the human computer interface has become a limiting factor. One such limitation is the demand for intuitive interfaces for non-technical users, a key obstacle to the widespread acceptance of computer automation [Landau, Norwich, and Evans 1989]. Another difficulty consists of hands-busy and eyes-busy restrictions, such as those found in the biomedical area during patient care or other data collection tasks. An approach that addresses both of these limitations is to develop interfaces using automated speech recognition. 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 human computer interface has yet to occur. This effort seeks to cultivate the speech modality by evaluating it in a multimodal environment with direct manipulation. Preliminary work on this effort has already been published [Grasso and Finin 1997]. The specific focus of this paper is to document how perceptual structure can effect the speed, accuracy, and acceptance of a multimodal interface.

Speech and Direct Manipulation Modalities

Compared to more traditional modalities, speech interfaces have a number of unique characteristics. The most significant is that speech is temporary. Once uttered, auditory information is no longer available. This can place extra memory burdens on the user and severely limit the ability to scan, review and cross-reference information. Speech can be used at a distance which makes it ideal for hands-busy and eyes-busy situations. It is omnidirectional and therefore can communicate with multiple users. However, this has implications related to privacy and security. Finally, more than other modalities, there is the possibility of anthropomorphism when using a speech interface. It has been documented that users tend to overestimate the capabilities of a system if a speech interface is used and that users are more tempted to treat the device as another person [Jones, Hapeshi, and Frankish 1990].

At the same time, speech recognition systems often carry technical limitations, such as speaker dependence, continuity, and vocabulary size. Speaker dependent systems must be trained by each individual user, but typically have higher accuracy rates than speaker independent systems, which can recognize speech from any person. Continuous speech systems recognize words spoken in a natural rhythm while isolated word systems require a deliberate pause between each word. Although more desirable, continuous speech is harder to process, because of the difficulty in detecting word boundaries. Vocabulary size can vary anywhere from 20 words to more than 40,000 words. Large vocabularies cause difficulties in maintaining recognition accuracy, but small vocabularies can impose unwanted restrictions. A more thorough review of this subject can be found elsewhere [Peacocke and Graf 1990].

Direct manipulation, made popular by the Apple Macintosh and Microsoft Windows graphical user interfaces, is based on the visual display of objects of interest, the selection by pointing, rapid and reversible actions, and continuous feedback [Shneiderman 1993]. The display in a direct manipulation interface should indicate a complete image of the application's environment, including its current state, what errors have occurred, and what actions are appropriate. A virtual representation of reality is created, which can be manipulated by the user through physical actions like pointing, clicking, dragging, and sliding.

While this approach has several advantages, arguments have been made that direct manipulation is inadequate for supporting fundamental transactions in applications such as word processing, CAD, and database queries. These comments were made in reference to the limited means of object identification and how the non-declarative aspects of direct manipulation can result in an interface that is too low-level [Buxton 1993; Cohen and Oviatt 1994]. Shneiderman [1993] points to ambiguity in the meanings of icons and limitations in screen display space as additional problems with direct manipulation.

Multimodal Input Tasks

Taking these observations into account, a framework of complementary behavior was proposed, suggesting that direct manipulation and speech interfaces have reciprocal strengths and weaknesses [Cohen 1992]. This suggests that user interface performance and acceptance may increase by adopting a multimodal approach that combines speech and direct manipulation. The objective of this study was to help define a theoretical basis for this framework by evaluating the effect of perceptual structure on this multimodal interface. Two research efforts are provided here for additional background information. The first reported how the perceptual structure of input devices can affect the performance of unimodal tasks. The second examines those conditions under which a person is most likely to combine two modalities.

Theory of Perceptual Structures

Perception occurs in the head, somewhere between the observable stimulus and the response. Perception consists of various kinds of processing that have distinct costs, so that this response can not be viewed as just a simple representation of a stimulus. 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. Garner documented that the dimensions of a structure can be characterized as integral or separable and that this relationship may affect performance under certain conditions [Garner 1974; Garner and Felfoldy 1970]. The dimensions of a structure are integral if they cannot be attended to individually, one at a time; otherwise, they are separable.

Structures abound in the real world and are used by people to perceive and process information. Structure can be defined as the way the constituent parts are arranged to give something its distinctive nature. It is not limited to shape or other physical stimuli, but is an abstract property transcending any particular stimulus. Information and structure are essentially the same in that they are the property of a stimulus which is perceived and processed.

A structured system is one that contains redundancy. The following examples illustrate that the principle of redundancy is pervasive in the world around us. A crude, but somewhat useful method for weather forecasting is that the weather today is a good predictor of the weather tomorrow. An instruction cache can increase computer performance because the address of the last memory fetch is a good predictor of the address of the next fetch. Consider a visual picture on a video screen. The adjacent pixels are usually similar to each other. Without this structure, the video screen would be perceived as meaningless noise or snow.

