The Integrality of Speech in Multimodal Interfaces

Michael A. Grasso, Ph.D.[1, 2], David Ebert, Ph.D.[2], Tim Finin, Ph.D.[2]
Segue Biomedical Computing, Laurel, Maryland[1] and
Department of Computer Science and Electrical Engineering at the
University of Maryland Baltimore County, Baltimore, Maryland[2]
grasso@cs.umbc.edu, ebert@cs.umbc.edu, finin@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 that 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.

Keywords

Direct manipulation, input devices, integrality, medical informatics, multimodal, natural language processing, pathology, perceptual structure, separability, speech recognition.

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, Ebert and Finin 1997]. The specific focus is to develop a theoretical model on the use of speech input with direct manipulation in a multimodal interface. Such information can be used to predict the success of multimodal interface designs using an empirically-based model. The specific objective of this study was to apply the theory of perceptual structure to multimodal interfaces using speech and mouse input. This was based on previous work with multimodal interfaces [Cohen 1992; Oviatt and Olsen 1994] and work that extended the theory of perceptual structure to unimodal interfaces [Jacob et al. 1994].

Multimodal Interfaces

The history of research in multimodal speech and direct manipulation interfaces has led to the identification of two key principles relevant to this research: the complementary framework between speech and direct manipulation, and contrastive functionality. Both principles are introduced along with general background information on speech and direct manipulation interfaces.

Speech Interface

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

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.

Complementary Framework

It has been suggested that direct manipulation and speech recognition interfaces have complementary strengths and weaknesses that could be leveraged in multimodal user interfaces [Cohen 1992]. By combining the two modalities, the strengths of one could be used to offset the weaknesses of the other. For simplicity, we used speech recognition to mean the identification of spoken words, not necessarily natural language recognition, and for direct manipulation we focused on mouse input.

The complementary advantages of direct manipulation and speech recognition are summarized in Figure 1. Note that the advantages of one are the weaknesses of the other. For example, direct engagement provides an interactive environment that is thought to result in increased user acceptance and allow the computer to become transparent as users concentrate on their tasks [Shneiderman 1983]. However, the computer can only become totally transparent if the interface allows hands-free and eyes-free operation. Speech recognition interfaces provide this, but intuitive physical actions no longer drive the interface.

Direct Manipulation	**Speech Recognition**
Direct engagement	Hands/eyes free operation
Simple, intuitive actions	Complex actions possible
Consistent look and feel	Reference does not depend on location
No reference ambiguity	Multiple ways to refer to entities

Figure 1: Complementary Strengths of Direct Manipulation and Speech

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 and Oviatt 1994]. This suggests that user interface performance and acceptance may increase by adopting a multimodal approach that combines speech and direct manipulation. Several applications were proposed where each modality would be beneficial. These are summarized in Figure 2. For example, direct manipulation interfaces wee believed to be best used for specifying simple actions when all references are visible and the number of references are limited, while speech recognition interfaces would be better at specifying more complex actions when references are numerous and not visible.

Direct Manipulation	**Speech Recognition**
Visible References	Non-Visible References
Limited References	Multiple References
Simple Actions	Complex Actions

Figure 2: Proposed Applications for Direct Manipulation and Speech

Contrastive Functionality

A study by Oviatt and Olsen [1994] examined how people might combine input from different devices in a multimodal computer interface. 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 handwriting, speech, or both during testing.

This study evaluated user preferences in modality integration using spoken and written input. Among the findings, it was noted that simultaneous input with both pen and voice was rare. Digits and proper names were more likely written. Also, structured interaction using a form-based approach were more likely written.

However, the most significant factor in predicting the use of integrated multimodal speech and handwriting was what they called contrastive functionality. Here, the two modalities were used in different ways to designate a shift in context or functionality. Input patterns observed were original versus corrected input, data versus command, and digits versus text. For example, one modality was used for entering original input while the other was reserved for corrections.

While this study identified user preferences, a follow-up study explored possible performance advantages [Oviatt 1996]. It was reported that multimodal speech and handwriting interfaces decreased task completion time and decreased errors for certain tasks.

Theory of Perceptual Structure

Along with key principles of multimodal interfaces, the work we present is also based on an extension of the theory of perceptual structure [Garner 1974]. Perception is a cognitive process that occurs in the head, somewhere between the observable stimulus and the response. This response is not just a simple representation of a stimulus, because perception consists of various kinds of cognitive processing with distinct costs. Pomerantz and Lockhead [1991] built upon Garner's work to show that by understanding and capitalizing on the underlying structure of an observable stimulus, it is believed that a perceptual system could reduce these processing costs.

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. Relying on this phenomenon has led to increased efficiency in various activities. For example, a crude 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. Software engineers use metrics from previous projects to predict the outcome of future efforts.