Considering these principles, one research effort tested the hypothesis that performance improves when the perceptual structure of the task matches the control structure of the input device [Jacob et al. 1994]. 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.

The theory of perceptual structures, integral and separable, was originally developed by Garner [1974]. The structure has to do with how the dimensions of the input task combine perceptually. This theory was extended with the hypothesis that the perceptual structure of an input task is key to the performance of unimodal input devices on multidimensional tasks. An appropriate follow-on question is the performance of integral and separable tasks on multimodal interfaces.

Integrating Input Modalities

Another effort examined how people might integrate input from different devices in a multimodal computer interface [Oviatt and Olsen 1994]. The study used a simulated service transaction system with verbal, temporal, and computational input tasks using both structured and unstructured interactions. Participants were free to use either handwriting, speech, or both during testing. User preferences were reported as follows: digits were more likely written than text, proper names were more likely written than other textual content, and structured interactions were more likely written than unstructured interactions.

The most significant factor in predicting the use of integrated multimodal speech and handwriting was contrastive functionality. 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 and referring description. Of all the transactions using writing and speech, 57% were due to one of the contrastive pattern. Also reported was the tendency toward certain combinations, such as written data and spoken command being roughly 3 times as likely as spoken data with written command.

While this study predicted users would prefer multimodal to unimodal interfaces, a follow-up study explored whether there were performance advantages as well [Oviatt 1996]. They evaluated speech, handwriting, and multimodal interfaces for map-based input tasks. It was reported that increasing the length of spoken utterances and using unstructured displays increased the number of disfluencies. Speech-only input also resulted in more performance errors and increased task completion time. Participants revealed a preference to using speech and writing for complementary functions. This was backed up by quantitative data showing the greatest speed advantages of multimodal input with pen-based pointing and gestures to identify location and speech for other data input.

Research Hypotheses

The general research question proposed for this effort is to determine what multidimensional tasks can best be integrated with multimodal speech and direct manipulation. Predicted results were that the speed, accuracy, and acceptance of multidimensional multimodal input will increase when the attributes of the task are perceived as separable, and for unimodal input will increase when the attributes are perceived as integral. Three null hypotheses were generated.

($H1_0$) The integrality of input attributes has no effect on the speed of the user.
($H2_0$) The integrality of input attributes has no effect on the accuracy of the user.
($H3_0$) The integrality of input attributes has no effect on acceptance by the user.

Studies which can provide theoretical models on the use of speech as an interface modality are significant in several ways. A foundational approach for research in human computer interaction calls for studies which replace anecdotal arguments with scientific evidence [Shneiderman 1993]. Bradford [1995] states that there are almost certainly applications where speech is the more natural medium and calls for comparative studies to determine where and when speech functions most effectively as a user interface. Cole et al. [1995] note the role that spoken language should ultimately play in multimodal systems is not well understood and call for the development of theoretical models from which predictions can be made about the strengths, weaknesses, and overall performance of different types of unimodal and multimodal systems.

Histopathologic data collection in animal toxicology studies was chosen as the application domain for user testing. Applications in this area include several significant hands-busy and eyes-busy restrictions during microscopy, necropsy, and animal handling. It is based on a highly structured, specialized, and moderately sized vocabulary with an accepted medical nomenclature. These and other characteristics make it a prototypical data collection task, similar to those required in biomedical research and clinical trials, and therefore a good candidate for a speech interface [Grasso 1995]. Also, the input tasks mainly involve reference identification, with little declarative, spatial, or computational data entry required, which should minimize any built-in bias toward either direct manipulation or speech.

Methodology

A set of software tools was developed to simulate a typical biomedical data collection task in order to test the validity of this hypothesis. The main data entry task the software supported was to project 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 prototype computer program was developed using Microsoft Windows 3.11 (Redmond, WA) and Borland C++ 4.51 (Borland International, Inc., Scotts Valley, CA).

The PE500+ was used for speech recognition (Speech Systems, Inc, Boulder, CO). The hardware came on a half-sized, 16-bit ISA card along with head-mounted microphone and speaker, and accompanying software development tools. Software to drive the PE500+ was written in C++ with the SPOT application programming interface. The Voice Match Tool Kit was used for grammar development. The environment supported speaker-independent, continuous recognition of large vocabularies, constrained by grammar rules. The vocabulary was based on the Pathology Code Table [1985] and was derived from a previous effort establishing the feasibility of speech input for histopathologic data collection [Grasso and Grasso 1994]. Roughly 1,500 lines of code were written for the prototype.