While the concept of structure has a dimensional connotation, Pomerantz and Lockhead [1991] state that structure is not limited to shape or other physical stimuli, but is an abstract property that transcends any particular stimulus. Using this viewpoint, information and structure are essentially the same in that they are the property of a stimulus that is perceived and processed. This allowed us to apply the concept of structure to a set of attributes that are more abstract in nature. That is, the collection of histopathology observations.

Integrality of Stimulus Dimensions

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; Shepard 1991]. The dimensions of a structure are integral if they cannot be attended to individually, one at a time; otherwise, they are separable.

Whether two dimensions are integral or separable can be determined by similarity scaling. In this process, similarity between two stimuli is measured as a distance. Subjects are asked to compare pairs of stimuli and indicate how alike they are. For example, consider the three stimuli, A, B, and C. Stimuli A and B are in dimension X (they differ based on some characteristic of X). Similarly, stimuli A and C are in the Y dimension. Given the values of d_x and d_y, which each differ in one dimension, the value of d_{xy} can be computed.

The distance between C and B, which are in different dimensions, can be measured in two ways as diagrammed in Figure 3. The city-block or Manhattan distance is calculated by following the sides of the right triangle so that $d_{xy} = d_x + d_y$. The Euclidean distance follows the Pythagorean relation so that $d_{xy} = (d_x + d_y)^{1/2}$. This value is then compared to the value between C and B given by the subjects. If the given value for d_{xy} is closer to the Euclidean distance, the two dimensions are integral. If it is closer to the city-block distance, the dimensions are separable.

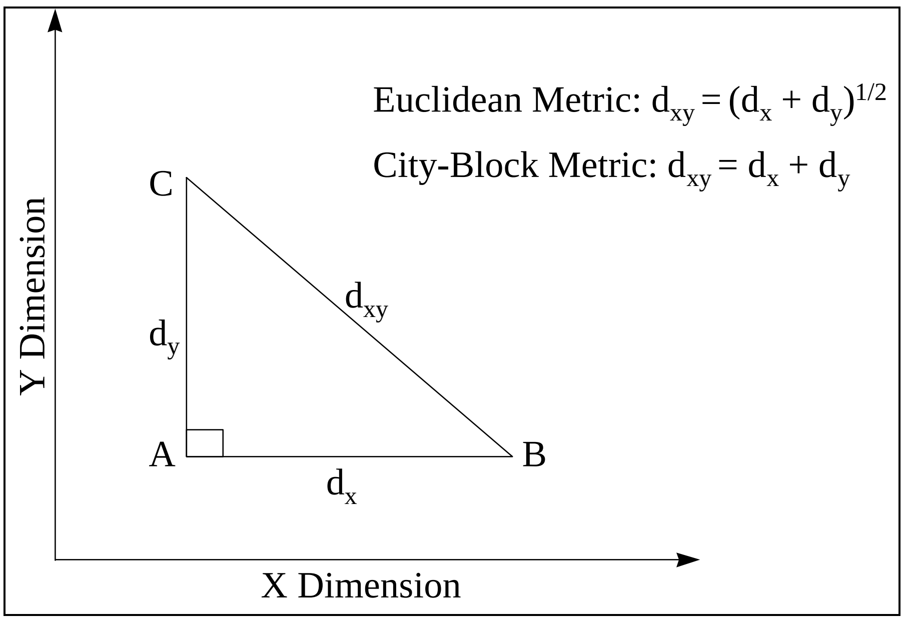

Figure 3: Euclidean Versus City-Block Metrics

Integrality of Unimodal Interfaces

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]. The concept of integral and separable was extended to interactive tasks by noting that the attributes of an input task correspond to the dimensions of an observable stimulus. Also, certain input attributes would be integral if they follow the Euclidean metric, and separable if they follow the city-block metric.

Each input task involved one multidimensional input device, either a two-dimensional mouse or a three-dimensional tracker. 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.

Application of Perceptual Structure to Multimodal Interfaces

Previous work on multimodal interfaces reported that such interfaces should result in performance gains [Cohen 1992]. Also, it was reported that a multimodal approach is preferred when an input task contains a shift in context [Oviatt and Olsen 1994]. This shift in context suggests that the attributes of those tasks were perceptually separable.

In addition, the theory of perceptual structures, integral and separable, was extended with the hypothesis that the perceptual structure of an input task is key to the performance of unimodal, multidimensional input devices on multidimensional tasks [Jacob et al. 1994]. Their finding that performance increased when a separable task used an input device with separable dimensions suggests that separable tasks should be entered with separate devices in a multimodal interface. Also, since performance increased when integral tasks were entered with an integral device suggests that a single device should be used to enter integral tasks in a multimodal interface.

Based on these results, a follow-on question was proposed to determine the effect of integral and separable input tasks on multimodal speech and direct manipulation interfaces. Predicted results were that the speed, accuracy, and acceptance of multidimensional multimodal input would increase when the attributes of the task are perceived as separable, and for unimodal input would 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.

In this experiment, the theory of perceptual structure was applied to a multimodal interface similar to Jacob et al. [1994]. One important difference is that Jacob et al. used a single multidimensional device while we used multiple single dimensional devices. Note that we viewed selecting items with a mouse as a one-dimensional task, while Jacob viewed selected an X and Y coordinate with a mouse as a two-dimensional task. The attributes of the input task correspond to the dimensions of the perceptual space. The structure or redundancy in these dimensions reflects the correlation in the attributes. Those dimensions that are highly correlated are integral and those that are not are separable. The input modality consists of two devices: speech and mouse input. Those input tasks that use one of the devices are using the input modality in an integral way and those input tasks that use both devices are using the input modality in a separable way. This is shown in Figure 4.

Input Device	Perception	Modality
Speech Only	Integral	Unimodal
Mouse Only	Integral	Unimodal
Speech and Mouse	Separable	Multimodal

Figure 4: Input Device Perception Versus Modality

Significance

Studies that 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 that 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].

Methodology

Independent Variables

The two independent variables for the experiment were the interface type and task order. Both variables were counterbalanced as described below. The actual input task was to enter histopathologic observations consisting of three attributes: topographical site, qualifier, and morphology. The site is a location on a given organ. For example, the alveolus is a topographical site of the lung. The qualifier is used to identify the severity or extent of the morphology, such as mild or severe. The morphology describes the specific histopathological observation, such as inflammation or carcinoma. Note that input task was limited to these three items. In normal histopathological observations, there may be multiple morphologies and qualifiers. These were omitted for this experiment. For example, consider the following observation of a lung tissue slide consisting of a site, qualifier, and morphology: *alveolus multifocal granulosa cell tumor*.

The three input attributes correspond to three input dimensions: site, qualifier, and morphology. After considering pairs of input attributes, it was concluded that qualifier and morphology (QM relationship) were related by Euclidean distances and therefore integral. Conceptually, this makes sense, since the qualifier is used to describe the morphology, such as *multifocal granulosa cell tumor*. Taken by itself, the qualifier had little meaning. Also, the site and qualifier (SQ relationship) were related by city-block distances and therefore separable. Again, this makes sense since the site identified what substructure in the organ the tissue was taken from, such as *alveolus* or *epithelium*. Similar to SQ, the site and morphology (SM relationship) was related by city-block distances and also separable. Based on these relationships and the general research hypothesis, Figure 5 predicted which modality would lead to performance, accuracy, and acceptability 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

Figure 5: 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 Figure 6. 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 *congruent* interface, because the choice of input devices was thought to best match the integrality of the attributes. The fifth alternative was the *baseline* interface, since the input devices least match the integrality of the attributes.

Modality	Site	Qual	Morph	SQ	SM	QM	Interface
1. Mouse	M	M	M	-	-	+	
2. Speech	S	S	S	-	-	+	
3. Both	M	S	S	+	+	+	Congruent
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	+	-	-	

Figure 6: Possible Interfaces Combinations for the Software Prototype

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. This was done to minimize any bias related to the layout of information on the computer screen.

It might have been useful to consider mouse-only and speech-only tasks (interface alternatives one and two). However, because of performance differences between mouse and speech input, any advantages due to perceptual structure could not be measured accurately.

The three input attributes mainly involve reference identification, with little declarative, spatial, or computational data entry required. This includes the organ sites, which may be construed as having a spatial connotation. However, most of the sites we selected are not spatial, such as the epithelium, a ubiquitous component of most organs. Also, sites were selected from a list as opposed to identifying a physical location on an organ. This should minimize any built-in bias toward either direct manipulation or speech.

There are some limitations in using the third and fifth alternatives. Note in Figure 4 and in Figure 5 that both the input device and the input attributes can be integral or separable. Figure 7 describes the interface alternatives in these terms. Note that the congruent interface compares a separable device with separable attributes and an integral device with integral attributes. The baseline interface compares a separable device with integral attributes and a separable device with separable attributes. However, neither interface compares an integral device with separable attributes.

	Relationship	Device	Attributes
Alternative 3 (Congruent)	SQ	Separable	Separable
	SM	Separable	Separable
	QM	Integral	Integral
Alternative 5 (Baseline)	SQ	Separable	Integral
	SM	Separable	Separable
	QM	Separable	Integral

Figure 7: Structure of Input Device and Input Attributes

One other comment is that using two input devices to enter histopathology observations would normally be considered counterproductive. These specific user-interface tasks were not meant to identify the optimal method for entering data, but to discover something about the efficiency of multimodal interfaces.