The software and speech recognition hardware were deployed on a portable PC-III computer with a 12.1 inch, 800x600 TFT color display, a PCI Pentium-200 motherboard, 32 MB RAM, and 2.5 GB disk drive (PC Portable Manufacture, South El Monte, CA). This provided a platform that could accept ISA cards and was portable enough to take to the participants' facilities for testing.

Independent Variables

The two independent variables for the experiment were the interface (baseline, perceptually structured) and task order (slide group 1, slide group 2). The input task was to enter histopathologic observations consisting of three attributes: topographical site, qualifier, and morphology. Consider the following observation consisting of an organ, site, qualifier, and morphology: *lung alveolus marked inflammation*. It was assumed that the qualifier/morphology (QM) relationship was integral, since the qualifier is used to describe or modify the morphology, such as *marked inflammation*. The site/qualifier (SQ) 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 (SM) relationship was assumed to be separable for the same reason. Based on these assumptions and the general research hypothesis, Table 1 predicted which modality would lead to improvements in the computer interface.

Data Entry Task		**Perception**	**Modality**
(SQ)	Enter Site and Qualifier	Separable	Multimodal
(SM)	Enter Site and Morphology	Separable	Multimodal
(QM)	Enter Qualifier and Morphology	Integral	Unimodal

Table 1: Predicted Modalities for Computer-Human Interface Improvements

The three input attributes (site, qualifier, morphology) and two modalities (speech, mouse) yielded a possible eight different user interface combinations for the software prototype as shown in Table 2. Also in this table are the predicted interface improvements for entering each pair of attributes (SQ, SM, QM) identified with a "+" or "-" for a predicted increase or decrease, respectively. The third alternative was selected as the *perceptually structured* interface, because the choice of input devices was thought to best match the perceptual structure of the attributes. The fifth alternative was the *baseline* interface, since the input devices least match the perceptual structure of the attributes. The third and fifth alternatives were selected over other equivalent ones, because they both required two speech inputs, one mouse input, and the two speech inputs appeared adjacent to each other on the computer screen.

Modality	Site	Qual	Morph	SQ	SM	QM	Interface
1. Mouse	M	M	M	-	-	+	
2. Speech	S	S	S	-	-	+	
3. Both	M	S	S	+	+	+	Perceptually Structured
4. Both	S	M	M	+	+	+	
5. Both	S	S	M	-	+	-	Baseline
6. Both	M	M	S	-	+	-	
7. Both	S	M	S	+	-	-	
8. Both	M	S	M	+	-	-	

Table 2: Possible Interfaces Combinations for the Software Prototype

Dependent Variables

The dependent variables for the experiment were speed, accuracy, and acceptance. The first two were quantitative measures while the latter was subjective.

Speed and accuracy were recorded both by the experimenter and the software prototype. Time was defined as the time it takes a participant to complete each of the 12 data entry tasks and was recorded to nearest millisecond. Three measures of accuracy were recorded: speech errors, mouse errors, and diagnosis errors. Speech errors were counted when the prototype incorrectly recognized a spoken utterance by the participant. This was because the utterance was misunderstood by the prototype or was not a valid phrase from the vocabulary. Mouse errors were recorded when a participant accidentally selected an incorrect term from one of the lists displayed on the computer screen and later changed his or her mind. Diagnosis errors were identified as when the input did not match the most likely diagnosis for each tissue slide. The actual speed and number of errors was determined by analysis of diagnostic output from the prototype, recorded observations of the experimenter, and review of audio tapes recorded during the study.

User acceptance data was collected with a subjective questionnaire containing 13 bi-polar adjective pairs which has been used in other human computer interaction studies [Casali, Williges, and Dryden 1990; Dillon 1995]. The adjectives are listed in Table 3. The questionnaire was given to each participant after testing was completed. An acceptability index (AI) was defined as the mean of the scale responses, where the higher the value, the lower the user acceptance.

User Acceptance Survey Questions					
1.	fast	slow	8.	comfortable	uncomfortable
2.	accurate	inaccurate	9.	friendly	unfriendly
3.	consistent	inconsistent	10.	facilitating	distracting
4.	pleasing	irritating	11.	simple	complicated
5.	dependable	undependable	12.	useful	useless
6.	natural	unnatural	13.	acceptable	unacceptable
7.	complete	incomplete			

Table 3: Adjective Pairs used in the User Acceptance Survey

Subjects

Twenty subjects from among the biomedical community participated in this experiment as unpaid volunteers between January and February 1997. Each participant reviewed 12 tissue slides, resulting in a total of 240 tasks for which data was collected. The target population consisted of veterinary and clinical pathologists from the Baltimore-Washington area. Since the main objective was to evaluate different user interfaces, participants did not need a high level of expertise in animal toxicology studies, but only to be familiar with tissue types and reactions. The participants came from the University of Maryland Medical Center (Baltimore, MD), the Veteran Affairs Medical Center (Baltimore, MD), the Johns Hopkins Medical Institutions (Baltimore, MD), the Food and Drug Administration Center for Veterinary Medicine (Rockville, MD), and the Food and Drug Administration Center for Drug Evaluation and Research (Gaithersburg, MD). To increase the likelihood of participation, testing took place at the subjects' facilities.