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. A speech error was 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 that has been used in other human computer interaction studies [Casali, Williges, and Dryden 1990; Dillon 1995]. The adjectives are listed in Figure 8. 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		

Figure 8: 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. 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 congruent-interface-first, baseline-interface-second group and were asked to complete six data entry tasks using the congruent interface and then complete six tasks using the baseline interface. The other half of the subjects were assigned to the baseline-interface-first, congruent-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 Figure 9. For example, subjects in group *B1C2* used the baseline interface with slides 1 through 6 followed by the congruent interface with slides 7 through 12. Counterbalancing into these four groups minimized unwanted effects from slide order and vocabulary. For example, during half of the tasks, observations for slides 1 through 6 were entered first while the other half entered slides 7 through 12 first.

	First Task		**Second Task**	
	Interface	**Slides**	**Interface**	**Slides**
B1C2	Baseline	1-6	Congruent	7-12
B2C1	Baseline	7-12	Congruent	1-6
C1B2	Congruent	1-6	Baseline	7-12
C2B1	Congruent	7-12	Baseline	1-6

Figure 9: Subject Groupings for the Experiment

Materials

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 prototype computer program was developed using Microsoft Windows 3.11 (Microsoft Corporation, 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 tissue slides for the experiment were provided by the National Center for Toxicological Research (Jefferson, AK). All the slides were from mouse tissue and stained with H&E. Pictures were taken at high resolution with the original dimensions of 36 millimeters by 24 millimeters. Each slide was cropped to show the critical diagnosis and scanned at two resolutions: 570 by 300 and 800 by 600. All scans were at 256 colors. The diagnoses for the twelve slides are shown in Figure 10.

	Slide	Diagnosis (Organ, Site, Qualifier, Morphology)
Group 1	1	Ovary, Media, Focal, Giant Cell
	2	Ovary, Follicle, Focal, Luteoma
	3	Ovary, Media, Multifocal, Granulosa Cell Tumor
	4	Urinary Bladder, Wall, Diffuse, Squamous Cell Carcinoma
	5	Urinary Bladder, Epithelium, Focal, Transitional Cell Carcinoma
	6	Urinary Bladder, Transitional Epithelium, Focal, Hyperplasia
Group 2	7	Adrenal Gland, Medulla, Focal, Pheochromocytoma
	8	Adrenal Gland, Cortex, Focal, Carcinoma
	9	Pituitary, Pars Distalis, Focal, Cyst
	10	Liver, Lobules, Diffuse, Vacuolization Cytoplasmic
	11	Liver, Parenchyma, Focal, Hemangiosarcoma
	12	Liver, Parenchyma, Focal, Hepatocelluar Carcinoma

Figure 10: Tissue Slide Diagnoses

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 Manufacturer, 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.

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. Normally, a pathologist would examine tissue slides with a microscope. However, to minimize hands-busy or eyes-busy bias, no microscopy was involved. Instead, the software projected images of tissue slides on the computer monitor while participants entered observations in the form of topographical sites, qualifiers, and morphologies. While this might have contributed to increased diagnosis errors, the difference in relative error rates from both interfaces can still be measured. Also, participants were allowed to review the slides and ask clarifying questions as described in the experimental procedure.

The software provided prompts and directions to identify which modality was to be used for which inputs. No menus were used to control the system. Instead, buttons could be pressed to zoom the slide to show greater detail, adjust the microphone gain, or go to the next slide. To minimize bias, all command options and nomenclature terms were visible on the screen at all times. The user did not need to scroll to find additional terms.

A sample screen is shown in Figure 11. In this particular configuration, the user would select a site with a mouse click and enter the qualifier and morphology by speaking a single phrase, such as *moderate giant cell*. The selected items would appear in the box above their respective lists on the screen. Note that the two speech terms were always entered together. If one of the terms was not recognized by the system, both would have to be repeated. A transcript for the congruent and baseline interfaces for one of the subjects is given in Figure 12 and Figure 13.

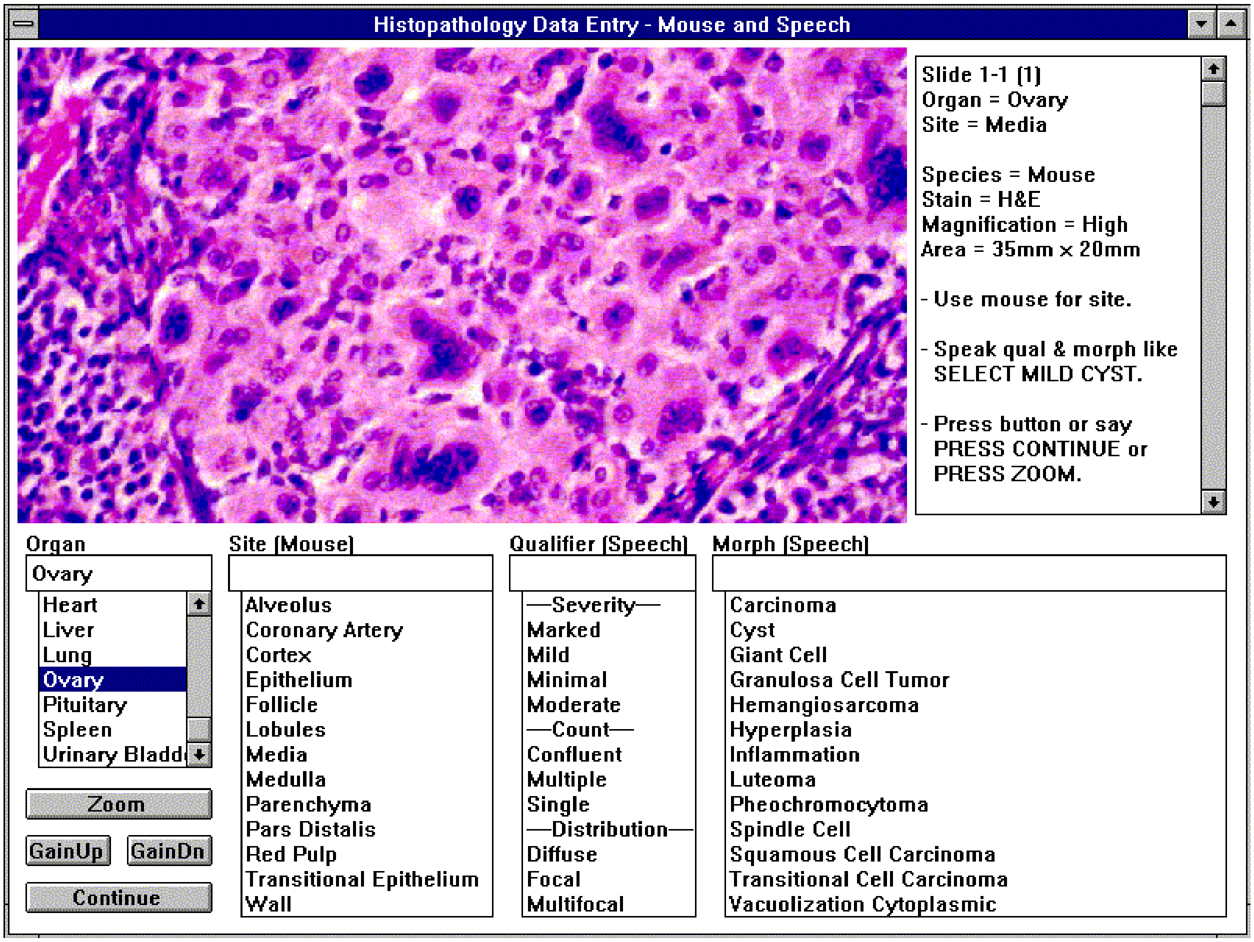

Figure 11: Sample Data Entry Screen

	Time	Device	Action	Comment
Task 1	0	Mouse	Press button to begin test.	
	3	Mouse	Click on “media”	
	7	Speech	“Select marked giant cell”	
	14	Mouse	Click on “press continue” button	
Task 2	20	Mouse	Click on “follicle”	
	29	Speech	“Select moderate hyperplasia”	Recognition error
	36	Speech	“Select moderate hyperplasia”	
	42	Mouse	Click on “press continue” button	
Task 3	44	Mouse	Click on “media”	
	50	Speech	“Select moderate inflammation”	
	57	Mouse	Click on “press continue” button	
Task 4	61	Mouse	Click on “wall”	
	65	Speech	“Select marked squamous cell carcinoma”	
	71	Mouse	Click on “press continue” button	
Task 5	74	Mouse	Click on “epithelium”	
	81	Speech	“Select moderate transitional cell carcinoma”	
	89	Mouse	Click on “press continue” button	
Task 6	94	Mouse	Click on “transitional epithelium”	
	96	Speech	“Select marked transitional cell carcinoma”	
	104	Mouse	Click on “press continue” button	

Figure 12: Congruent Interface Transcript

	Time	Device	Action	Comment
Task 1	0	Mouse	Press button to begin test.	
	15	Mouse	Click on "medulla"	Incorrect action
	20	Speech	"Select medulla mild"	
	21	Mouse	Click on "pheochromocytoma"	
	27	Mouse	Click on "press continue" button	
Task 2	35	Speech	"Select cortex marked"	Recognition error
	39	Mouse	Click on "pheochromocytoma"	
	42	Speech	"Select cortex marked"	
	51	Mouse	Click on "press continue" button	
Task 3	70	Speech	"Select pars distalis moderate"	
	76	Mouse	Click on "granulosa cell tumor"	
	77	Mouse	Click on "press continue" button	
Task 4	82	Speech	"Select lobules marked"	
	88	Mouse	Click on "vacuolization cytoplasmic"	
	89	Mouse	Click on "press continue" button	
Task 5	97	Speech	"Select parenchyma moderate"	Recognition error
	101	Mouse	Click on "hemangiosarcoma"	
	103	Speech	"Select parenchyma moderate"	
	109	Mouse	Click on "press continue" button	
Task 6	114	Speech	"Select parenchyma marked"	Recognition error
	118	Mouse	Click on "hepatocellular carcinoma"	
	124	Speech	Click on "press continue" button	
	128	Mouse	Click on "press continue" button	