The 20 participants were distributed demographically as follows, based on responses to the pre-experiment questionnaire. 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 1 had any significant speech recognition experience.

The subjects were randomly assigned to the experiment using a within-group design. Half of the subjects were assigned to the perceptually-structured-interface-first, baseline-interface-second group and were asked to complete six data entry tasks using the perceptually structured interface and then complete six tasks using the baseline interface. The other half of the subjects were assigned to the baseline-interface-first, perceptually-structured-interface-second group and completed the tasks in the reverse order. Also counterbalanced were the tissue slides examined. Two groups of 6 slides with roughly equivalent difficulty were randomly assigned to the participants. This resulted in 4 groups based on interface and slide order as shown in Table 4. For example, subjects in group *BIP2* used the baseline interface with slides 1 through 6 followed by the perceptually structured interface with slides 7 through 12.

Group	Interface Order	Slide Order
B1P2	Baseline, Perceptually Structured	1-6, 7-12
B2P1	Baseline, Perceptually Structured	7-12, 1-6
P1B2	Perceptually Structured, Baseline	1-6, 7-12
P2B1	Perceptually Structured, Baseline	7-12, 1-6

Table 4: Subject Groupings for the Experiment

Procedure

A within-groups experiment, fully counterbalanced on input modality and slide order was performed. Each subject was tested individually in a laboratory setting at their place of employment or study. Participants were first asked to fill out the pre-experiment questionnaire and were told that the objective of this study was to evaluate several user interfaces in the context of collecting histopathology data. They were instructed that a computer program will project images of tissue slides on a computer monitor while they enter observations in the form of topographical sites, qualifiers, and morphologies. This was followed by a brief training session to become familiar with the test environment.

Before the actual test, participants were allowed to review the two sets of tissue slides. The goal here was to make sure participants were comfortable reading the slides before the test began. This was to ensure the experiment was measuring data input and not the ability of the subjects to read slides. While reviewing slides, participants were encouraged to ask questions about possible diagnoses. During testing, participants entered two groups of six histopathologic observations based on the group they were randomly assigned to. They were encouraged to work at a normal pace that was comfortable for them and to ask questions before the actual test began. After the test, the user acceptance survey was administered.

Results

For each participant, speed was measured as the time to complete the 6 baseline interface tasks, the time to complete the 6 perceptually structured interface tasks, and time improvement (baseline interface time - perceptually structured interface time). The mean improvement for all subjects was 41.468 seconds. A ***t*** test on the time improvements was significant ($t(19) = 4.791$, $p < .001$, two-tailed). A single-factor ANOVA comparing the baseline and perceptually structured interface times was significant ($F(1,38) = 4.719$, $p < .05$, two-tailed). A comparison of mean task completion times is in Figure 1.

ANOVA was also used to show that interface order and task order had no significant effect on the results. A single-factor ANOVA comparing the baseline-interface-first-group and base-interface-second groups was not significant ($F(1,18) = 0.123$, $p = 0.730$, two-tailed). A single factor ANOVA comparing the perceptually-structured-interface-first and perceptually-structured-interface-second groups was not significant ($F(1,18) = 0.723$, $p = 0.406$, two-tailed). A single factor ANOVA comparing the slide-group-two-first and slide-group-two-second groups was not significant ($F(1,18) = 3.440$, $p = 0.080$, two-tailed). A single factor ANOVA comparing the slide-group-two-first and slide-group-two-second groups was not significant ($F(1,18) = 1.650$, $p = 0.215$, two-tailed).

Three types of user errors were recorded: speech recognition errors, mouse errors, and diagnosis errors. The baseline interface had a mean speech error rate of 5.35, and the perceptually structured interface had mean of 3.40. The reduction in speech errors was significant (paired $t(19) = 2.924$, $p < .01$, two-tailed). Mouse errors for the baseline interface had mean error rate of 0.35, while the perceptually structured interface had mean of 0.45. Although the baseline interface had fewer mouse errors, these results were not significant (paired $t(19) = 0.346$, $p = .733$, two-tailed). For diagnosis errors, the baseline interface had mean error rate of 1.80, and the perceptually structured interface had mean of 1.85. Again, although the rate for the baseline interface was slightly better, these results were not significant (paired $t(19) = 0.181$, $p = 0.858$, two-tailed). A comparison of mean error rates by task is shown in Figure 2.

For analyzing the subjective scores, an acceptability index by question was defined as the mean scale response for each question across all participants. A lower AI was indicative of higher user acceptance. The overall AI was 3.81 for the baseline interface and 3.72 for the perceptually structured interface, with 10 of 13 questions showing improvement. The results were not significant ($p = .187$) using a 2x13 ANOVA with repeated measures, comparing the 2 interfaces for the 13 questions. However, one subject's score was more than 2 standard deviations outside the mean AI (subject 17). With this outlier removed, the baseline interface AI was 3.99 and the perceptually structured interface was 3.63, which was a modest 6.7% improvement. All 13 questions showed improvement, and the result was significant using the 2x13 ANOVA ($p = .014$). A comparison of these values is shown in Figure 3.

Discussion

The results of this study support the general findings that for multimodal speech and direct manipulation interfaces, multidimensional input tasks work best when the input attributes are perceived as separable, and unimodal interfaces work best when the inputs are perceived as integral. Of the three null hypotheses identified before the study began, two were rejected in favor of an alternate hypothesis based on predicted results. One of the null hypotheses was rejected in part, in favor of predicted results. In addition, several significant relationships between dependent variables were observed.

The first null hypothesis stated: **($H1_0$)** *The integrality of input attributes has no effect on the speed of the user*. As reported, a significant improvement in task completion time was observed when integral input attributes used the same modality and separable attributes used different modalities. The improvement in total time was 41.468 seconds, or about 22.5%, which was significant ($p < .001$). Of the 20 participants, 18 saw improvement with the perceptually structured interface. Strengthening this finding was a significant ANOVA that times from the baseline and perceptually structured groups were from different populations. ANOVA also showed that interface order and task order had no significant effect on the results. The null hypothesis was rejected in support of an alternate hypothesis: **($H1_A$)** *The speed of multidimensional, multimodal interfaces will increase when the attributes of the task are perceived as separable, and for unimodal interfaces will increase when the attributes of the task are perceived as integral*.

The second null hypothesis stated: **($H2_0$)** *The integrality of input attributes has no effect on the accuracy of the user*. As reported, there were 1.95 less speech errors with the perceptually structured group, or a 36% improvement, with 16 of the 20 subjects making less speech errors using the perceptually structured interface. The reduction in speech errors was significant (p <

.01). For mouse errors and diagnosis errors, there was a slight improvement with the baseline group, but these were not significant.

The reason why mouse errors did not follow predicted results was possibly because there were few such errors recorded. Across all subjects, there were only 16 mouse errors compared to 175 speech errors. A mouse error was recorded only when a subject clicked on the wrong item from a list and later changed his or her mind, which was a rare event.

There were 77 diagnosis errors, but these also did not follow predicted results. Diagnosis errors were really a measure of the subject's expertise in identifying tissue types and reactions. Ordinarily, this type of finding would suggest that there is no relationship between perceptual structure of the input task and the ability of the user to apply domain expertise. However, this cannot be concluded from this study, since efforts were made to avoid measuring a subject's ability to apply domain expertise by allowing them to review the tissue slides before the actual test.

The null hypothesis was accepted in part: **($H2'_0$)** *The integrality of input attributes has no effect on accuracy of the user, regarding mouse errors and applying domain expertise*. The null hypothesis was rejected with respect to speech errors in support of the modified alternate hypothesis: **($H2'_A$)** *With respect to speech input, the accuracy of multidimensional, multimodal interfaces will increase when the attributes of the task are perceived as separable, and for unimodal interfaces will increase when the attributes of the task are perceived as integral.*

The third null hypothesis stated: **($H3_0$)** *The integrality of input attributes has no effect on acceptance by the user*. As stated earlier, once the outlier was removed, the overall AI was 3.99 for the baseline group and 3.63 for the perceptually structured group, an improvement of 6.7%, which was significant using a 2x13 ANOVA ($p = .014$). The null hypothesis was rejected in support of the alternate hypothesis: **($H3_A$)** *The acceptance of multidimensional, multimodal interfaces will increase when the attributes of the task are perceived as separable, and for unimodal interfaces will increase when the attributes of the task are perceived as integral.*

The Pearson correlation coefficient was computed to reveal possible relationships between the dependent variables. This includes relationships between the baseline and perceptually structured interface, relationships with task completion time, and relationships with user acceptance.