Figure 13: Baseline Interface Transcript

Procedure

A within-groups experiment, fully counterbalanced on nput modality and slide order was performed. Each subject was tested individually in a laboratory setting at the participant's place of employment or study. Participants were first asked to fill out the pre-experiment questionnaire to collect demographic information. The subjects were told that the objective of this study was to evaluate several user interfaces in the context of collecting histopathology data and was being used to fulfill certain requirements in the Ph.D. Program of the Computer Science and Electrical Engineering Department at the University of Maryland Baltimore County. They were told that a computer program would project images of tissue slides on a computer monitor while they enter observations in the form of topographical sites, qualifiers, and morphologies.

After reviewing the stated objectives, each participant was seated in front of the computer and had the headset adjusted properly and comfortably, being careful to place the microphone directly in front of the mouth, about an inch away. Since the system was speaker-independent, there was no need to enroll or train the speech recognizer. However, a training program was run, to allow participants to practice speaking typical phrases in such a way that the speech recognizer could understand. The objective was to become familiar speaking these phrases with reasonable recognition accuracy. Participants were encouraged to speak as clearly and as normally as possible.

Next, each subject went through a training session with the actual test program to practice reading slides and entering observations. Participants were instructed that this was not a test and to feel free to ask the experimenter about any questions they might have.

The last step before the test was to review the two sets of tissue slides. The goal was to make sure participants were comfortable reading the slides before the test. This was to ensure that the experiment was measuring the ability of subject to enter data, not their ability to read slides. During the review, participants were encouraged to ask questions about possible diagnoses.

For the actual test, participants entered two groups of six histopathologic observations in an order 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 as a post-experiment questionnaire. A summary of the experimental procedure can be found in Figure 14.

	Task
Step 1	Pre-experiment questionnaire and instructions
Step 2	Speech training
Step 3	Application training
Step 4	Slide review
Step 5	Evaluation and quantitative data collection
Step 6	Post-experiment questionnaire and subjective data collection

Figure 14: Experimental Procedure

Statistical Analysis

Basic assumptions about the distribution of data were used to perform the statistical analysis. The Central Limit Theorem states that for a normal population with mean μ and standard deviation σ, the sample mean observed during data collection is normally distributed with mean μ and standard deviation $\sigma / n^{1/2}$, provided the number of observations n in the sample is sufficiently large and the sample mean is genuinely unbiased by the random allocation of conditions [Noether 1976]. Several null hypotheses were derived from the general research hypothesis stating that there was no difference between the subject groups (i.e, that the experimental manipulation did not effect the results). Each null hypothesis was tested by computing the probability of randomly obtaining those same results. If the probability indicates that the result did not occur simply by chance, then the null hypothesis could be safely rejected [Johnson 1992].

As stated earlier, a within-groups experiment, fully counterbalanced on input modality and slide order was performed. The data collected consisted of pairs of measurements taken on the same subjects, with the results analyzed as a single sample of differences. The F test and t test were used to determine if different samples came from the same population, for example, the baseline-interface-first and the baseline-interface-second groups. Finally, regression analysis was used to identify relationships between any of the dependent variables.

Results

For each participant, speed was measured as the time to complete the 6 baseline interface tasks, the time to complete the 6 congruent interface tasks, and time improvement (baseline interface time - congruent 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 comparison of mean task completion times is in Figure 15. For each subject, the 6 baseline and 6 congruent tasks are graphed.

A two-factor ANOVA with repeated measures was run as well to show if the results were significant. A 2 x 4 ANOVA was set up to compare the 2 interfaces with the 4 treatment groups. The sample variation comparing the baseline interface times to the congruent interface times was significant ($p = .028$). The ANOVA showed that the interaction between interface order and task order had no significant effect on the results ($p = .903$).

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 congruent interface had mean of 3.40. The reduction in speech errors was significant (paired $t(19) = 2.924$, $p < .009$, two-tailed). Mouse errors for the baseline interface had mean error rate of 0.35, while the congruent 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 congruent interface had mean of 1.85. Although the rate for the congruent interface was slightly better, these results were not significant (paired $t(19) = 0.181$, $p = 0.858$, two-tailed). A comparison of mean speech error rates by task is shown in Figure 16. Similar to task completion times, a two-factor ANOVA with repeated measures was run for speech errors to show that the sample variation was significant ($p = .009$) and that the interaction between interface order and task order had no significant effect on the results ($p = .245$).