A positive correlation of time between the baseline interface and perceptually structured interface was probably due to the fact that a subject who works slowly (or fast) will do so regardless of the interface ($p < .001$). The positive correlation of diagnosis errors between the baseline and perceptually structured interface suggests that a subject's ability to apply domain knowledge was not effected by the interface ($p < .001$). This was probably due to the fact that subjects were allowed to review the slides before the actual test. The lack of correlation for speech errors was notable. Under normal circumstances, one would expect there to be a positive correlation, implying that a subject who made errors with one interface was predisposed to making errors with the other. Having no correlation agrees with the finding that the user was more likely to make speech errors with the baseline interface, because the interface did not match the perceptual structure of the input task.

When comparing time to other variables, several relationships were found. There was a positive correlation between the number of speech errors and task completion time ($p < .01$). This was expected, since it takes time to identify and correct these errors. There was also a positive correlation between time and the number of mouse errors. However, due to the relatively few mouse errors which were recorded, nothing was inferred from these results. No correlation was observed between task completion time and diagnosis errors. Normally, one could assume that a lack of domain knowledge would lead to a higher task completion time. For this experiment, subjects were allowed to review slides before the actual test. This was to ensure that the experiment was measuring data entry time and other attributes of user interface performance, and not the ability of participants to read tissue slides. Finding no correlation suggests that this goal was accomplished.

Several relationships were identified between the acceptability index and other variables. Note that for the acceptability index, a lower score corresponds to higher user acceptance. A significant positive correlation was observed between acceptability index and the number of speech errors ($p < .01$). An unexpected result was that no correlation was observed between task completion time and the acceptability index. . This suggests that accuracy is more critical than speed, with respect to whether a user will embrace the computer interface. No correlation was found between the acceptability index and mouse errors, most likely due to the lack of recorded mouse errors. A significant positive correlation was observed between the acceptability index and 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 approve of the computer interface.

Summary

A research hypothesis was proposed for multimodal speech and direct manipulation interfaces. It stated that multimodal multidimensional interfaces work best when the input attributes are perceived as separable, and that unimodal multidimensional interfaces work best when the inputs are perceived as integral. This was based on previous research that extended the theory of perceptual structure [Garner 1972] to show that performance of multidimensional, unimodal, graphical environments improves when the structure of the perceptual space matches the control space of the input device [Jacob et al. 1994]. Also influencing this study was the finding that contrastive functionality can drive a user's preference of input devices in multimodal interfaces [Oviatt and Olsen 1994] and the framework for complementary behavior between speech and direct manipulation [Cohen 1992].

The results of this experiment support the hypothesis when using a multimodal interface on multidimensional biomedical tasks. Task completion time, accuracy, and user acceptance all increased when a single modality was used to enter attributes which were integral and two modalities were used to enter attributes which were separable. 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.

A group of 20 clinical and veterinary pathologists evaluated the interface in an experimental setting, where data on task completion time, speech errors, mouse errors, diagnosis errors, and user acceptance was collected. Task completion time improved by 22.5%, speech errors were reduced by 36%, and user acceptance increased 6.7% for the interface that best matched the perceptual structure of the attributes. Mouse and diagnosis errors decreased slightly for the baseline interface, but these were not statistically significant. User acceptance was shown to be related to speech recognition errors and domain errors, but not task completion time.

Additional research into theoretical models which can predict the success of speech input in multimodal environments are needed. Future directions could include additional studies on domain expertise and minimizing speech errors. The reduction of speech errors is typically viewed as a technical problem. However, this effort successfully reduced the rate of speech errors by applying certain user-interface principles based on perceptual structure. Others have reported a reduction in spoken disfluencies by applying other user-interface techniques [Oviatt 1996]. Also, noting the strong relationship between user acceptance and domain expertise, additional research on how to build domain knowledge into the user interface might be helpful.

Acknowledgements

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.