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. One subject’s score was more than 2 standard deviations outside the mean AI and was rejected as an outlier. This person answered every question with the value of 1, resulting in a mean AI of 1. No other subject answered every question with the same value, suggesting that this person did not give ample consideration. With this outlier removed, the baseline interface AI was 3.99 and the congruent 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 ($p = .014$) and the interaction between groups was not ($p > .999$). A comparison of these values is shown in Figure 17.

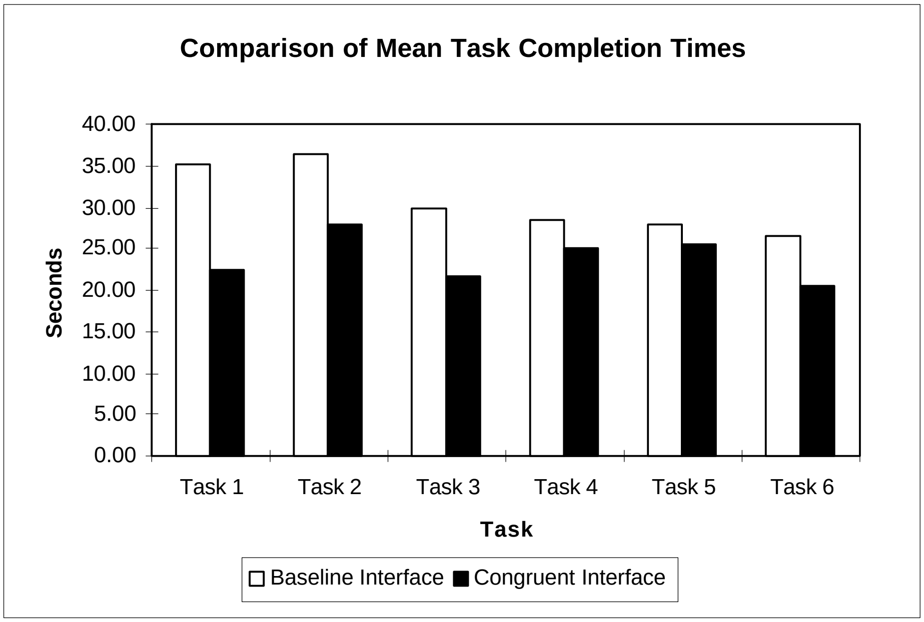

Figure 15: Comparison of Mean Task Completion Times

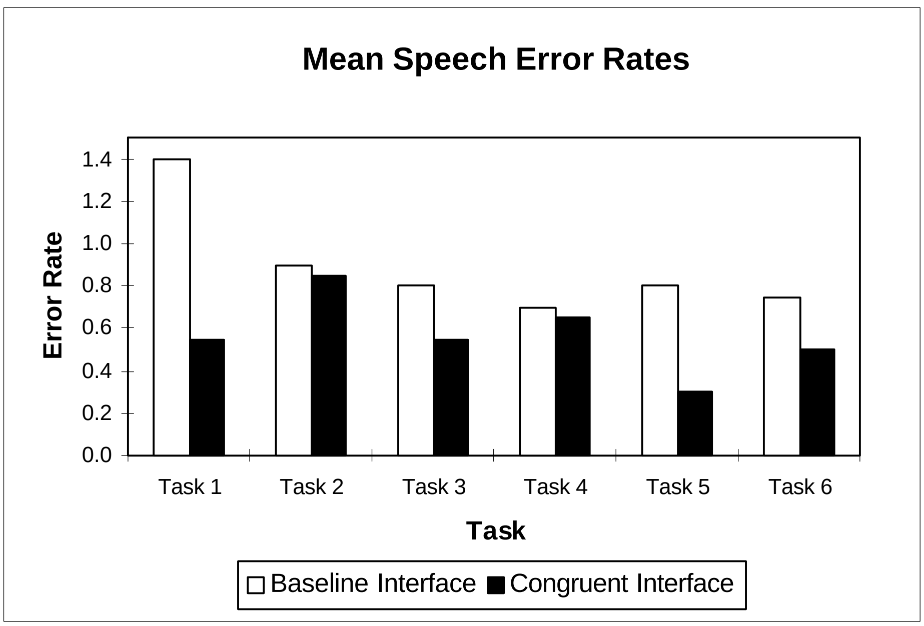

Figure 16: Comparison of Mean Speech Errors

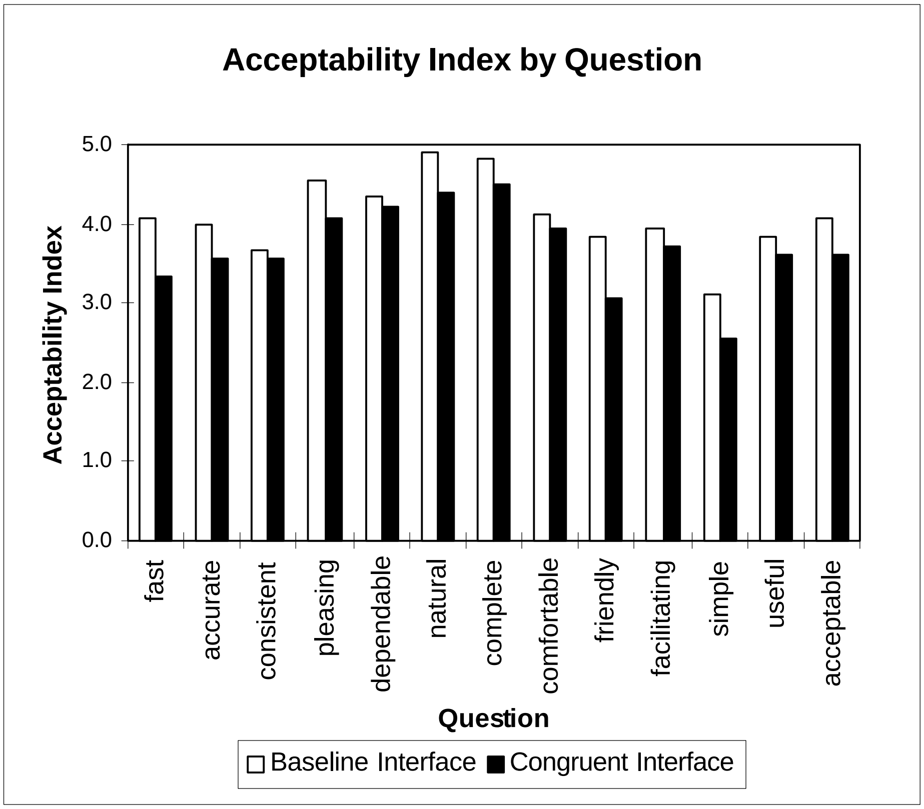

Figure 17: Comparison of Acceptability Index by Question

Discussion

The results of this study show that the congruent interface was favored over the baseline interface. This supports the hypothesis that the perceptual structure of an input task is an important consideration when designing a multimodal computer interface. As shown in Figure 7, the QM relationship compared entry of integral attributes with an integral device in the congruent interface and a separable device in the baseline interface. Based on this, the three null hypotheses were rejected in favor of alternate hypotheses stating that performance, accuracy, and user acceptance were shown to improve when integral attributes are entered with a single device. However, since separable attributes were not tested with both integral and separable devices, no conclusion can be made about whether it was advantageous to enter separable attributes with either a single device or multiple devices. In addition, several significant relationships between dependent variables were observed.

With respect to accuracy, the results were only significant for speech errors. Mouse and diagnosis errors showed a slight improvement with the baseline group, but these were not significant. This 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 the results were not statistically significant. 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.

Pearson correlation coefficients were computed to reveal possible relationships between the dependent variables. This includes relationships between the baseline and congruent interface, relationships with task completion time, and relationships with user acceptance.

A positive correlation of time between the baseline interface and congruent 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 congruent 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 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 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 1974] 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 supported the hypothesis that the perceptual structure of an input task is an important consideration when designing a multimodal computer interface. Task completion time, accuracy, and user acceptance all increased when a single modality was used to enter attributes that were integral. A biomedical 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 congruent interface used speech and mouse input in a way that best matched the perceptual structure. It should be noted that this experiment did not determine whether or not a unimodal speech-only or mouse-only interface would perform better overall. It also did not show whether separable attributes should be entered with separate input devices or one device. However, for input tasks that use a multimodal approach, this work provided evidence that integral attributes should be entered with a single device.

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 errors decreased slightly and diagnosis errors increased slightly for the baseline interface, but these were not statistically significant. User acceptance was 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. This could include a more direct evaluation of perceptual structure on separable data. Another approach could include studies on minimizing speech errors. The reduction of speech errors has typically been 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 speech errors 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.

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.

Cohen, P. R. and Oviatt, S. L. (1994). The Role of Voice in Human-Machine Communication. In Voice Communication Between Humans and Machines, pp. 34-75, National Academy Press.

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.

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

Grasso, M. A. Ebert, D. S. and Finin, T. W. (1997). Acceptance of a Speech Interface for Biomedical Data Collection. AMIA 1997 Annual Fall Symposium.

Grasso, M. A. and 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.

Johnson, P. (1992). Evaluations of Interactive Systems. In Human-Computer Interaction. McGraw-Hill, New York, pp. 84-99.

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., and 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.

Noether, G. E. (1976). Introduction to Statistics: A Nonparametric Approach. Houghton Mifflin Company, Boston, page 213.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. (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.

Shepard, R. N. (1991). Integrality Versus Separability of Stimulus Dimension: From an early Convergence of Evidence to a Proposed Theoretical Basis. In The perception of Structure, pp. 53 - 71, American Psychological Association, Washington, DC.

Shneiderman, B. (1983). Direct Manipulation: A Step Beyond Programming Languages. IEEE Computer, 16(8):57-69.

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