Appendix

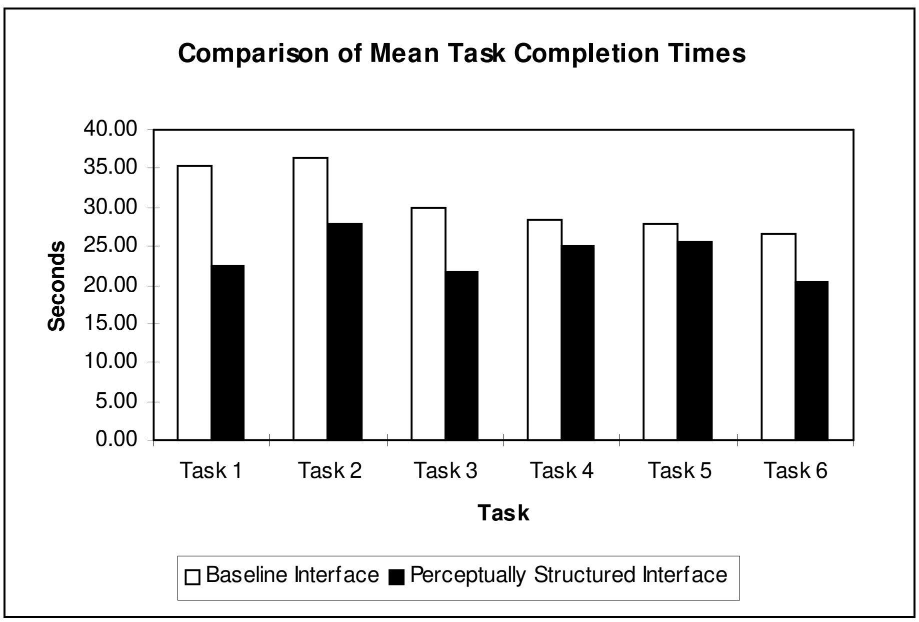

Figure 1: Comparison of Mean Task Completion Times

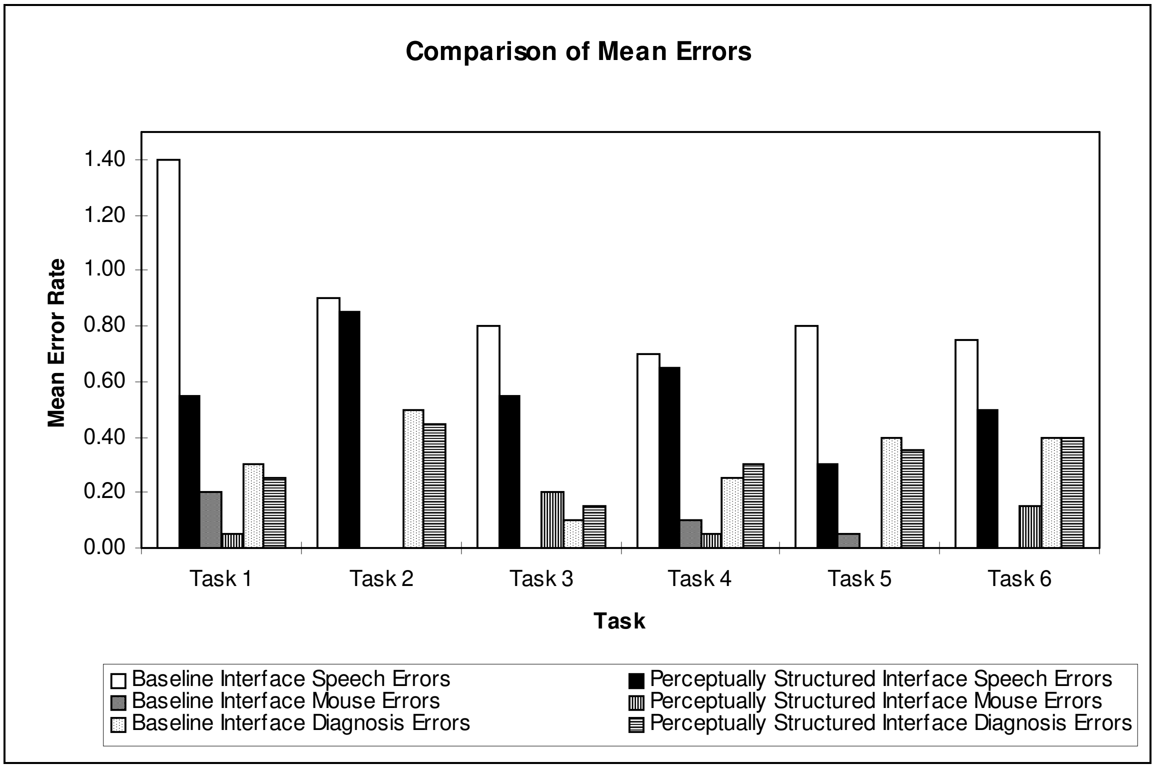

Figure 2: Comparison of Mean Errors

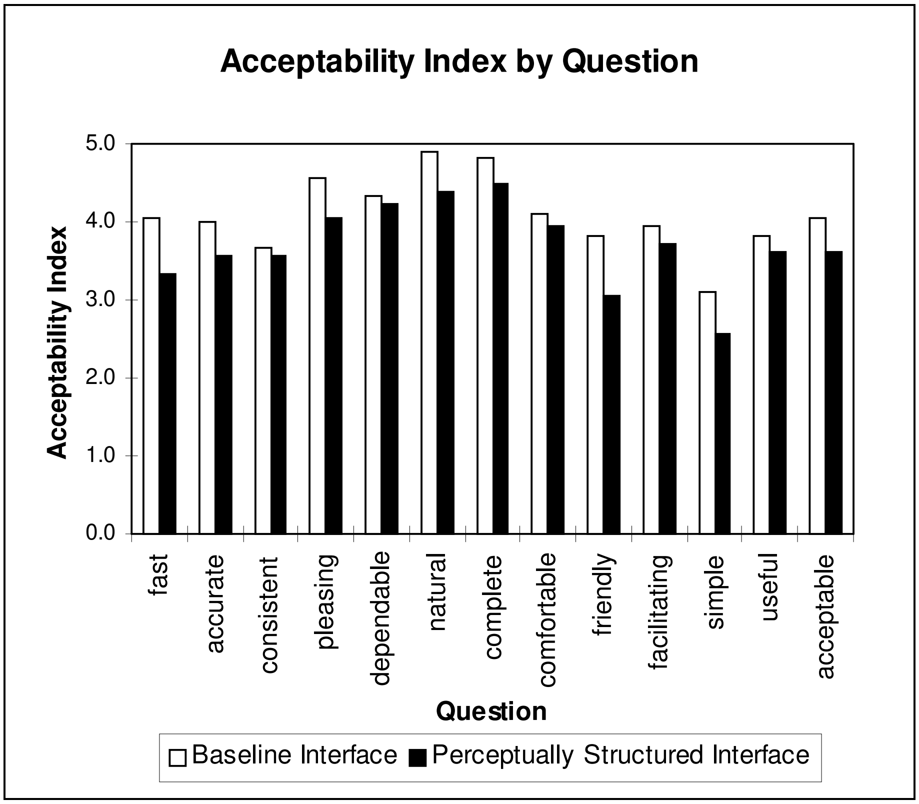

Figure 3: Comparison of Acceptability Index by Question

References

Bradford, J. H. (1995). The Human Factors of Speech-Based Interfaces: A Research Agenda. ACM SIGCHI Bulletin, 27(2):61-67.

Buxton, B. (1993). HCI and the Inadequacies of Direct Manipulation Systems. SIGCHI Bulletin, 25(1):21-22.

Casali, S. P., Williges, B. H., and 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(2):183-196.

Cohen, P. R. (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.

Cole, R., et al. (1995). The Challenge of Spoken Language Systems: Research Directions for the Nineties. IEEE Transactions on Speech and Audio Processing, 3(1):1-21.

Dillon, T. W. (1995). Spoken Language Interaction: Effects of Vocabulary Size, User Experience, and Expertise on User Acceptance and Performance. Doctoral Dissertation, University of Maryland Baltimore County.

Garner, W. R. (1974). The Processing of Information and Structure. Lawrence Erlbaum, Potomac, Maryland.

Garner, W. R. and Felfoldy, G. L. (1970). Integrality of Stimulus Dimensions in Various Types of Information Processing. Cognitive Psychology, 1:225-241.

Grasso, M. A. and Finin, T. (1997). Task Integration in Multimodal Speech Recognition Environments. Crossroads, publication pending.

Grasso, M. A. (1995). Automated Speech Recognition in Medical Applications. M.D. Computing, 12(1):16-23.

Grasso, M. A., Grasso, C. T. (1994). Feasibility Study of Voice-Driven Data Collection in Animal Drug Toxicology Studies. Computers in Biology and Medicine, 24(4):289-294.

Jacob, R. J. K. et al. (1994). Integrality and Separability of Input Devices. ACM Transactions on Computer-Human Interaction, 1(1):3-26.

Jones, D. M., Hapeshi, K., and Frankish, C. (1990). Design Guidelines for Speech Recognition Interfaces. Applied Ergonomics, 20:40-52.

Landau, J. A., Norwich, K. H., Evans, S. J. (1989). Automatic Speech Recognition - Can it Improve the Man-Machine Interface in Medical Expert Systems? International Journal of Biomedical Computing, 24:111-117.

Oviatt, S. L. (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.

Oviatt, S. L. and 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.

Pathology Code Table Reference Manual, Post Experiment Information System (1985). National Center for Toxicological Research, TDMS Document #1118-PCT-4.0, Jefferson, Ark.

Peacocke, R. D. and Graf, D. H. (1990). An Introduction to Speech and Speaker Recognition. IEEE Computer, 23(8):26-33.

Pomerantz, J. R. and Lockhead, G. R. (1991). Perception of Structure: An Overview. In The perception of Structure, pp. 1 - 20, American Psychological Association, Washington, DC.

Shneiderman, B. (1993). Sparks of Innovation in Human-Computer Interaction, Ablex Publishing Corporation, Norwood, NJ.